Discours
Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics

16 | 2015
Varia

Correspondences between Czech and English Coreferential Expressions

Michal Novák and Anna Nedoluzhko

Electronic version
URL: http://journals.openedition.org/discours/9058
DOI: 10.4000/discours.9058
ISSN: 1963-1723

Publisher:
Laboratoire LATTICE, Presses universitaires de Caen

Electronic reference
Michal Novák and Anna Nedoluzhko, « Correspondences between Czech and English Coreferential Expressions », *Discours* [Online], 16 | 2015, Online since 09 September 2015, connection on 30 April 2019. URL : http://journals.openedition.org/discours/9058 ; DOI : 10.4000/discours.9058

discours
Revue de linguistique, psycholinguistique et informatique
http://discours.revues.org/

Correspondences between Czech and English Coreferential Expressions

Michal Novák
Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Charles University in Prague
mnovak@ufal.mff.cuni.cz

Anna Nedoluzhko
Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Charles University in Prague
nedoluzko@ufal.mff.cuni.cz

Michal Novák et Anna Nedoluzhko, « Correspondences between Czech and English Coreferential Expressions », *Discours* [En ligne], 16 | 2015, mis en ligne le 9 septembre 2015.

URL : http://discours.revues.org/9058

Titre du numéro : *Varia*
Coordination : Lydia-Mai Ho-Dac et Frédéric Landragin

Correspondences between Czech and English Coreferential Expressions

Michal Novák
Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Charles University in Prague

Anna Nedoluzhko
Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Charles University in Prague

In this work, we present a comprehensive study on correspondences between certain classes of coreferential expressions in English and Czech. We focus on central pronouns, relative pronouns, and anaphoric zeros. We designed an alignment-refining algorithm for English personal and possessive pronouns and Czech relative pronouns that improves the quality of alignment links not only for the classes it aimed at but also in general. Moreover, the instances of anaphoric expressions we focus on were manually annotated with their alignment counterparts, which served as a basis for this empirical study. The collected statistics of correspondences are contrasted with theoretical assumptions regarding the use of anaphoric means in the languages under analysis, such as pro-drop properties, the use of finite and non-finite constructions, etc. Finally, we present the ways how the observed correspondences can be exploited in cross-lingual coreference resolution.

Keywords: coreference, coreferential expressions, cross-lingual study, parallel corpus, alignment

1. Introduction

1 Coreference is one of the main pillars of maintaining coherence in a discourse. As far as we know, the fundamentals of this concept, i.e., repeated references to entities playing more or less important roles in the discourse as well as references to previous discourse segments, are shared across all languages. However, as we start to examine this concept in a given language more closely, we find that languages may vary considerably in the means they usually use to express coreference relations.

2 Our work focuses on the comparison of coreferential expressions in two typologically different languages – Czech and English. The differences between these two languages also concern the means of expressing coreference. For illustration, let us sketch out the differences in the following example [1]:

1. All the examples in the following text are presented in the same four-part format:
Line 1: Sentence in the language which is primary to the phenomenon under consideration (in bold).
Line 2: Its translation in the other language.
Line 3-4: Aligned words or phrases of the English and Czech sentence, which are usually reordered. Special symbols may be inserted: "∅" (possibly followed by its semantic role) stands for an ellipsis (zero), i.e., a full-fledged member of the sentence present in its meaning but not expressed on the surface; "—" stands for no counterpart. Some English phrases may be extended with a literal English translation of its Czech counterpart (in square brackets) if the original phrase is not literal enough.

[1] It switched to a caffeine-free formula using its new Coke in 1985.
V roce 1985 přešla na bezkofeinovou recepturu, kterou používá pro svojí novou kolu.

It	switched	to	a caffeine-free	formula	∅[which]	using[it uses]
∅	přešla	na	bezkofeinovou	recepturu,	kterou	používá

its[for its]	new	Coke	in	—[the year]	1985.
pro svojí	novou	kolu	v	roce	1985.

3 Let us look at the coreferential means represented in this sentence pair. The first difference between English and Czech can be seen in the subject of the main clause. While expressed by the personal pronoun *it* in English, the subject in Czech is elided. This is a common difference between 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. Second, we have a participle construction *using its new Coke* that is translated into 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 into Czech with the reflexive possessive pronoun *svůj*, a category missing in English.

4 In this work, we collected coreferential expressions in both languages, along with their translation counterparts in a parallel corpus, to form comprehensive statistics of translation correspondences. We concentrate on the coreferential expressions that are tied closely to syntactic rules of grammar, such as different types of pronouns and anaphoric zeros, and disregard, for instance, nouns, which are less affected by the syntactic patterns of a language. Therefore, throughout this work, we mainly emphasize the differences in terms of syntax and deep syntax.

5 To ensure that the statistics are as reliable as possible, we aligned coreferential expressions in the underlying parallel corpus manually. In addition, we designed a word-alignment algorithm that served as an automatic pre-annotation step prior to the manual annotation. The algorithm focuses on selected types of coreferential expressions and takes advantage of word alignment obtained in a standard unsupervised manner, syntactic structures, and certain regularities observed in the data.

6 All in all, the contribution of this work is threefold. First, we propose a rule-based aligner that performs for the selected coreferential expressions better than the unsupervised approach. Second, we create manually annotated alignments of coreferential expressions, and third, we collect comprehensive statistics of what the nature of the correspondences is. We consider the latter two contributions the most valuable for future work, since their combination in a supervised machine learning approach has a potential to outperform the presented rule-based approach to word-alignment.

7 High-quality aligner of coreferential expressions and the statistics of their translation regularities are also valuable for what is the main motivation for this work. It stems primarily from computational processing of language, especially from the tasks of machine translation and coreference resolution. While the motivation

for machine translation is straightforward, i.e., to observe the patterns of typical translations of given constructions, let us explain the motivation for coreference resolution, which may not be so clear. First, a comparison of how we defined the classes of coreferential expressions and how many such expressions the classes actually contain might enhance the quality of anaphor detection, i.e., a subtask of coreference resolution that deals with identifying the words that can be anaphoric. This issue arises, for instance, in English relative pronouns, many of which are homonymous with conjunctions, interrogatives and fused pronouns, none of them being anaphoric (see more in Section 4 and 7.4). Second, for some expressions it may be easier to find their antecedents than for others, e.g., reflexive pronouns usually corefer with the subject of the sentence, which does not necessarily hold for personal pronouns. The complexity of finding the antecedent may also vary across languages within the same class of expressions. For instance, English reflexive pronouns might be easier to resolve than Czech ones because Czech pronouns do not carry additional information on the antecedent's gender. These varying levels of complexity may be exploited by training a cross-lingual coreference resolution system for parallel texts that performs better than using a monolingual system for each of the languages. Since a cross-lingual system takes advantage of features from both languages, the quality of the alignment of potentially coreferential expressions is essential. Even though this kind of system can be applied solely to parallel texts, we believe that better automatic coreference annotation on a larger parallel dataset may be exploited to improve the quality of monolingual resolution as well. The techniques of semi-supervised learning, e.g., self-training or co-training (Blum & Mitchell, 1998) can be used for this purpose. Although coreference resolution is the main motivation of this work, we do not address this task here and leave it for future work.

8 From the perspective of theoretical linguistics, the comprehensive statistics of corresponding means of coreference is a unique source for comparative research into anaphoric expressions in the languages under analysis. The resulting English-Czech counterparts will make it possible to address such typologically interesting linguistic problems as pro-drop qualities of Czech, the expression of possessivity in Czech and English and its correspondence with the grammatical category of definiteness, the competition of relative clauses and non-finite constructions in English and Czech, and so on. Moreover, we believe that analyzing coreferential means in a language from a multilingual perspective is not only beneficial for cross-lingual comparisons, but also helps to understand this phenomenon more deeply in each individual language.

9 This study is based on English texts and their Czech translations. Even though the translation direction might introduce some bias, we believe that the basic shape of the statistics would remain the same even if it was collected on texts with the opposite translation direction.

10 The paper is structured as follows. Section 2 introduces the studies that have focused on similar phenomena from both a theoretical and a computational perspective. In Section 3, we describe the data on which the subsequent study was carried out.

The classes of potentially coreferential expressions are formally defined in Section 4. Then we proceed to the crucial part of this work in Section 5 – presenting three approaches to cross-lingual word alignment: the original alignment obtained by an unsupervised machine learning method, a rule-based algorithm that builds upon the original alignment, and manual annotation of the alignment links. Having the manual annotation at our disposal, we evaluate the former two approaches to alignment in Section 6. The most extensive part of the work follows in Section 7. We comprehensively examine all the classes of potentially coreferential expressions and assess their most frequent counterparts in the other language. In Section 8, we discuss the results obtained, and we conclude the work in Section 9.

2. Related work

11 The fact that anaphoric expressions function differently in typologically different languages is at the heart of the theory of topicality introduced in (Givón, 1983) and widely used in linguistic typology.

12 During the last few decades, the development of parallel corpora made it possible to compare coreferential expressions in various languages on the basis of large-scale annotated data. However, with the exception of the Romanian-English corpus (Postolache et al., 2006), coreference-annotated parallel corpora have only recently emerged: the manually annotated Prague Czech-English Dependency Treebank 2.0 (Hajič et al., 2012) and German-English ParCor 1.0 (Guillou et al., 2014), and automatically annotated CzEng 1.0 (Bojar et al., 2012). Hence, as far as we know, there is a very small number of bilingual studies on anaphoric expressions based on large-scale annotated parallel corpora, even though the need for such research was pointed out in several works, e.g., Kunz (2010), Kibrik (2011) and Nedoluzhko et al. (2015).

13 Several case studies on anaphoric expressions were recently reported, e.g., a detailed study of abstract pronominal anaphors and label nouns in German and English by Zinsmeister et al. (2012), an analysis of variation in English and German nominal coreference (Kunz, 2010) and an analysis of coreferential chains for English and German parallel and comparable corpora across various registers (Kunz & Lapshinova-Koltunski, 2015). While very interesting from the linguistic point of view, these studies are more oriented towards textual phenomena, focusing on contextual and stylistic factors rather than syntactic ones, which are the point of our present analysis. The comparison of possessive pronouns in Czech and English fiction texts, with a special focus on those used with the parts of human body, by Onderková (2009) is more syntactically oriented and proposes a series of inspiring assumptions that can be proved by corpus analysis on large-scale parallel data.

14 As mentioned in Section 1, one of the motivations of this work is using parallel data to improve the quality of coreference resolution. This task was addressed in (Souza & Orăsan, 2011), where coreference resolution was applied to a corpus with no manual

coreference annotation, with coreferential chains being automatically projected from parallel English texts. A similar technique was also applied on the parallel English-Romanian corpus (Postolache et al., 2006). Regarding the Czech-English language pair, Veselovská et al. (2012) examined the functions of the English pronoun *it* and reasons for a missing subject in Czech. Furthermore, they built a system for detecting anaphoricity of *it* and Czech subjects, experimenting also with information from parallel texts. A recent work by Novák and Žabokrtský (2014) motivated the present study to a certain extent. The authors took advantage of a parallel treebank and built a resolver of English pronoun coreference operating on a bitext, using the aligned Czech text to aid the resolution. The work also presents a supervised word aligner, which was trained on the data annotated within the present study.

15 High quality word alignment is crucial to most cross-lingual techniques. Pronouns resemble function words in that they usually carry several functions, which makes them more difficult to be correctly aligned using the standard unsupervised approach based on IBM models (Brown et al., 1993), represented by its most popular implementation GIZA++ (Och & Ney, 2000). The idea of taking advantage of syntax and better alignment of content words used in the present work was already presented by, e.g., Hermjakob (2009) and Zhang and Zhao (2013). Whereas both the former and the present work extensively use linguistic knowledge for alignment filtering (knowledge of English and Arabic in case of the former work), the latter work resembles the present one in the way how syntactic trees (phrase trees in case of the latter work) are used to select alignment candidates.

16 Our research on translation correspondences of coreferential expressions can also be beneficial for the task of machine translation, where coreferential relations are a recurring issue. It has been addressed in Le Nagard and Koehn (2010) and Hardmeier and Federico (2010) with unsatisfying results. Guillou (2012) advanced this topic by conducting an experimental investigation on Czech-English data into why coreference information fails to improve the quality of translation. The work by Novák et al. (2013a and b) proposed specialized models for translating English reflexive pronouns and the pronoun *it* within a syntax-based English-to-Czech machine translation system TectoMT (Žabokrtský et al., 2008), taking advantage of some of the correspondences that we observe and quantify in the present work.

17 Anaphoric devices may be used differently depending on whether the text is original or translated. The research on the differences between translations and original texts, which can be quite striking, is presented in detail, e.g., by Baker (1995) and Baroni and Bernardini (2006).

3. Parallel data

18 The Prague Czech-English Dependency Treebank 2.0 (PCEDT) is a Czech-English parallel corpus of 1.2 million words comprising almost 50,000 sentences for each language. The English part consists of the Wall Street Journal (WSJ) section of

the Penn Treebank (Marcus et al., 1999). The Czech part was manually translated from the English source sentence by sentence.

19 The linguistic annotation in PCEDT draws on the framework of Functional Generative Description (FGD) (Sgall, 1967; Sgall et al., 1986) and is divided into the following annotation layers: the lowermost "word" layer (w-layer) representing the tokenized plain text, the morphological layer (m-layer) containing automatic part-of-speech tagging and lemmatization, the analytical layer (a-layer) representing surface dependency syntax, and the deep syntax or tectogrammatical layer (t-layer). The t-layer includes semantic labeling of content words (nouns, adjectives, adverbs, and verbs) and coordinating conjunctions, ellipsis reconstruction, coreference annotation, and argument structure description based on a valency lexicon.

20 While annotations on the Czech m-layer and a-layer were performed automatically, the English dependency trees on the a-layer were converted from the original phrase-structures in Penn Treebank. On the other hand, the t-layer in both languages was annotated manually.

21 An overview of the underlying linguistic theory (tectogrammatical annotation) with some details on the most important features such as valency annotation, ellipsis reconstruction, etc. can be found in (Hajič et al., 2012). Samples of the data visualized in a web browser are available on the PCEDT web site [2]. Figure 1 shows a tectogrammatical representation of the sentence pair from example [1].

22 Coreference links in PCEDT have been annotated manually, with an individual treatment of the Czech and English parts (Nedoluzhko et al., 2014). Following FGD, two types of coreference relations are distinguished in PCEDT: *grammatical* and *textual coreference*.

23 **Grammatical coreference.** It is denoted by normal solid arrows in Figure 1. It includes the following subtypes of relations, which appear as a consequence of language-dependent grammatical rules:

1. Reflexive pronoun coreference. In this case, the anaphoric pronoun mostly refers to the closest subject, cf. *My daughter likes to dress herself without my help*, where the reflexive pronoun *herself* corefers with the subject *daughter*.
2. Coreference with relative elements. Relative pronouns and pronominal adverbs introducing relative clauses are linked to their antecedent in the governing clause, cf. *Alex is the boy who kissed Mary*, where the relative pronoun *who* corefers with the noun *boy* modified by the dependent relative clause.
3. Control – a type of grammatical coreference that arises with certain verbs, called control verbs, such as *begin*, *let*, *want*, etc. The control relation arises, for example, with the elided subject of the infinitive *sleep* and the subject *Peter* in the sentence *Peter wants to sleep*.

2. See: http://ufal.mff.cuni.cz/pcedt2.0.

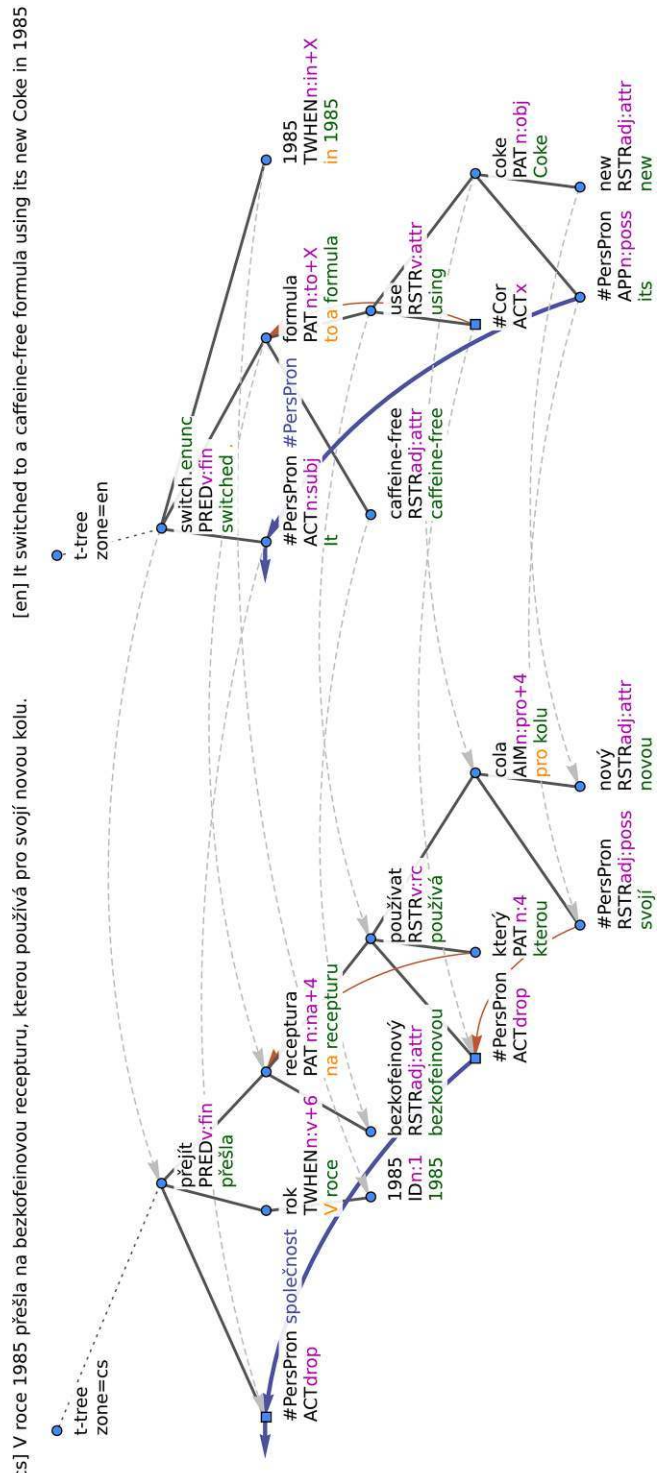

Figure 1. A tectogrammatical representation of a sample sentence pair from PCEDT with grammatical and textual coreference links and node alignment links denoted by normal solid, bold solid, and dashed arrows, respectively

4. Coreference with verbal modifications that have dual dependency. In this case, grammatical coreference concerns non-expressed arguments of verbal modifications with the so-called dual dependency (e.g., passive participles, gerunds, infinitives). This is, for example, the case of coreference of the unexpressed subject with the infinitive *run* with the object *Mary* of the governing verb *saw* in *John saw Mary run around the lake*.
5. Coreference in constructions with reciprocity, cf. the elided object in *John and Mary kissed*.

24 All of the above types of grammatical coreference are the subject of our present study, with the exception of reciprocal constructions.

25 **Textual coreference.** It is denoted by bold solid arrows in Figure 1. Its arguments are not realized by grammatical means alone, but also via context (e.g., central pronouns in the third person, demonstrative pronouns and some cases of anaphoric zeros). In this work, we are concerned only with grammatical coreference and those cases of textual coreference where anaphoric expressions are represented by third person pronouns (including anaphoric zero pronouns) [3].

26 The English and Czech sections of PCEDT are aligned on both sentence and word levels. The sentence alignment is a natural consequence of the fact that the Czech side was created by translating the English one. The words in each sentence pair are aligned automatically on the a-layer as well as the t-layer (denoted by dashed arrows in Figure 1). We will describe the alignment in greater detail in Section 5.

3.1. Data subset under analysis

27 The present work involved manual annotation of word alignment. Given the size of PCEDT, processing the entire PCEDT would be extremely time-demanding. We therefore limited the dataset to only the first half of the PCEDT section 19, i.e., the 50 documents from `wsj_1900` to `wsj_1949`. Table 1 shows some of the basic statistics related to the present work calculated on this dataset.

	English	**Czech**
Sentences	1,078	1,078
T-layer nodes	18,611	20,696
Coreferential	1,362 (7.3%)	1,440 (6.9%)
Grammatical	763 (56%)	568 (40%)
Textual	599 (44%)	872 (60%)

Table 1. The basic statistics of the dataset used in this work

3. The newest version of PCEDT also includes the annotation of pronoun coreference for the first and second person, as well as nominal coreference, see Nedoluzhko et al. (2014).

4. Classes of inspected nodes

28 We focus here on three special classes of anaphoric nodes: central pronouns[4] in the third person, relative pronouns, and anaphoric zeros.

29 For the purpose of the task of coreference resolution, the potentially anaphoric nodes must be identified in the data without using the coreference information. A typical approach is to use heuristic rules that define the set of such nodes more broadly, in order to ensure as high recall as possible. On the other hand, precision must also be kept high, since the inclusion of many non-anaphoric nodes would negatively affect the quality as well as the time complexity of resolution.

30 We decided to select the potentially anaphoric nodes mostly based on their surface-level and deep-level lemmas and grammatical categories. In a few cases, an additional constraint that takes advantage of a syntactic structure was imposed, e.g., in the case of the English relative pronoun *that* introducing a subordinate declarative clause, which is never anaphoric.

31 Our rules defining the three classes of potentially anaphoric nodes cover 99% and 95% of coreferential nodes in English and Czech, respectively. The rest amounts to nodes representing reciprocity and Czech demonstrative pronouns, which were deliberately excluded due to time reasons. Table 2 shows the number of the nodes covered and how many of them are coreferential per each class in both languages. The precision of coverage of the coreferential nodes reaches around 95%, with the average value being 89% and 92% in English and Czech, respectively. The only outliers are English relative pronouns, which will be justified in Section 7.4.

	English		**Czech**	
	Covered	Coreferential	Covered	Coreferential
Central pronouns	578	537 (93%)	286	284 (99%)
Relative pronouns	234	151 (65%)	341	302 (89%)
Anaphoric zeros	702	659 (94%)	850	777 (91%)
Total	**1,514**	**1,350 (89%)**	**1,477**	**1,363 (92%)**

Table 2. Node counts per each class of coreferential expressions, containing the number of nodes covered by the proposed rules and how many of them are coreferential

4. *Central pronouns* is a term coined by Quirk et al. (1985) embracing English personal (e.g., *he*, *she*, *him*, *her*), possessive (e.g., *his*, *her*, *mine*), and reflexive pronouns (e.g., *myself*, *themselves*). Using this term for Czech pronouns we mean the class consisting of personal (e.g., *on*, *jemu*, *ní*), possessive (e.g., *jeho*, *jejich*), reflexive (*se*, *si*), and reflexive possessive (*svůj*) pronouns.

4.1. Central pronouns

32 The rules we used to select central pronouns use mainly the t-layer (deep syntax) since we can rely on manual annotation in the PCEDT[5]. In this approach, a node is considered potentially coreferential if it is a personal pronoun (its deep lemma is `#PersPron`) in the third person and there is a corresponding word on the surface for this node. Equivalent rules using just the surface layer would be:

1. In Czech, a central pronoun is a word where:
   - The surface lemma is one of the following: *on*, *jeho*, *se*, *svůj* (corresponding to personal, possessive, reflexive and reflexive possessive pronouns).
   - The word must be in the third person or the person is undefined[6].
2. In English:
   - The word form must be one of the following: *he*, *she*, *it*, *they*, *him*, *her*, *them*, *his*, *its*, *their*, *himself*, *herself*, *itself*, *themselves*.

33 Using gold-standard t-layer annotation helps us avoid disambiguation errors introduced by automatic processing, such as filtering out expressions homonymous with anaphoric central pronouns, e.g., the Czech reflexive particle *se* in *reflexiva tantum* verbs such as *smát se* (lit. *to laugh*) or reciprocal usage of the pronoun *se*, which is not within the scope of this work.

34 On the other hand, we did not attempt to avoid including the pleonastic usage of the English pronoun *it* in constructions such as *It is possible that…*

4.2. Relative pronouns

35 Our rules define this class more broadly than its name suggests. It has been extended by a group of adverbs that act like relative pronouns (in English, e.g., *how*, *where*, *why*). We refer to the whole class as *relative pronouns* for the sake of convenience.

36 Most expressions used to introduce a relative clause in both languages are homonymous with conjunctions, interrogative, or fused pronouns. Except for the case of the English subordinating conjunction *that* mentioned below, we do not disambiguate these pronouns and leave this task for future work.

37 There is no straightforward way to distinguish relative pronouns on the t-layer. Furthermore, in both languages some relative pronouns are not represented by their own node on the t-layer. The rules used to select relative pronouns therefore use only surface-level constraints:

---

5. Note that while the deep layer in the PCEDT is annotated manually, the surface layer is automatic (see Section 3).
6. The latter case concerns reflexives which, unlike the English reflexives, do not carry the person information themselves.

1. In Czech, a relative pronoun is a word where one of the following holds:
   - Its part-of-speech tag corresponds to a relative or interrogative pronoun or the numeral *kolik* (*how much/many*).
   - Its lemma is *kde* (*where*) or *kdy* (*when*) (these adverbs also function as relative or interrogative pronouns).
2. In English:
   - Its part-of-speech tag corresponds to a wh-determiner, a wh-adverb or a (possibly possessive) wh-pronoun.
   - Since the tags were assigned automatically, some occurrences of the relative pronoun *that* were falsely labeled as a subordinating conjunction. In these cases we decided to believe the automatic parse trees more and filter out a potential conjunction *that*, if it was not a leaf node [7].

4.3. Anaphoric zeros

38 Both languages operate with ellipsis, i.e., with elements missing on the surface but present in the meaning of the utterance. We focus on those cases of ellipsis which take part in coreferential relations – the so-called *anaphoric zeros*. Since they are not visible in the text, the decision whether and when they should be introduced into linguistic description varies across different theories. In PCEDT, anaphoric zeros are introduced in the t-layer with a newly established node, which is assigned the t-lemma `#Cor` and `#PersPron` for the ellipsis representing grammatical and textual coreference respectively.

39 The node with the t-lemma `#Cor` should be used to represent an elided controlled argument in control constructions and in constructions with dual dependencies (see Section 3). This holds for Czech. Indeed, the antecedents of such syntactic constructions in Czech are mostly easily reconstructed based on language-dependent grammatical rules. The situation for English is different. The majority of English nodes with t-lemma `#Cor` are arguments of *-ing* and *-ed* participles (see example [2] below) which are coreferential with one of the arguments of the parent of this participle.

[2] The company had sought increases `#Cor`**.ACT** totaling $80.3 million, or 22%.

40 The problem is that English grammar does not require that the argument of the participle in such a position occupying the semantic role of Actor [8] be coreferential with the Actor of the governing node. For example, in the sentence *John bumped into Mary riding a bike* both John and Mary could be the person riding a bike

7. The reason is that on the a-layer of PCEDT, which the automatic parse trees try to mimic, relative pronouns cannot have children.
8. The detailed description of semantic roles used in the Prague-style tectogrammatical annotation can be found in Panevová et al. (2014) and Mikulová et al. (2006).

before the incident. Thus, strictly speaking, this case cannot be considered to be grammatical coreference. This led us not to differentiate between these two types and to denote them with the common term *anaphoric zeros*.

41 Taking all of this into account, a Czech or English t-node is considered an anaphoric zero if the following constraints are fulfilled:

- its deep lemma is `#Cor` or `#PersPron`;
- it is not expressed (as a separate word) on the surface;
- its person[9] cannot be first or second.

5. Aligning Czech and English nodes

42 At this point, we have the classes of coreferential expressions properly defined. In order to examine what kinds of expressions in the other language are their probable translations, alignment between surface words and t-nodes on both language sides of the PCEDT is required.

43 In the following sections, we will present three stages of improving word alignment in PCEDT. We started from the originally provided automatic alignment, which had been built using an unsupervised machine learning method (see Section 5.1), then we applied a rule-based refinement tailored to two subclasses of coreferential expressions (see Section 5.2), and finally, we corrected the alignments manually for all nodes considered by the constraints introduced in Section 4 (see Section 5.3).

5.1. The original PCEDT alignment

44 As mentioned in Section 3, the PCEDT 2.0 includes a one-to-one sentence alignment between its language parts. The treebank also contains alignment between Czech and English nodes in both surface and t-layer trees.

45 Since the nodes in the surface dependency tree correspond one-to-one to tokens of the sentence, it was possible to employ a standard GIZA++ unsupervised word alignment (Och & Ney, 2000). The authors of PCEDT applied this tool in both directions, including the intersection of the two alignments and the result of the popular symmetrization heuristics (grow-diag-final-and) in the treebank.

46 The alignment of t-layer nodes was obtained by a projection of the alignment from the analytical layer, followed by rule-based heuristics for nodes that remained unaligned. This included aligning the nodes with the same semantic roles whose

9. It is possible to identify the person of anaphoric zeros using the governing verb (or if need be the auxiliary verbs) for Czech. However, we decided rather to annotate a few more examples than to miss some valuable occurrences by potentially erroneous heuristics. We expected the number of these superfluous examples not to be high, as the PCEDT texts are in the news domain that generally prefers using the third person to the other ones.

parents were already aligned. This technique was designed to cover unexpressed subject pronouns (mostly in Czech), which were reconstructed on the t-layer.

5.2. Rule-based improvements on top of the original alignment

47 One can spot at first glance that the automatic alignment performs much worse for function words than for content words. Pronouns are not usually considered to be function words, but, similarly to them, they are more tied by the syntactic rules of a particular language and their interpretation often depends on the context. Inspired by the final rule-based stage of the original PCEDT alignment, we wanted to take advantage of the manual monolingual t-layer annotation and exploit it to refine the existing alignment links and introduce new ones.

48 The algorithm we propose builds upon the original PCEDT alignment (mostly) obtained by GIZA++. It consists of a sequence of multiple rules in the form of *selector-filter* processing pairs, where the selector creates a selection of nodes, which are subsequently filtered based on certain criteria using the filter.

49 The selector works as follows: making use of the dependency relations within the trees and the original alignment links, it suggests a set of possible candidates for the input node's counterpart in the second language. For instance, the simplest selector picks all the nodes aligned with the input node itself. Another possible selector could use the parent of a given node and return children of every node aligned with this parent as a set of candidates.

50 The purpose of the filter is the following: given the candidates obtained by the selector and certain criteria, it filters out the nodes that do not meet the criteria. A filtering criterion typically depends on the selector that precedes it. A selector which uses an input node's parents is usually coupled with a filter that discards all the candidates but the one which shares the semantic role with the input node. However, the criterion is also often tied to the type of the input node, which makes this algorithm less universal. More examples of filters are shown in the following sections.

51 Several selector-filter pairs are applied sequentially on the same node: if a selector-filter pair does not yield any alignment counterpart nodes, the next pair in the sequence is applied. If none of the processing pairs outputs any counterpart nodes, the node is kept unaligned.

52 In the following, we describe the particular alignment-refining rules which we implemented for English personal and possessive pronouns (Section 5.2.1) and Czech relative pronouns (Section 5.2.2). The reader will probably notice that the rules for aligning Czech relative pronouns seem to be much more complicated than the ones for English personal and possessive pronouns. The complexity of the constructed heuristics was the main factor why we did not continue in building rule-based refining methods for the other classes (e.g., anaphoric zeros) and decided instead to annotate the data manually.

Rule 1: ***Self-Pronoun***
Selector Return all nodes directly aligned to N.
Filter Return nodes that are expressed on the surface as a Czech personal, possessive or reflexive possessive pronoun.

Rule 2: ***Parents-SemRole***
Selector Let P be the parent of N and $R_1, \ldots, R_n$ the Czech nodes aligned to P. Return all children of $R_1, \ldots, R_n$.
Filter Pick the child with the same semantic role that N has.

Rule 3: ***Siblings-SemRole***
Selector Let $S_1, \ldots, S_m$ be the siblings of N. Return a union of the Czech nodes aligned to each S_i.
Filter All input nodes must share the same parent E. Pick one of its children that agrees with N on the semantic role. Note that the true counterpart does not necessarily have to be aligned with any of the S_i in the original alignment. That is the reason why it is looked for among the all children of E.

Rule 4: ***Ancestors-Dative***
Selector Let A be the nearest verbal ancestor of N. Return all children of A.
Filter Pick the node expressed on the surface as a non-possessive pronoun in the dative case.

Algorithm 1. The selector-filter pairs used for refining alignment of English personal and possessive pronouns

5.2.1. Refining alignment for English personal and possessive pronouns

53 The first class addressed with the rule-based refining algorithm is the class of English personal and possessive pronouns in the third person, which corresponds to the class of English central pronouns described in Section 4.1, excluding reflexive pronouns. The main reason for not including reflexives in the rules was that since they are infrequent (see Table 6), manual annotation of the small number of occurrences was less costly than creating the selector and filter rules [10].

54 The alignment refining algorithm itself consists of four selector-filter pairs: *Self-Pronoun*, *Parents-SemRole*, *Siblings-SemRole* and *Ancestors-Dative* (see Algorithm 1). Variable N denotes the node representing the currently processed English pronoun.

55 The selector of the *Self-Pronoun* rule forms a set consisting of exactly the same counterparts as the original alignment would return. However, its filter deliberately reduces the coverage of this rule by excluding all generated and non-pronominal nodes. Moreover, relative and non-possessive reflexive pronouns are excluded because they rarely become a true translation of an English personal pronoun, though often misclassified by GIZA++, as illustrated by the words in bold in examples [3] and [4]:

[3] At night he returns to the condemned building **he** calls home.
Na noc se vrací do opuštěné budovy, **kterou** nazývá domovem.

10. Reflexive pronouns were excluded by discarding central pronoun nodes whose lemma ends with *-self* or *-selves*.

At	night	he	returns	—[himself]	to the	condemned	building	which
Na	noc	∅	vrací	se	do	opuštěné	budovy	**kterou**

he	calls	home.
∅	nazývá	domovem.

[4] These individuals may not necessarily be under investigation when **they** hire lawyers.
Tito jednotlivci nemusí být nutně v době, kdy **si** najímají právníky, ve vyšetřování.

These	individuals	may not	necessarily	be	under	investigation
Tito	jednotlivci	nemusí	nutně	být	ve	vyšetřování

—[at the time]	when	**they**	hire	—[to themselves]	lawyers.
v době	kdy	∅	najímají	**si**	právníky.

56 Observing the data, we found that English central pronouns often occupy the same semantic roles as their Czech counterparts. *Parents-SemRole* and *Siblings-SemRole* processors aim at capturing these counterparts via the pronoun's parent and its siblings, respectively. The technique similar to *Parents-SemRole* was employed in the t-layer projection of the original PCEDT alignment (see Section 5.1).

57 The last rule, *Ancestors-Dative*, attempts to find the cases where the possessive relationship, represented in English by a possessive pronoun, is expressed by a non-possessive pronoun in dative case in Czech. This phenomenon is illustrated in example [5]:

[5] Residents picked **their** way through glass-strewn streets.
Obyvatelé města **si** razili cestu ulicemi zasypanými sklem.

Residents	—[of the city]	picked	—[to themselves]	**their**	way
Obyvatelé	města	razili	**si**	—	cestu

through	glass-strewn	streets.
—	sklem zasypanými	ulicemi.

	No. of instances
Rule 1: *Self-Pronoun*	241
Rule 2: *Parents-SemRole*	190
Rule 3: *Siblings-SemRole*	18
Rule 4: *Ancestors-Dative*	4
Total	**453**

Table 3. Number of English central pronoun instances,
for which the heuristics was able to find the probable Czech counterpart

58 Out of all English central pronouns in the dataset, i.e., 578 instances (see Table 2), this method targeted 549 (95%) which are non-reflexive. For 453 of them, the method was able to find a Czech counterpart. Table 3 illustrates how many instances were covered by each of the rules. It can be seen that the first two rules are responsible for over 95% of the resulting alignments. This does not say anything about the true performance of the algorithm, though. The evaluation can be found in Section 6.

5.2.2. *Refining alignment for Czech relative pronouns*

59 The other class we addressed with the refining heuristics was Czech relative pronouns. We collected the relative pronouns in almost the same manner as described in Section 4.2, the only difference being that here we excluded instances not represented on the t-layer.

60 The alignment refinement was carried out in the following four selector-filter pair rules: *Self-Pronoun*, *Parents-Coref-SemRole*, *Siblings-SemRole* and *Self-Siblings-Apps-EmpVerb* as described in Algorithm 2. The *N* variable again denotes the node whose alignment counterparts are to be found, i.e. an instance of a Czech relative pronoun.

61 Some of the rules may output so-called *indirect counterparts* if the rule fails to find a standard counterpart (denoted as *direct* here). Unlike the direct counterparts, the indirect ones are aligned with a high probability to the antecedent of *N* rather than to *N* itself. Such counterparts can be found only for specific syntactic constructions, e.g., when the relative clause introduced by the Czech relative pronoun is expressed by a simple modifier depending on a noun (as in example [6]) or by a predicative complement or other construction depending on a verb (see example [7]) in English.

[6] To mírně přesáhlo odhad společnosti Sotheby's před aukcí, **který** byl III milionů dolarů.
That was slightly above Sotheby's presale **estimate** of $III million.

That	was above[exceeded]	slightly	Sotheby's	—[company]	presale[before sale]
To	přesáhlo	mírně	Sotheby's	společnosti	před aukcí

estimate	—[which]	—[was]	of	$III million.
odhad	**který**	byl	—	dolarů III milionů.

[7] Libra zaznamenala kurz 1,5920 dolaru, **což** bylo zvýšení z 1,5753 dolaru v úterý večer.
Sterling **was quoted** at $1.5920, up from $1.5753 late Tuesday.

Sterling	**was**	**quoted** at	—[rate]	$1.5920,	—[which]	—[was]
Libra	—	zaznamenala	kurz	dolaru 1,5920	**což**	bylo

up[an increase]	from	$1.5753	late[evening]	—[on]	Tuesday.
zvýšení	z	dolaru 1,5753	večer	v	úterý.

62 The *Self-Pronoun* rule is based on direct links from the original alignment, filtering the collected counterparts to English relative pronouns only. Relative pronouns exist and behave practically the same in both the languages, so GIZA++ is expected to perform well in this case.

Rule 1: ***Self-Pronoun***
Selector Return all nodes directly aligned with N.
Filter Return nodes that are expressed as an English relative pronoun.

Rule 2: ***Parents-Coref-SemRole***
Selector Let P be the parent of N and $R_1, \ldots, R_n$ English nodes aligned to P. Return all children of $R_1, \ldots, R_n$.
Filter Return the node that fulfill any of the following:

1. The node is grammatically coreferential with its grandparent.
2. The node shares the same semantic role with N, if the node is a relative pronoun or its deep lemma is `#Cor` or `#PersPron`.
3. The node is grammatically coreferential. It is returned solely if it is the only such node among the selected nodes.

Rule 3: ***Siblings-SemRole***
Selector Let $S_1, \ldots, S_m$ be the siblings of N. Return a union of the English nodes aligned to each S_i.
Filter All input nodes must share the same parent P, which must be a verb. Pick one of its children that fulfill any of the following as a direct counterpart of N:

1. The node is grammatically coreferential with its grandparent.
2. The node is an English relative pronoun.
3. The node's deep lemma is `#Cor` and there is no other such node among the children of P.

If the sieves fail to find any node or if P is a noun, return P as its indirect counterpart. Likewise, return P if it is a root of apposition.

Rule 4: ***Self-Siblings-Apps-EmpVerb***
Selector Let $S_1, \ldots, S_m$ be the siblings of N. Return a union of the English nodes aligned to N and each S_i.
Filter In contrast to the *Siblings-SemRole* processor, the selected nodes do not need to share the same parent. In case the parent P_i of any of the selected nodes is a root of apposition, return it. If an unexpressed reconstructed verb P_j (its deep lemma is `#EmpVerb`) appears among the selected nodes, denote its parent R. Return R as a direct counterpart if it is a root of apposition and as an indirect counterpart if it is either a verb or noun.

Algorithm 2. The selector-filter pairs used for refining alignment of Czech relative pronouns

	No. of instances	
	Direct counterparts	Indirect counterparts
Rule 1: *Self-Pronoun*	178	–
Rule 2: *Parents-Coref-SemRole*	66	–
Rule 3: *Siblings-SemRole*	14	35
Rule 4: *Self-Siblings-Apps-EmpVerb*	10	3
Total	**268**	**38**

Table 4. Number of Czech relative pronoun instances
for which the heuristics was able to find the probable English direct or indirect counterpart

63 The three remaining rules are more structured and more fine-grained. The selectors collect their candidates via parents as well as siblings, whereas the filters combine information about deep lemmas, grammatical coreference with

indication of English relative pronouns, apposition, and elided verbs reconstructed on the t-layer.

64 Out of the 341 Czech relative pronouns in the dataset (see Table 2), this method focuses only on the 335 instances represented on the t-layer. It succeeded in finding a counterpart in 306 cases (including indirect ones). The contribution of the individual rules is shown in Table 4. The evaluation of the performance of this approach follows in Section 6.

5.3. Manual alignment between Czech and English nodes

65 In the final step, the data were processed manually to obtain as correct alignments as possible. Manual annotation of alignment was carried out only on a subsection of PCEDT (see Section 3.1). The original and heuristically refined alignment served as pre-annotation to speed up the manual work.

66 The alignment links were labeled by two annotators – the authors of this paper. Each instance was annotated only once by one of the annotators, i.e., there is no instance with duplicate annotations.

67 Both direct and indirect alignment were annotated. Furthermore, additional comments were added to the annotation, especially to examples which remained unaligned. There were no strict rules regarding these comments, the annotators were just asked to be consistent in their judgments. Afterwards, these comments were gradually merged in subclasses that we introduce in the analysis of counterparts in Section 7.

68 The alignment was manually annotated for all the classes introduced in Section 4. Although Table 2 shows that the total sum of expressions covered for Czech and English is 2,991, annotating only 2,036 of them sufficed. We took advantage of the fact that many expressions from one language are aligned to the expressions that belong to one of the classes in the other language, i.e., by covering an English expression, we also cover a Czech one, so there is no need to do it again the other way round.

6. Evaluation of the original and rule-based alignment

69 With manual alignment at our disposal, it is possible to evaluate and compare the quality of the original PCEDT alignment of coreferential nodes and its rule-based refinement described in Sections 5.1 and 5.2, respectively. We used the following four metrics for the evaluation:

- Accuracy (A) – the ratio of correctly guessed instances (both positive and negative) to all instances;
- Precision (P) – the ratio of correctly guessed positive instances to all instances predicted as positive;

– Recall (R) – the ratio of correctly guessed positive instances to true positive instances;
– F-score (F) – the harmonic mean of precision and recall: $\frac{2\,PR}{P+R}$

70 Here, a *positive instance* is one that has at least one alignment counterpart, whereas a *negative* one does not have any counterparts. An instance is considered to be *correctly guessed* if at least one predicted alignment counterpart matches one of its true counterparts. For accuracy, instances where both prediction and truth are empty sets are counted as correctly guessed as well.

71 The results, measured on the manually aligned subset of the PCEDT (see Sections 3.1 and 5.3), for both languages are shown in Table 5. The scores for the two classes addressed by our rule-based refinement show that it succeeded in improving over the original PCEDT alignment in terms of all four metrics, especially recall. This is also reflected in the overall numbers, which are better for the rule-based refinement in terms of all metrics, e.g., the average improvement in F-score is 5% points. Interestingly, the refinement algorithm positively affected also the scores on the classes the algorithm did not target. This may happen if a correctly resolved link aligns a node from one of the targeted classes and another node, which does not belong to one of the targeted classes. Since the targeted classes contain a single class for each language, such a result suggests that the alignment links between Czech and English coreferential expressions often cross class boundaries. We will support this hypothesis by detailed statistics of the aligned counterparts in Section 7.

		CS				EN			
		A	P	R	F	A	P	R	F
Central pronouns	orig	88.11	93.80	89.02	91.35	76.47	83.15	80.21	81.65
	rule	89.16	94.26	90.20	92.18	83.74	88.15	88.33	88.24
Relative pronouns	orig	67.16	86.96	66.87	75.60	96.15	96.52	97.00	96.76
	rule	83.87	90.29	84.80	87.46	97.44	98.01	98.50	98.25
Anaphoric zeros	orig	78.71	98.89	71.18	82.78	75.93	98.60	62.58	76.57
	rule	81.76	98.75	75.32	85.46	79.63	99.03	68.37	80.90
Total	**orig**	**77.86**	**94.40**	**73.76**	**82.82**	**79.26**	**90.62**	**76.17**	**82.77**
	rule	**83.68**	**95.16**	**81.02**	**87.52**	**83.95**	**93.55**	**82.20**	**87.51**

Table 5. Evaluation of the original PCEDT alignment (orig) and its rule-based refinement (rule) measured on the manually annotated data set, per class as well as in total

7. Counterparts of the nodes in the other language

72 The following sections present the main results of this work – a detailed study of how the means of expressing coreference change when moving from English to Czech and vice versa [11]. We will go through all the classes introduced in Section 4 and their correspondences in the other language; frequent and interesting cases will be exemplified.

73 Comparing the number of instances covered in Table 2 with the total numbers in Tables 6-11 one can see an occasional discrepancy in the numbers. This arises because the numbers in Tables 6-11 count links, not nodes, and a single node may have multiple counterparts.

7.1. English central pronouns

74 Table 6 shows how frequently English central pronouns, particularly the personal, possessive, and reflexive pronouns, form alignment pairs with Czech nouns, anaphoric zeros, personal, possessive, reflexive possessive, reflexive, or demonstrative pronouns. For cases where the English central pronoun had no Czech counterpart, Table 6 also indicates the reason for its absence: missing Czech possessive pronoun, pleonastic usage of the pronoun *it*, or substantial rewording.

CS / EN	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	rew	
pers	49	190	3		1	21	18	7		29	16	**334**
poss	2	1	94	80	2		6	1	46		4	**236**
refl					3			8				**11**
Total	**51**	**191**	**97**	**80**	**6**	**21**	**24**	**16**	**46**	**29**	**20**	**581**

Table 6. Statistics on the correspondence of English central pronouns to their Czech counterparts. The last three Czech categories indicate the reason why there is no corresponding word in Czech for an English pronoun [12]

75 **Personal pronouns.** As for English personal pronouns, most of them (57%) turn into Czech anaphoric zeros, as in example [8] (99% of these cases occur in the subject position).

[8] **He** left a message accusing Mr. Darman of selling out.
Zanechal mu zprávu, ve které viní Darmana ze zaprodanosti.

11. Note that we still operate on the PCEDT data, i.e., originally English sentences translated to Czech (see Section 3), even if it may appear to be the other way round in some places.

12. The abbreviated names stand for the following: personal (pers), possessive (poss), reflexive (refl), reflexive possessive (refl poss), and demonstrative (demon) pronouns, missing Czech possessive pronoun (no poss), pleonastic usage of the pronoun *it* (pleo), and rewording (rew).

He	left	a	message	—[to him]	∅[in which]	—[he]	accusing[accuses]	
∅	zanechal	—	zprávu	mu	ve které	∅	viní	

Mr.	Darman	of	selling out.
—	Darmana	ze	zaprodanosti.

76 Translations to Czech personal pronouns expressed on the surface account only for 15%. Even though these pronouns are mainly in non-subject positions, still over 35% of them are subjects. These are expressed in Czech mostly either due to their shift away from the subject position or due to topic-focus articulation reasons. Another reason for this is that Czech grammar requires coordinated subject pronouns to be expressed as well.

77 Except for one case, the English personal pronouns aligned with Czech demonstrative pronouns, represented by the pronoun *ten*, are represented by the pronoun *it* (see example [9]).

[9] It endorsed the White House strategy, believing **it** to be the surest way to victory.
Ta přijala strategii Bílého domu v domnění, že je **to** nejjistější cesta k vítězství.

It	endorsed	the	White	House	strategy	believing[in the belief that]
Ta	přijala	—	Bílého	domu	strategii	v domnění, že

it	to	be[is]	the	surest	way	to	victory.
to	—	je	—	nejjistější	cesta	k	vítězství.

78 In Czech, if one refers to a sentence or a longer utterance, the pronoun *ten* is the one most often used. Besides this, the English pronoun *it* occurs also in its pleonastic usage (see example [10]).

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

It	wasn't	known	to	what	extent	—[and]	if	any	the
—	Nebylo	známo,	do	jaké	míry	a	jestli	vůbec	—

facility	was	damaged.
zařízení	bylo	poškozeno.

79 In that case, the pronoun has no counterpart in the Czech sentence. These different means to express the individual functions of the overloaded English pronoun *it* in Czech motivated a cross-lingual approach to disambiguation of *it* (Veselovská et al., 2012), machine translation (Novák et al., 2013a) as well as automatic coreference resolution (Novák & Žabokrtský, 2014).

80 **Possessive pronouns.** Unlike personal pronouns, possessive pronouns often remain in the same class when translated to Czech. In 40% of cases they are translated as possessive pronouns, in almost 35% they become the Czech reflexive possessive *svůj*, a pronoun that shares some features with reflexive pronouns and

substitutes Czech possessive pronouns in some positions when referring to the subject [13]. This category is missing in English, the pronoun *svůj* being translated to English with the possessive pronouns *his*, *her*, *my*, *your* (example [11]).

[11] While the book amply justifies **its** subtitle, the title itself is dubious.
Zatímco **svůj** podtitul kniha dostatečně ospravedlňuje, samotný název je zavádějící.

While	the	book	amply	justifies	**its**	subtitle	the
Zatímco	—	kniha	dostatečně	ospravedlňuje	**svůj**	podtitul	—

title	itself	is	dubious.
název	samotný	je	zavádějící.

81 A substantial proportion of possessive pronouns (20%) disappear in Czech (example [12]).

[12] As a result of **their** illness, they lost $1.8 million in wages and earnings.
Důsledkem nemoci, přišli na mzdách a výdělcích o 1.8 milionu dolarů.

As a result of	**their**	illness,	they	lost	$1.8 million	in
Důsledkem	—	nemoci,	∅	přišli o	dolarů 1.8 milionu	na

wages	and	earnings.
mzdách	a	výdělcích.

82 The relation of possession is then understood intuitively from the context and as in the case of reflexive possessive pronouns, it relates mostly to the subject of the sentence (37 out of the 46 instances). Besides, we found three interesting cases where the benefactor of the predicate and the possessor of the direct object are identical. Then, it is sufficient for a language to express only one of these positions explicitly. For instance, in example [5] (Section 5.2.1), the possessor of the direct object *their* is expressed in English and only the benefactor of the predicate *si* is expressed in Czech, which is exclusively in the dative case.

83 From the point of view of coreference resolution, we can draw an interim conclusion that using personal or reflexive (reflexive possessive) pronouns in Czech increases the probability that the antecedent of the English personal pronoun is a subject, and this fact can be exploited in cross-lingual coreference resolution (Novák & Žabokrtský, 2014) as well as in machine translation.

84 **Reflexive pronouns.** According to Quirk et al. (1985: 356), English reflexive pronouns have two distinct uses: basic and emphatic. Whereas the former functions as object or complement and its antecedent is the subject of the clause, the latter

13. The fact that their antecedent is usually the subject of the same sentence is the main reason why we divide them into a specific subcategory. The rules of use for the reflexive possessive *svůj* in Czech have been addressed in multiple linguistic studies, e.g., by Daneš and Hausenblas (1962), Daneš (1985), and Pit'ha (1992).

is in apposition [14] with its antecedent. The function of the emphatic reflexive is to put special stress on its antecedent. This distinction shows up nicely when moving to Czech: the counterparts of basic reflexives are reflexive pronouns, but emphatic reflexives are expressed by different means in Czech, e.g., by the pronoun *sám* or the adjective *samotný* (lit. *alone*, see example [13]). This fact has been previously exploited in machine translation (Novák et al., 2013b).

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

As	Mr.	Bronner	**himself**	says	the	smell	of "raw meat"	was
Jak	pan	Bronner	**sám**	říká	—	zápach	"syrového masa"	byl

—[smelled]	in	the	air.
cítit	ve	—	vzduchu.

7.2. Czech central pronouns

85 The statistics of Czech central pronouns, namely the personal, possessive, reflexive possessive, and reflexive pronouns and their English counterparts are illustrated in Table 7. The most important counterpart categories are English personal, possessive, and reflexive pronouns, definite article *the*, and anaphoric zeros.

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

Table 7. The statistics on the correspondence of Czech central pronouns to their English counterparts [15]

86 English counterparts of Czech central pronouns are not as diverse as those for English central pronouns. The majority of personal and possessive pronouns remain in the same category and the reflexive possessive *svůj*, which does not exist in English, is, not surprisingly, most often translated as a possessive pronoun (see Section 7.1).

14. This is not annotated as an apposition in PCEDT.

15. The abbreviated names are explained in the note 12 linked to the caption of Table 6.

87 **Personal pronouns.** While translation of personal pronouns to zero is common in the English-to-Czech direction, one expects it to be less frequent in the opposite direction. The collected data support this expectation, as we found only 10% of such cases. A closer look at the individual examples reveals that Czech personal pronouns are realized as zeros in English mostly in the case of infinite clauses, where the argument occupied by the personal pronoun in Czech does not have to (or must not) be expressed in English (see example [14]).

[14] Poslanec Bates prohlásil, že dopisy napíše tak, jak **mu** bylo nařízeno.
Rep. Bates said he would write the letters as ordered.

Rep.[deputy]	Bates	said	—[that]	he	would write	the	letters
Poslanec	Bates	prohlásil	že	∅	napíše	—	dopisy

as[in the way how]	—[it was]	ordered	**∅[to him]**.
tak, jak	bylo	nařízeno	**mu**.

88 **Possessive pronouns.** Czech possessive pronouns mostly translate as English possessives (94 of 107 instances). Among the cases where the translation is different, their co-occurrence with the definite article is especially interesting. Unlike in English, there is no grammatical category of definiteness in Czech. Determination in Czech is expressed by other means, e.g., demonstrative pronouns, intonation, word order, etc. As we can see from our data, in a few instances, the Czech possessive and reflexive possessive pronouns are introduced for this purpose (see example [15]).

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

The[this]	retailer	was unable to	find	a	buyer	for
Tento	maloobchodník	nebyl schopen	najít	—	kupce	pro

the[his]	building.
svoji	budovu.

89 **Reflexive pronouns.** The majority of Czech reflexive pronouns remain unaligned. In 10 out of 14 such cases, the pronoun carries the semantic role of Benefactor or Addressee. In some of these cases, its missing counterpart can be attributed to the phenomenon shown in example [5]. While in example [5], the English possessive pronoun is replaced by a Czech personal or reflexive pronoun in the dative with the semantic role of Benefactor, in example [16], the Czech sentence contains a reflexive pronoun occupying the Benefactor role as well as a reflexive possessive pronoun, both referring to the same entity. Then, having aligned the possessive pronouns together, there is no node left to be aligned to the Czech reflexive pronoun. In such cases, Czech tends to be more pleonastic than English.

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

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

90 Finally, a Czech reflexive can be part of some longer phrase which is translated into English by a completely different expression, e.g., *po sobě (jdoucí)* (lit. *going after one another*) and *proti sobě (jdoucí)* (lit. *going against each other*) to *consecutive* and *contradictory*, respectively (see example [17]).

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

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

Last year's	figure	was down	from	13.4%	in	—[the year]	1987	and	marked
Loňská	hodnota	klesla	z	13.4%	z	roku	1987	a	ukázala

—[that]	in	the	poverty	rate	decline[declined]	the fifth	**consecutive**	
že	—	—	chudoby	míra	klesala	pátý	**po sobě jdoucí**	

annual[year].
rok.

7.3. Czech relative pronouns

91 As for the relative pronouns, we start with the Czech ones since their English counterparts are more diverse. Table 8 gives a picture of how Czech relative pronouns and relative determiners are represented in English. Czech relative pronouns map to the English pronoun *that*, wh-words used in relative clauses, wh-words used in fused relative or interrogative constructions, zeros, roots of appositive constructions, and (rarely) to personal pronouns. Some Czech relative pronouns have no English counterpart: most frequently, relative clauses introduced by Czech relative pronouns are replaced with modifiers of a noun phrase or with verb phrase modifiers.

92 As the anaphoric functions of the Czech relative pronoun *což* differ from other relative pronouns (*což* can refer both to noun phrases and sentences), we cover it separately from the rest.

93 **The relative pronoun *což*.** The expression *což* is a specific relative pronoun frequently used in Czech to refer to a clause or a longer utterance. The wh-words aligned with it are exclusively instances of the pronoun *which*, commonly used as an introducing element of so-called *sentential relative clauses* (Quirk et al., 1985: 1118). However, more often (44% of cases) apposition is used instead, as in example [18] and Figure 2.

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

Table 8. The statistics on the correspondence of Czech relative pronouns to their English counterparts. The last three English categories indicate the reason why there is no corresponding word in English for a Czech pronoun [16]

[18] Akcie včera uzavřely na Neworské burze na 28.75 dolaru, **což** je pokles o 12.5 centu.
The stock closed yesterday on the Big Board at $28.75, down 12.5 cents.

The	stock	closed	yesterday	on	the	Big Board	at	$28.75	
—	akcie	uzavřely	včera	na	—	Neworské burze	na	dolaru	28.75

,[which]	—[is]	down	12.5	cents.	
což	je	pokles	o	12.5	centu.

94 Another way of translating the relative *což* referring to a clause is using a non-finite or verbless clause (Quirk et al., 1985: 992-997), often occupying the role of Effect, Result, or Complement (example [19]).

[19] Společnost Whitbread z Británie dala na prodej svoji divizi lihovin, **čímž** rozpoutala boj mezi lihovary.
Whitbread of Britain **put** its spirits division **up** for sale, setting off a scramble among distillers.

Whitbread	—[company]	of	Britain	**put up**	its	spirits	division	for	
Whitbread	společnost	z	Británie	dala	svoji	lihovin	divizi	na	

sale	—[by which]	setting off[it set off]	a	scramble	among	distillers.	
prodej	**čímž**	rozpoutala	—	boj	mezi	lihovary.	

95 The relative pronoun *což* may also refer to noun phrases. This occurred in two cases in our data (see example [20]), where the relative clause introduced by this pronoun translates as a verbless clause postmodifying a noun phrase.

16. The abbreviated names stand for wh-words used in relative clauses (wh-word relat), wh-words used in fused relative or interrogative constructions (wh-word inter & fused), roots of appositive constructions (appos), personal pronouns (pers), modifiers of a noun phrase (NP modif), and verb phrase modifiers (VP modif).

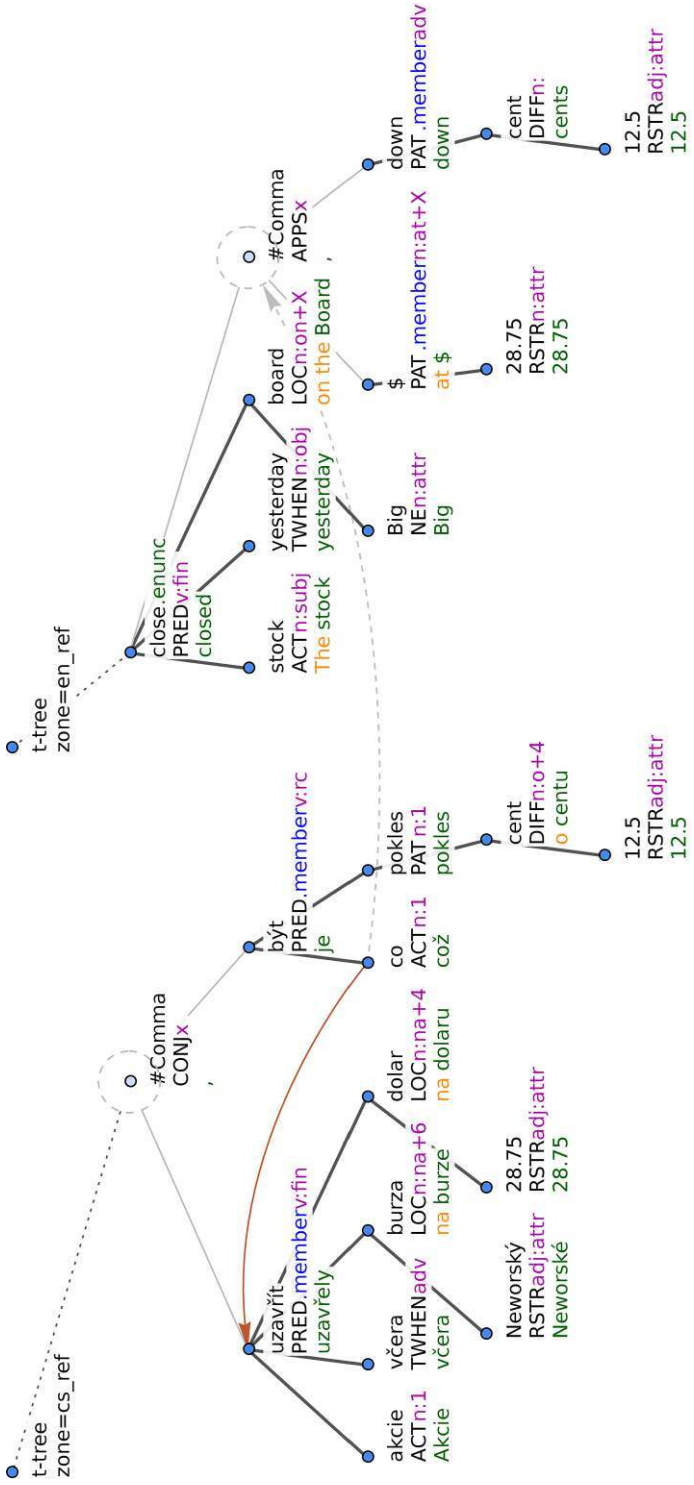

Figure 2. A tectogrammatical representation of the sentence pair from example [18], where Czech *což* turns into an English root of apposition. The alignment is denoted by a dashed arrow. The solid arrow identifies the grammatical coreference

[20] Komise schválila společnosti Tucson zvýšení sazby o 11.5%, **což** je méně, než doporučoval úředník.
The commission authorized an 11.5% rate **increase** at Tucson, lower than recommended by an officer.

The	commission	authorized	an	11.5%	rate	**increase**	at	Tucson
—	komise	schválila	—	o 11.5%	sazby	zvýšení	—	Tucson

—[company]	—[which]	—[is]	lower	than	recommended	by	an	officer.
společnosti	**což**	je	méně	než	doporučoval	—	—	úředník.

96 **Other relative pronouns.** Other Czech relative pronouns are used mainly within *adnominal relative clauses*, i.e., clauses post-modifying a noun phrase. In 50% of cases, the English counterpart is a relative pronoun (see example [21]).

[21] Mohou se objevit síly, **které** tento scénář pozdrží.
There may be forces **that** would delay this scenario.

There	may	be[appear]	forces	**that**	would delay	this	scenario.
—	mohou	se objevit	síly	**které**	pozdrží	tento	scénář.

97 Over 23% of the instances are translated by an anaphoric zero. The reason why this happens is twofold: Czech relative clauses introduced by a pronoun are replaced either with English relative clauses using a zero relative pronoun (example [22]), or with a non-finite clause, specifically with a to-infinitive, *-ing* or *-ed* participles (see example [23]). In both cases, the PCEDT t-layer representation of the subordinate clause contains an anaphoric zero node coreferring with the modified noun.

[22] To je otázka, **na níž** nemůže Východní Německo odpovědět snadno.
That's a question East Germany can't answer easily.

That's	a	question	**∅[which]**	East	Germany	can't	answer
To je	—	otázka	**na níž**	Východní	Německo	nemůže	odpovědět

easily.
snadno.

[23] Zanechal mu zprávu, **ve které** viní Darmana ze zaprodanosti.
He left a message accusing Mr. Darman of selling out.

He	left	a	message	—[to him]	**∅[in which]**	—[he]
∅	zanechal	—	zprávu	mu	**ve které**	∅

accusing[accuses]	Mr.	Darman	of	selling out.
viní	—	Darmana	ze	zaprodanosti.

98 In over 18% of cases, an English counterpart could not be found. In the majority of these cases, the relative clause is transformed into a form not using a verb, thus not having a zero argument on the t-layer that could be aligned with the pronoun.

These forms include premodifiers (adjectives, nouns, participles treated as adjectives) as in example [24], prepositional post-modifiers and post-modifiers using a verbless clause [17] as in example [25].

[24] Dvě zbývající dosud nedosáhly stádia, **kdy** se zjišťují fakta.
The two that remain haven't yet reached the fact-finding **stage**.

The	two	that	remain[remaining]	yet	haven't reached	the	**stage**
—	dvě	—	zbývající	dosud	nedosáhly	—	stádia

—[when]	fact-finding[facts are being found].
kdy	fakta se zjišťují.

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

Imports	—	then	—[was]	six	million	barrels	a day	came
Dovoz	**který**	tehdy	činil	šest	milionů	barelů	denně	přicházel

from	Canada.
z	Kanady.

99 We have not yet mentioned a special subclass of Czech relative pronouns which maps to the English pronouns introducing *interrogative* (see example [26]) and *fused (nominal) relative clauses* (example [27]).

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

It	wasn't	clear	**when**	the	normal	750-car-a-day	pace
—	nebylo	jasné	**kdy**	—	normální	750 vozů za den	tempo

will resume.
se znovu obnoví.

[27] Na tom, **co** máme, je třeba udělat hodně práce.
There is plenty of work to be done on **what** we have.

There	is	plenty	of work	to be	done	on	**what[that, what]**	we	have.
—	je	hodně	práce	třeba	udělat	na	tom, **co**	∅	máme.

100 While the pronoun in the former example does not have any antecedent, the pronoun in the latter is fused with its antecedent. However, it is often very difficult to distinguish which of the two categories a particular occurrence belongs to. All

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

in all, from the computational point of view it is more important to find reliable formal differences between these two categories and the "real" relative pronouns in order to avoid looking for their antecedents in the task of coreference resolution.

7.4. English relative pronouns

101 In Table 9, we show the statistics of English relative pronouns, consisting of the pronoun *that* and wh-words used in adnominal and sentential relative clauses, interrogative and fused clauses, and as a conjunction. Their Czech counterparts have been categorized into four main classes: the Czech relative pronoun *což*, other relative pronouns, conjunctions, and other expressions.

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

Table 9. The statistics on the correspondence of English relative pronouns to their Czech counterparts [18]

102 About 68% of all instances of English relative pronouns can be attributed to alignments between similar categories of true relative pronouns, i.e., the pronoun *that* [19] and relative wh-words on the English side, and the pronoun *což* and other relative pronouns on the Czech side (see example [21]).

103 The majority of wh-words that appear in interrogative or fused relative constructions turn into relative pronouns other than *což* on the Czech side. Over 43% of them are expressed using a so-called *correlative pair*, which in our case consists of a demonstrative pronoun and the following relative pronoun introducing a dependent clause. The antecedent of the relative pronoun is the demonstrative pronoun itself, added to the sentence only for syntactic and stylistic reasons (see example [27]). The 13 occurrences of interrogative or fused pronouns not aligned to a Czech relative pronoun mostly contain the instances of the wh-adverbs *why*

18. The abbreviated names are partly explained in the note 16 linked to the caption of Table 8, the rest stand for wh-words used as conjunctions (wh-words conj), Czech relative pronouns other than *což* (other relat), and conjunctions (conj).

19. One would expect the numbers of English *that* translated to other relative pronouns in Table 9 and of the same case in the opposite direction in Table 8 to be the same. The discrepancy (49 vs. 51 instances) arose due to incorrect part-of-speech tags assigned to two instances of *that*, which prevented the automatic selection method described in Section 4.2 from including these examples.

and *how*. While for English we included them in the class of relative pronouns, their Czech translations *proč* and *jak*, which are never anaphoric in PCEDT, did not meet the specification of the class introduced in Section 4.2.

104 We also spotted 17 occurrences of wh-words, consisting solely of the adverbs *when* and *where* used as a subordinating conjunction (see example [28]). Since this class is irrelevant for the task of coreference resolution, they should be excluded from the set of English relative pronouns. To identify them, we would have to include more syntax-based constraints into the specification of the class presented in Section 4.2. However, the Czech translation can be used to reliably identify wh-words used as conjunctions, as they tend to be translated consistently using the Czech conjunction *když*, which is not ambiguous.

[28] In 1956, **when** Britain, France and Israel invaded Egypt, Arab producers cut off supplies to Europe.

V roce 1956, **když** Británie, Francie a Izrael napadly Egypt, zastavili arabští výrobci dodávky do Evropy.

In	—[the year]	1956	**when**	Britain	France	and	Israel	invaded	Egypt
V	roce	1956	**když**	Británie	Francie	a	Izrael	napadly	Egypt

Arab	producers	cut off	supplies	to	Europe.
arabští	výrobci	zastavili	dodávky	do	Evropy.

105 To sum up, let us recall that Table 2 paints a bleak picture of the precision of the method for selecting coreferential English relative pronouns: 35% of the selected nodes are in fact non-anaphoric. Nonetheless, a deeper investigation summarized in Table 9 discloses that 26% of the nodes labeled as English relative pronouns are in fact wh-words used in interrogative and fused constructions or as a conjunction. Inspecting the non-anaphoric nodes, we found that 72% of them are in fact used in these constructions. The rest might be attributed to some special cases and annotation errors.

7.5. English anaphoric zeros

106 As described in Section 4.3, we decided not to distinguish between different types of anaphoric zeros in this work. Table 10 gives an overview of how English anaphoric zeros map to their Czech counterparts.

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

Table 10. The statistics on the correspondence of English anaphoric zeros to their Czech counterparts [20]

20. The abbreviated names stand for relative (relat) and personal (pers) pronouns.

107 Unsurprisingly, the most frequent aligned counterparts for anaphoric zeros in English are Czech anaphoric zeros. In most cases, missing valency arguments of a verbal predicate are aligned, cf. the unexpressed Actor of the verbs *do* and *ride* in example [29].

[29] Their reaction was to do nothing and ride it out.
Jejich reakcí bylo nedělat nic a nechat to odeznít.

Their	reaction	was	to	**ø.ACT**	do nothing	and	**ø.ACT**	ride it out.
Jejich	reakcí	bylo	—	**ø.ACT**	nedělat nic	a	**ø.ACT**	nechat to odeznít.

108 About 10% of English anaphoric zeros correspond to Czech relative pronouns. These cases represent relative clauses with a zero relative pronoun or non-finite clauses in English (see the description in Section 7.3 and examples [22] and [23]).

109 Almost 50% of anaphoric zeros in English have no Czech counterparts. The most frequent reasons for such an absence are either substantial rewording in the translation, or the absence of corresponding verbal arguments from the t-layer annotation of Czech. Some of these unaligned cases have more or less technical reasons. For example, the verb *chtít* (*want*) is considered to be modal in Czech, so it does not have its own node in the tectogrammatical representation. In English, the verb *want* is represented in the t-layer as a separate node, so its arguments are reconstructed, but cannot have Czech counterparts (see example [30]).

[30] "I want to publish one that succeeds," said Mr. Lang.
"Já chci vydávat takový, který uspěje," řekl Lang.

"I	want	to	**ø.ACT**	publish	one	that	succeeds,"	said	Mr.	Lang.
"Já	chci	—	—	vydávat	takový	který	uspěje,"	řekl	—	Lang.

7.6. Czech anaphoric zeros

110 Table 11 shows a statistic of alignment counterparts for Czech anaphoric zeros.

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

Table 11. The statistics on the correspondence of Czech anaphoric zeros to their English counterparts [21]

21. The abbreviated names stand for personal pronouns in the third (pers), first and second person (pers 1st & 2nd), and possessive pronouns (poss).

111 The cases where Czech zeros correlate to English anaphoric zeros have been exemplified in the previous section. The difference between the two languages as concerns the use of anaphoric zeros is the pro-drop character of Czech, which results in a large number of zeros in subject position. These positions in English are occupied by personal pronouns in the third person (190 cases, see example [8] in Section 7.1) or in the first and second person (40 cases in our data, see example [31]).

[31] Nemáme pasivní čtenáře.
We don't have passive readers.

We	don't have	passive	readers.
∅	nemáme	pasivní	čtenáře.

112 Czech anaphoric zeros are not aligned in about 33% of cases. Similarly as in Section 7.5, the most frequent reasons for that are substantial rewording of the translation or missing arguments in the PCEDT t-layer annotation of English.

8. Discussion

113 The comparison of coreferential pairs in Czech and English has revealed that the alignment counterparts for a single group of coreferential expressions in one language typically come from a wide variety of groups in the other. Some of the counterparts coming from a different group reflect a different use of anaphoric expressions in these two languages (e.g., a Czech demonstrative pronoun *ten* suggests that its English counterpart *it* corefers with a text segment, see Section 7.1), some point out their typological differences. Others reflect different vocabulary and semantics of words (e.g., the emphatic use of English reflexives, see Section 7.1), and some cases indicate different syntactic tendencies (e.g., more frequent usage of non-finite constructions in English than in Czech, see Section 7.3). There are also many cases of rewording or just occasional changes of anaphoric expressions, which could be theoretically interesting for a linguistic investigation but the number of cases was so small that it was not possible to verify our hypotheses. In this work, we have pointed out and exemplified only a few types of coreferential pairs in Czech and English, but still they open many theoretical questions, far more than we are able to address here.

114 One of the most interesting points is addressed in Section 7.1 and concerns the expression of possessivity in English and Czech. The statistic on the correspondence of English possessive pronouns to their Czech counterparts confirms the general tendency of Czech to express personal possessives less frequently than English. Indeed, in Czech, it is not common to use a possessive (or a reflexive possessive) pronoun in sentences like example [12]. However, it is not ungrammatical. The Czech sentence in example [12] would remain grammatically correct after adding a reflexive possessive (see example [12′]).

[12′] As a result of **their** illness, they lost $1.8 million in wages and earnings.
Důsledkem **své** nemoci, přišli na mzdách a výdělcích o 1.8 milionu dolarů.

As a result of	**their**	illness	they	lost	$1.8 million	in	wages
Důsledkem	**své**	nemoci	∅	přišli o	dolarů 1.8 milionu	na	mzdách

and	earnings.
a	výdělcích.

115 The high frequency of possessives in English is connected with the grammatical category of definiteness. English has a strong tendency to avoid using bare nouns, i.e., noun phrases (especially in the singular) should be mostly specified by either an article or other means of determination. Possessive pronouns in cases such as *their* in example [12] express determination even more explicitly than the definite article, giving a monosemantic reference to the possessor. Czech does not have such a strong tendency to express determination. On the other hand, it has a means of expressing it that is unknown to English – the Dative possessor [22], which occurs in our examples parallel to English possessive pronouns, cf. example [5] in Section 5.2.1.

116 The collected statistics of correspondences also give us valuable information that can be exploited within the task of automatic coreference resolution and its subtask of anaphor detection on parallel texts. As mentioned in Section 7.1, Czech texts may provide several hints about the coreference of English central pronouns, e.g.:

- the pleonastic usage of the pronoun *it* is indicated by no counterpart in Czech;
- the pronoun *it* referring to a larger segment is usually translated as the demonstrative pronoun *ten*;
- a reflexive possessive or no Czech counterpart indicates that the antecedent of the English pronoun is probably the subject of the sentence.

117 Another fact that can be exploited is that in both the languages, the gender of the pronoun must agree with the gender of its antecedent and the distribution of genders over nouns differs across these languages. While in English, most nouns are referred to by a pronoun in the neuter gender, Czech genders are distributed more evenly. These differences in Czech and English central pronouns were already taken into account in previous cross-lingual coreference resolution experiments by Novák and Žabokrtský (2014).

118 Concerning English relative pronouns, Table 2 shows that the precision of our selection method (see Section 4.2) is much lower for this class than for the others. However, the analysis in Section 7.4 shows that their correspondence with a Czech correlative pair or the non-ambiguous conjunction *když* can be used to reliably indicate wh-words which are not used as relative pronouns.

22. See, e.g., Payne and Barshi (1999) and Křivan (2007).

119 The correspondence of Czech anaphoric zeros in the subject position and English personal pronouns demonstrated in Section 7.6 illustrates the pro-drop nature of Czech and suggests that the English pronouns can be used to facilitate identification of the places where to reconstruct a Czech zero. Bojar et al. (2012) reported that 25% of all Czech pronouns unexpressed on the surface are reconstructed incorrectly or not at all, which substantially contributes to a 27 percentage point decrease in F-score of coreference resolution if gold linguistic annotation is replaced by an automatic one. English personal pronouns can also help the disambiguation by providing additional information on gender in cases where a verb governing the Czech anaphoric zero is in the present tense, having the same form in any gender.

120 Although coreference resolution of Czech relative pronouns is not as difficult task as the resolution of personal pronouns and anaphoric zeros, we believe it can be slightly improved if the information from its English counterparts as presented in Section 7.3 is taken into account (especially those counterparts which are not relative pronouns, e.g., *-ing* or *-ed* participles, or noun modifiers).

121 One more important consideration for the interpretation of our results is that the collected statistics are influenced by the translation direction since all our English texts are originals and the Czech texts are translations from English. We expect that if the original texts were in Czech, we would see, e.g., fewer nominalizations, non-finite clauses, and appositions in English. It is also important to mention that our results should be understood as valid only for the particular domain represented in the PCEDT, namely English journalistic texts and their translations to Czech. This holds mostly for the differences between original and translated texts but it can also concern the properties of anaphoric expressions that we have identified.

9. Conclusion

122 This work presents a comprehensive study on how certain classes of expressions used to establish coreferential relations are represented in English and Czech and what the most frequent mappings between them are. The study was carried out on the parallel data of the Prague Czech-English Dependency Treebank, focusing on central pronouns, relative pronouns, and anaphoric zeros. We formally defined these classes in order to capture the coreferential expressions in PCEDT with very high recall and sufficient precision.

123 To obtain a reliable word alignment between coreferential expressions for our studies, we designed a rule-based alignment refining algorithm that improves the quality of the original PCEDT word alignment links not only for the classes it aims at, but in general. Starting from the improved automatic alignment, we manually annotated word alignment on a subset of the PCEDT data.

124 Our study of the aligned coreferential expression pairs has confirmed many theoretical assumptions on, e.g., a different frequency of possessives in Czech and English, dropping the subject pronoun when moving from English to Czech, or

English nominalization of a Czech relative pronoun. Furthermore, we found that the aligned Czech relative pronoun can be reliably used to determine whether the English pronoun refers to an entity or a text segment. We also discovered a high diversity in the translations of reflexive pronouns in both directions. All the findings can be also applied in feature engineering for cross-lingual coreference resolution on parallel texts, which was the central motivation of this study.

125 In our future work, we plan to concentrate on how to improve the precision of selecting the coreferential nodes, especially for the class of English relative pronouns, which contained many instances in fused and interrogative constructions. We will also apply the results of this study in improving automatic coreference resolution. Our goal is to combine improved alignment techniques (either by using the presented rule-based aligner or by exploiting the manually aligned dataset in a supervised machine learning approach) and the observed correspondences to build a coreference resolution system that takes advantage of the cross-lingual information. Such a system can then be applied to a much larger bilingual dataset in the hope that it performs better than two separate monolingual systems. The system annotations of coreference obtained in this way can be subsequently used to enrich manual annotation in a semi-supervised manner, providing more training data for monolingual systems in each of the two languages.

Acknowledgments

126 We gratefully acknowledge support from the Grant Agency of the Czech Republic (grant P406/12/0658 "Coreference, discourse relations and information structure in a contrastive perspective"), the Foundation of Vilem Mathesius, GAUK 3389/2015, EU (grant FP7-ICT-2013-10-610516 – QTLeap) and SVV project number 260 224. This work has used language resources developed, stored, and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013). The authors also thank prof. Eva Hajičová, assoc. prof. Zdeněk Žabokrtský, Ondřej Dušek and three anonymous reviewers for their valuable comments and suggestions to improve the paper.

References

Baker, M. 1995. Corpora in Translation Studies: An Overview and Suggestions for Future Research. *Target* 7 (2): 223-243.

Baroni, M. & Bernardini, S. 2006. A New Approach to the Study of Translationese: Machine-Learning the Difference between Original and Translated Text. *Literary and Linguistic Computing* 21 (3): 259-274.

Blum, A. & Mitchell, T. 1998. Combining Labeled and Unlabeled Data with Co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*. New York: Association for Computer Machinery: 92-100.

Bojar, O. et al. 2012. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*. Stroudsburg: Association for Computational Linguistics: 3921-3928. Available online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/645_Paper.pdf.

Brown, P.F. et al. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19 (2): 263-311.

Daneš, F. 1985. Zwei Anmerkungen zu den Personalpronomen. *Zeitschrift für Slawistik* 30: 53-60.

Daneš, F. & Hausenblas, K. 1962. Privlastňovací zájmena osobní a zvratná ve spisovné češtině. *Slavica Pragensia* 4: 191-202.

Givón, T. (ed.) 1983. *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. Typological studies in language 3. Amsterdam: J. Benjamins.

Guillou, L. 2012. Improving Pronoun Translation for Statistical Machine Translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 1-10. Available online: http://aclweb.org/anthology-new/E/E12/E12-3001.pdf.

Guillou, L. et al. 2014. ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*. Stroudsburg: Association for Computational Linguistics: 3191-3198. Available online: http://www.lrec-conf.org/proceedings/lrec2014/pdf/298_Paper.pdf.

Hajič, J. et al. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*. Stroudsburg: Association for Computational Linguistics: 3153-3160. Available online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/510_Paper.pdf.

Hardmeier, C. & Federico, M. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT)*. 283-289. Available online: http://uu.diva-portal.org/smash/get/diva2:420761/FULLTEXT01.

Hermjakob, U. 2009. Improved Word Alignment with Statistics and Linguistic Heuristics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics: 229-237. Available online: https://aclweb.org/anthology/D/D09/D09-1024.pdf.

Kibrik, A.A. 2011. *Reference in Discourse*. Oxford – New York: Oxford University Press.

Křivan, J. 2007. *Externí posesivita v češtině: v typologické a areální perspektivě*. Master's thesis. Charles University in Prague, Faculty of Arts, Prague.

Kunz, K.A. 2010. *Variation in English and German Nominal Coreference: A Study of Political Essays*. Berlin – Bern: P. Lang.

Kunz, K.A. & Lapshinova-Koltunski, E. 2015. Cross-Linguistic Analysis of Discourse Variation Across Registers. *Nordic Journal of English Studies* 14 (1): 258-288.

Le Nagard, R. & Koehn, P. 2010. Aiding Pronoun Translation with Co-reference Resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*. Stroudsburg: Association for Computational Linguistics: 252-261. Available online: http://aclweb.org/anthology-new/W/W10/W10-1737.pdf.

Marcus, M. et al. 1999. *Treebank-3 LDC99T42*. Philadelphia: Linguistic Data Consortium. Available online: https://catalog.ldc.upenn.edu/LDC99T42.

Mikulová, M. et al. 2006. *Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation Manual*. Technical report 2006/30. Prague: ÚFAL MFF UK. 1287 p.

Nedoluzhko, A. et al. 2014. *Annotation of Coreference in Prague Czech-English Dependency Treebank*. Technical report 2014/57. Prague: ÚFAL MFF UK. 41 p.

Nedoluzhko, A., Toldova, S. & Novák, M. 2015. Coreference Chains in Czech, English and Russian: Preliminary Findings. *Computational Linguistics and Intellectual Technologies* 14: 456-469.

Novák, M. & Žabokrtský, Z. 2014. Cross-Lingual Coreference Resolution of Pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Stroudsburg: Association for Computational Linguistics: 14-24. Available online: http://www.aclweb.org/anthology/C/C14/C14-1003.pdf.

Novák, M., Nedoluzhko, A. & Žabokrtský, Z. 2013a. Translation of "It" in a Deep Syntax Framework. In *Proceedings of the Workshop on Discourse in Machine Translation*. Stroudsburg: Association for Computational Linguistics: 51-59. Available online: http://www.aclweb.org/anthology/W/W13/W13-3307.pdf.

Novák, M., Žabokrtský, Z. & Nedoluzhko, A. 2013b. Two Case Studies on Translating Pronouns in a Deep Syntax Framework. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*. Stroudsburg: Association for Computational Linguistics: 1037-1041. Available online: http://www.aclweb.org/anthology/I/I13/I13-1142.pdf.

Och, F.J. & Ney, H. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics. Available online: http://www.aclweb.org/anthology/P/P00/P00-1056.pdf.

Onderková, K. 2009. *Possessive Pronouns in English and Czech Works of Fiction. Their Use with Parts of Human Body and Translation*. Master's thesis. Masaryk University, Faculty of Arts, Brno.

Panevová, J. et al. 2014. *Mluvnice současné češtiny 2. Syntax na základě anotovaného korpusu*. Prague: Karolinum. Vol. 2.

Payne, D.L. & Barshi, I. (eds.) 1999. *External Possession*. Typological studies in language 39. Amsterdam – Philadelphia: J. Benjamins.

Piťha, P. 1992. *Posesivní vztah v češtině*. Prague: AVED.

Postolache, O., Cristea, D. & Orăsan, C. 2006. Transferring Coreference Chains through Word Alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*. Stroudsburg: Association for Computational Linguistics: 889-892. Available online: http://www.lrec-conf.org/proceedings/lrec2006/pdf/224_pdf.pdf.

Quirk, R. et al. 1985. *A Comprehensive Grammar of the English Language*. London – New York: Longman.

Sgall, P. 1967. *Generativní popis jazyka a česká deklinace*. Prague: Academia.

Sgall, P., Hajičová, E. & Panevová, J. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: D. Reidel.

Souza, J.G.C. (de) & Orăsan, C. 2011. Can Projected Chains in Parallel Corpora Help Coreference Resolution? In I. Hendrickx et al. (eds.), *Anaphora Processing and Applications (8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2011, Faro, Portugal, October 6-7, 2011)*. Berlin – Heidelberg: Springer: 59-69.

Veselovská, K., Nguy, G.L. & Novák, M. 2012. Using Czech-English Parallel Corpora in Automatic Identification of "It". In *The 5th Workshop on Building and Using Comparable Corpora*. Allschwil: European Association for Machine Translation: 112-120. Available online: http://www.mt-archive.info/10/BUCC-2012-Veselovska.pdf.

Žabokrtský, Z., Ptáček, J. & Pajas, P. 2008. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 167-170. Available online: http://www.aclweb.org/anthology/W/W08/W08-0325.pdf.

Zhang, J. & Zhao, H. 2013. Improving Function Word Alignment with Frequency and Syntactic Information. In F. Rossi (ed.), *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*. Palo Alto: AAAI Press: 2211-2217. Available online: http://ijcai.org/papers13/Papers/IJCAI13-326.pdf.

Zinsmeister, H., Dipper, S. & Seiss, M. 2012. Abstract Pronominal Anaphors and Label Nouns in German and English: Selected Case Studies and Quantitative Investigations. *Translation: Computation, Corpora, Cognition* 2 (1): 47-80.