A Fine-grained Large-scale Analysis of Coreference Projection

Michal Novák
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
{mnovak}@ufal.mff.cuni.cz

Abstract

We perform a fine-grained large-scale analysis of coreference projection. By projecting gold coreference from Czech to English and vice versa on Prague Czech-English Dependency Treebank 2.0 Coref, we set an upper bound of a proposed projection approach for these two languages. We undertake a detailed thorough analysis that combines the analysis of projection's subtasks with analysis of performance on individual mention types. The findings are accompanied with examples from the corpus.

1 Introduction

Projection has been for a long time seen as an alternative way to build a linguistic tool for resource-poor languages. Coreference projection has been no exception. Despite its mostly mediocre results, only some works perform a proper error analysis.

In this work, we conduct a fine-grained large-scale analysis of coreference projection. We adopt a corpus-based projection approach and apply it on Czech-English parallel texts in Prague Czech-English Dependency Treebank 2.0 Coref (Nedoluzhko et al., 2016) in both projection directions. We project manually annotated coreference links within texts enriched with mainly manually annotated linguistic annotation. Results obtained on manual (i.e. gold) annotation can be then considered as an upper bound for projection techniques where the gold annotation is replaced by the one obtained automatically.

We took inspiration from two works that have previously focused on projection of gold coreference. Even though both of them provided an analysis of collected projections, they treated it in a completely different way. Postolache et al. (2006) concentrated on factorized analysis. They split the task of projection into subtasks, such as mention matching, mention span overlapping and antecedent selection, and inspected their effect on the final result separately. Alternatively, in their multilingual projection approach Grishina and Stede (2017) carried out an analysis across mention types. They split all mentions to categories such as noun phrases, named entities and pronouns, and evaluated projection on these mention types separately.

Our work combines both these views on analysis, providing a factorized fine-grained analysis of projected coreference. In addition, we include new categories of mentions – zeros. These have been often neglected as they are not expressed on the surface. However, by ignoring them we would lose valuable information, especially in pro-drop language such as Czech and Spanish. Furthermore, our analysis is based on about 100-times bigger corpus than in the two related works, which makes the findings and conclusions more reliable.

The paper is structured as follows. In Section 2 we describe two main projection approaches with a special emphasis on corpus-based projection of gold coreference. Section 3 presents the corpus that we use for making projections and Section 4 describes the projection method we propose. The main projection experiments and its results are presented in Section 5 and analyzed in detail using a factorized view in Section 6. Finally, we conclude in Section 7.

2 Related Work

Approaches to cross-lingual projection are usually aimed to bridge the gap of missing resources in the target language. So far, they have been quite successfully applied to part-of-speech tagging (Täckström et al., 2013), syntactic parsing (Hwa et al., 2005), semantic role labeling (Padó and Lapata, 2009), opinion mining (Almeida et al.,

2015), etc. Projection techniques are generally grouped into two types with respect to how they obtain the translation to the source language, which is usually a resource-rich language. *MT-based approaches* apply a machine-translation service to create synthetic data in source language. *Corpus-based approaches* take advantage of the human-translated parallel corpus of the two languages.

MT-based approaches. The workflow of these approaches is as follows. Starting with a text in the target language to be labeled with coreference, it first must be machine-translated to the source language. A coreference resolver for the source language is then applied on the translated text and, finally, the newly established coreference links are projected back to the target language. Flexibility of this approach lies in the fact that it can be applied in both train and test time, and no linguistic processing tools for the target language are required. To the best of our knowledge, this approach has been applied to coreference only twice, by Rahman and Ng (2012) on projection from English to Spanish and Italian, and by Ogrodniczuk (2013) on projection from English to Polish.

Corpus-based approaches. In these approaches, a human-translated parallel corpus of the two languages is available and the projection mechanism is applied within this corpus. Coreference annotation in the source-language side of the corpus may be both labeled by humans or a coreference system. The target-language side of the corpus then serves as a training dataset for a coreference resolver. This approach thus must be applied in train time and, moreover, it requires a coreference resolver trainable on the target-language data. As a consequence, linguistic processing tools should be available for the target language as most of the resolvers depend on some amount of additional linguistic information. On the other hand, human translation and gold coreference annotation, if available, should increase the quality of the projected coreference. This approach has been used to create a coreference resolver by multiple authors, e.g. de Souza and Orăsan (2011), Martins (2015), Wallin and Nugues (2017), and Novák et al. (2017). However, since the present work employs the corpus-based approach on gold annotations of coreference, we offer more details on works of Postolache et al.

(2006) and Grishina and Stede (2015, 2017).

Postolache et al. (2006) followed corpus-based approach using a small English-Romanian corpus of 638 sentence pairs in order to create a bilingually-annotated resource. They projected manually annotated coreference, which was then post-processed by linguists to acquire high quality annotation in Romanian. Based on the gold coreference annotation of the Romanian side of the corpus, they evaluated the F-scores of mention heads' matching, mention spans' overlapping, and coreference clusters on all as well as only correctly projected mentions. A factorized error analysis they carried out shows that the majority of errors in coreference projection stems from a lower recall (around 70%) caused by missing alignment due to alignment errors or language differences introduced in the translation.

Yulia Grishina with her colleagues also investigate possibilities of corpus-based coreference projection. In (Grishina and Stede, 2015), they introduced a "generalizable" annotation schema that they tested on parallel texts of three languages (English, Russian and German) and three genres (newswire articles, short stories, medical leaflets). Using this dataset consisting of less than 500 sentence triples, they conducted experiments on projection from English to the two other languages. In (Grishina and Stede, 2017), they pursue a goal of multi-source projection of manual coreference annotation. They propose several strategies of combining projections from multiple languages, with some of them slightly improving the F-score of the best-performing projection source. They also provide a qualitative analysis on individual mention types suggesting that pronouns have much higher projection accuracy[1] than nominal groups. They justify their unsatisfactory results especially for German nominal groups by problems with inclusion of an unaligned German determiner in definite descriptions.

3 Data Source

We employ a slightly modified version of the Prague Czech-English Dependency Treebank 2.0 Coref (Nedoluzhko et al., 2016, PCEDT 2.0

---

[1]As far as we are concerned, it is misleading to call their projection measure "accuracy". There is another measure that could be calculated as a proportion of target mentions covered by projection among all target mentions, i.e. "recall". Therefore, whenever we apply this measure in our experiments in Section 4, we rather denote it as "precision".

Coref) for our projection experiments.

PCEDT 2.0 Coref is a coreferential extension to the Prague Czech-English Dependency Treebank 2.0 (Hajič et al., 2012). It is a Czech-English parallel corpus, consisting of almost 50k sentence pairs (more on its basic statistics is shown in the upper part of the Table 1). The English part originally comes from the Wall Street Journal collected in the Penn Treebank (Marcus et al., 1999) and the Czech part was manually translated. It has been annotated at multiple layers of linguistic representation up to the layer of deep syntax (or *tectogrammatical layer*), based on the theory of Functional Generative Description (Sgall et al., 1986). The tectogrammatical representation of a sentence is a dependency tree with semantic labeling, coreference, and argument structure description based on a valency lexicon. The nodes of a tectogrammatical tree comprise merely auto-semantic words. Furthermore, some surface-elided expressions are reconstructed at this layer. They include anaphoric zeros (e.g. zero subjects in Czech, unexpressed arguments of non-finite clauses in both English and Czech) that are introduced in the tectogrammatical layer with a newly established node.

The coreference annotation of PCEDT 2.0 Coref takes place on the tectogrammatical layer to allow for marking zero anaphora. Coreference is technically annotated as links connecting two mentions: *the anaphor* (the referring expression) and *the antecedent* (the referred expression). The coreference links then form *chains*, which correspond to coreference entities. In tectogrammatics, the mention is determined only by its head. No mention boundaries are specified. Therefore, a coreference link always connects two nodes on a tectogrammatical layer.

In order to provide a fine-grained qualitative analysis, we divide mentions into multiple categories in this paper: (1) personal pronouns, (2) possessive pronouns, (3) reflexive possessive pronouns, (4) reflexive pronouns, all four types of pronouns in the 3rd or ambiguous person, (5) demonstrative pronouns, (6) zero subjects, (7) zeros in non-finite clauses, (8) relative pronouns, (9) the pronouns of types (1)-(4) in the 1st or 2nd person, (10) named entities, (11) common nominal groups, and (12) other expressions. Note that categories (3) and (6) are defined only in Czech. The last category contains coordination roots, verbs, adjectives, but the majority is formed by other

Mention type	Czech	English
Sentences	49,208	49,208
Tokens	1,151,150	1,173,766
Tecto. nodes	931,846	838,212
Mentions (total)	183,277	188,685
Personal pron.	3,038	14,887
Possessive pron.	3,777	9,186
Refl. poss. pron.	4,389	—
Reflexive pron.	1,272	484
Demonstr. pron.	3,429	1,492
Zero subject	16,875	—
Zero in nonfin. cl.	6,151	29,759
Relative pron.	15,198	8,170
1st/2nd pers. pron.	4,415	4,557
Named entities	18,874	36,833
Nominal group	80,124	68,866
Other	25,735	14,451

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

nodes restoring ellipsis, e.g. zeros in other than subject positions or missing arguments in reciprocal relation. We do not focus on this category in the rest of the paper. The statistics of coreferential mentions is collected in the bottom part of Table 1.

The treebank is aligned on the level of tectogrammatical nodes. The alignment is based on unsupervised word alignment by GIZA++ (Och and Ney, 2000), augmented with a supervised method (Novák and Žabokrtský, 2014) for selected coreferential expressions. The supervised alignment has been trained on a section of PCEDT 2.0 Coref comprising 1,078 sentence pairs with manual annotation of alignment. To ensure that the whole PCEDT 2.0 Coref is aligned in the same way for our experiments, we make a slight modification to it and replace the manual alignment in this particular section with the supervised one, obtained by 10-fold cross-validation.

4 Coreference Projection

Our approach to coreference projection belongs to the corpus-based methods as introduced in Section 2. We work with manually translated English-Czech parallel corpus with word alignment and project coreference from one language side to the other. In fact, our approach is similar to the one

adopted by multiple previous works (Postolache et al., 2006; de Souza and Orăsan, 2011; Wallin and Nugues, 2017; Grishina, 2017, i.a.). Nevertheless, there is a substantial difference of our work compared to the others: our projection system operates on tectogrammatical representation. It leads to the two following consequences.

Firstly, our system is able to address zero anaphora. Thorough cross-lingual analysis by Novák and Nedoluzhko (2015) showed that many counterparts of Czech or English coreferential expressions are zeros. This likely holds for the other pro-drop languages, too. It is thus surprising that the previous work on projection to Spanish (Rahman and Ng, 2012; Martins, 2015) or Portuguese (de Souza and Orăsan, 2011; Martins, 2015) did not accent this problem at all. In tectogrammatics, generated nodes serve this purpose instead.

Secondly, mention spans are not specified in tectogrammatical trees, as mentioned in Section 3. Concerning projection, many of the previous works (Rahman and Ng, 2012; Postolache et al., 2006; Wallin and Nugues, 2017, i.a.) devote considerable space to answering the question of the proper strategy for determining boundaries of a projected mention. If a mention is solely defined by its head as in the present work, this question does not need to be answered.

The projection algorithm is schematized in Algorithm 1. An input of the algorithm are two aligned lists of tectogrammatical trees representing the same text in the source and the target language. First, a list of coreferential chains must be extracted from source trees (line 1). Every coreference chain is projected independently (lines 2-18) mention by mention, starting with the first one, the one that has no outcoming link. For each mention, at the moment viewed as an anaphor, its counterpart in the target language is returned using the alignment (line 5). In case there are several nodes aligned to the anaphor, those which do not yet participate in a different chain are interlinked and only the very last mention is returned by the function `GetAlignedAndInterlink`. If no aligned counterpart to the anaphor is found, the anaphor is skipped and its outgoing coreference link thus remains unprojected. Otherwise (lines 6-16), counterparts of anaphor's direct antecedents are retrieved (lines 7-8) and the algorithm adds a link between the anaphor's and antecedents' counterparts in the target language (line 10). If there

are no antecedents' counterparts, the last successfully projected anaphor from any of the previous iterations is used instead (line 13).

5 Experiments and Results

In the following experiment, we project gold coreference between gold trees in two directions: from English to Czech and vice versa. The experiment is carried out on the dataset presented in Section 3, PCEDT 2.0 Coref with supervised alignment in all its sections.

One of the objectives of this work is to analyze performance of coreference projection for individual mention types. Standard evaluation metrics (e.g. MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998)) are not suitable for our purposes, though, since they do not allow for scoring only a subset of mentions. Instead, we use a measure similar to scores proposed by (Tuggener, 2014) that we denote as *anaphora score*.

Let $K = \{K_1, \ldots, K_m\}$ be the set of true and $S = \{S_1, \ldots, S_n\}$ the set of predicted coreferential chains. From K and S we derive sets of true anaphors $\text{ANAPH}(K)$ and predicted anaphors $\text{ANAPH}(S)$ using the following definition:

$$\text{ANAPH}(Z) = \{x | \exists i : x \in Z_i \text{ and } \exists y : y \in \text{ANTE}_{Z_i}(x)\}$$

where the set $\text{ANTE}_{Z_i}(x)$ contains all direct antecedents of x in chain Z_i. We also define an indicator function $both(a, K, S)$ as follows:

$$both(a, K, S) = \begin{cases} 1 & \begin{aligned} &\text{if } \exists i, j : a \in K_i \cap S_j \\ &\text{and } \exists e \in K_i : e \in \text{ANTE}_{K_i}(a) \\ &\text{and } \exists f \in \text{ANTE}_{S_j}(a) : f \in K_i \end{aligned} \\ 0 & \text{otherwise} \end{cases}$$

In other words, it fires only if a has an antecedent in both the truth and the prediction and the predicted antecedent of anaphor a belongs to a true coreferential chain associated with a. Precision (P), Recall (R) are then computed by averaging the function $both(a, K, S)$ over all predicted and true anaphors, respectively, and F-score (F) traditionally as a harmonic mean of P and R:

$$P = \frac{\sum\limits_{a \in \text{ANAPH}(S)} both(a, K, S)}{|\text{ANAPH}(S)|} \quad R = \frac{\sum\limits_{a \in \text{ANAPH}(K)} both(a, K, S)}{|\text{ANAPH}(K)|}$$

$$F = \frac{2PR}{P + R}$$

To evaluate only a particular anaphor type, both sets $\text{ANAPH}(K)$ and $\text{ANAPH}(S)$ must be restricted only to anaphoric mentions of the given

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

Algorithm 1: Algorithm for coreference projection

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

Table 2: Anaphora scores of gold coreference projected on PCEDT 2.0 Coref with thorough supervised alignment.

type. In the following tables, we use ^P_R F to format the three components of the anaphora score.

Table 2 shows results of gold coreference projection. The main observation is that with the overall F-scores around 50%, coreference projection between EN and CS seems to be a difficult problem. Moreover, let us emphasize that this experiment is supposed to set an upper bound for our projection approach since most of the annotation it exploits is manual. Comparing the two directions, the CS→EN projection appears to be a bit easier, yet still not reaching 60% F-score. Although precision rates are rather low, it is even lower recall rates that seem to have a more important effect on the weak performance of projection.

Note that our absolute projection scores are not easy to be directly compared with the numbers reported in other works performing projection of gold coreference (e.g. by Postolache et al. (2006) and Grishina and Stede (2017)). There are several factors affecting the score values, in which these experiments certainly differ: a target language, a range of expressions annotated with coreference, quality of alignment, evaluation measure, etc.

To slightly facilitate comparison, we can judge relative performance on individual mention types. In both languages, coreference information is obviously best preserved for central pronouns (ex-

cept for basic reflexives). It agrees with findings by Grishina and Stede (2017), where they observed higher precision for pronouns than for nominal groups. They suggest that inferior performance for nominal groups may be a result of errors in mention matching. To find out if our results can be justified in this way, we undergo a detailed analysis of factors influencing the projection score.

6 Analysis of Factors

There are three main factors that contribute to the quality of coreference projection: quality of (1) alignment, (2) mention matching, and (3) antecedent selection. Every projection error can be associated with a factor that caused it. Table 3 shows the results of the analysis of factors that are elaborated in more details in the following paragraphs.

Proportion of aligned mentions. No coreference link can be projected to an unaligned mention. Missing alignment on target-language mentions thus causes errors of the first type. The left-hand side of Table 3 shows the proportion of aligned target-language mentions. Extremely low proportion of aligned mentions is observed for Czech basic reflexive pronouns. In the vast majority of cases, unaligned Czech basic reflexives are a result of not expressing the corresponding argument of the proposition in English. For instance, the Czech translation of the verb *to rent* in Example 1 requires explicit reflexive pronoun to signal the meaning that Exxon will pay for using the tower, not that Exxon will receive money as its owner.

(1) *Exxon **si** pronajme část výškové budovy.*
Exxon **[to it]** will rent part of a tower.
Do dokončení stavby si společnost Exxon pronajme část stávající kancelářské výškové budovy.
Until the building is completed, Exxon will rent part of an existing office tower.

Surprisingly, Czech personal pronouns are also less frequently aligned than the other mention types. Similarly to the previous case, the reason is often that some arguments of the English proposition are not explicitly mentioned (see Example 2). In general, missing English counterparts are a result of compact formulation of English sentences, like in Example 5. Compact language is, in our view, an inherent property of English as well as a feature of the specific journalistic style used in

Wall Street Journal (WSJ). Moreover, one should not neglect the factor of the so-called *Explicitation Hypothesis* as formulated by Blum-Kulka (1986): the redundancy expressed by a rise of cohesive explicitness in the target-language text might be caused by the nature of the translation process itself.

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

As for English, we can see lower scores for zeros in non-finite clauses and reflexive pronouns, again. The non-finite clauses mainly consist of past and present participles. All the missing Czech counterparts of zeros in the past participle are due to the participle being represented as an adjective in Czech, thus having no valency arguments annotated. The reasons behind a missing Czech counterpart of a zero in the present participle are more diverse. The counterpart is often missing even for the governing verb, not just for its zero argument (see Example 3). As opposed to the previous case of explicitation, this is an example of implicitation in the EN→CS translation.

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

Missing alignment for English reflexives stems from three prevailing reasons. In the first group, there is no counterpart at all. The second group has surface counterparts, however they are not represented in the tectogrammatical tree. This concerns Czech basic reflexive pronouns, which are often hard to distinguish whether they are tightly bound to a verb or they fill an argument of the verb. The last group are English reflexive pronoun in its emphatic use. As shown in Example 4, they are often translated as words *samotný* or *sám* (*alone*), for which the automatic alignment often fails.

(4) *the ringers **themselves** will be drawn into*
zvonící **samotní** budou vtaženi do
the life
života

Mention type	aligned (%)		ment. match. ($P\ R$)				ante. sel. (^P_R F)		
	CS	EN	CS		EN		CS	EN	
Personal pron.	68.30	85.46	99.69	93.88	98.68	94.02	$^{95.34}_{93.92}$ 94.63	$^{95.20}_{90.59}$ 92.83	
Possessive pron.	84.06	75.42	99.78	98.24	99.80	94.66	$^{95.64}_{94.76}$ 95.20	$^{95.45}_{90.61}$ 92.97	
Refl. poss. pron.	87.08	—	100.00	98.56	—		$^{97.25}_{96.53}$ 96.89	—	
Reflexive pron.	17.69	69.21	100.00	79.56	100.00	37.31	$^{95.60}_{92.68}$ 94.12	$^{93.86}_{89.92}$ 91.85	
Demonstr. pron.	79.29	87.47	91.76	70.87	81.76	71.11	$^{84.79}_{82.40}$ 83.58	$^{91.41}_{88.25}$ 89.81	
Zero subject	84.31	—	99.76	89.37	—		$^{95.71}_{94.78}$ 95.25	—	
Zero in nonfin. cl.	80.15	49.90	100.00	85.07	99.84	85.06	$^{94.96}_{93.40}$ 94.17	$^{92.94}_{89.24}$ 91.05	
Relative pron.	86.88	91.62	99.05	80.48	97.29	90.26	$^{95.43}_{91.95}$ 93.66	$^{97.35}_{93.80}$ 95.54	
1st/2nd pers. pron.	85.14	84.46	88.42	88.53	94.20	79.73	$^{91.77}_{87.34}$ 89.50	$^{94.26}_{89.01}$ 91.56	
Named entities	79.60	96.83	50.09	87.51	91.24	48.61	$^{95.82}_{94.15}$ 94.98	$^{95.03}_{91.88}$ 93.43	
Nominal group	72.99	95.35	75.08	72.76	79.22	59.27	$^{90.52}_{87.46}$ 88.96	$^{90.60}_{84.37}$ 87.38	
Other	53.43	73.54	32.90	54.14	41.81	53.24	$^{80.08}_{78.42}$ 79.24	$^{70.72}_{70.05}$ 70.38	
Total	73.88	84.49	70.86	77.51	83.51	65.20	$^{92.91}_{90.69}$ 91.79	$^{92.73}_{88.31}$ 90.47	

Table 3: Results of the analysis of the three factors directly affecting projection quality. The scores are always measured from a target-language perspective.

> I live in hopes that the ringers themselves will be drawn into that fuller life.
>
> Žiji v naději, že i samotní zvonící budou vtaženi do tohoto plnějšího života.

Mention matching. A coreference relation cannot be correctly projected unless both the anaphor and the antecedent match a mention in the target language. Not matching a target-language mention is an error of the second type. To check what is the impact of mention matching, we measure it solely on aligned target language mentions and show the results in the middle part of Table 3.

In agreement with findings of (Grishina and Stede, 2017), we observe that pronouns and zeros in the top part of the table clearly approach matching precision of 100% in both projection directions. At the same time, named entities, nominal and other coreferential expressions in the bottom part of the table exhibit drops in precision. We presume that the precision score grows with decreasing length of the mention span.[2]

An interesting behavior is displayed by named entities. Whereas in Czech their precision is much lower than recall, these rates are very similar but swapped in English. A closer insight to the data gives us a clear explanation illustrated in Example 5. A modifier, such as *společnost* (*company*), *firma* (*firm*), *trh* (*market*) etc., is added to many named entities in Czech. It sounds more natural

and is easier to comprehend, especially if you are not familiar with the WSJ domain. This modifier is in fact a head of the complete named entity and, more importantly, it is the node that may corefer with others. Since it has no counterpart in English, no coreference is transferred to English, which results in recall errors for corresponding named entities. In the opposite projection, the English coreference link that is connected directly to one of the words in a given named entity finds its Czech counterpart, which is not a head of the mention, though. Hence, the Czech counterpart is in fact not coreferential, which causes a precision error. And because the head of the mention, the true coreferential node, is a word like *společnost*, the recall error incurred by not covering it falls into the category of nominal groups, not named entities.

(5)   *stake*   *that*   ***Burmah***   *announced*
podílu, o kterém **společnost Burmah** prohlásila,
   –[that it]   ***SHV***   *held*
že ho   **společnost SHV** držela

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

Vlastnictví 13.6 milionu akcií je nárůst oproti 6.7% podílu, o kterém společnost Burmah prohlásila, že ho společnost SHV držela k minulému pondělí.   (6)

Moreover, English reflexives see a dramatic fall in recall. These errors are again incurred for instances that translate to the Czech expressions *sám* or *samotný* (*alone*). Even if they are correctly aligned, these Czech expressions do not carry any

---

[2]Note that even if the tectogrammatical theory does not predetermine the mention span, it still exists.

coreference annotation. Therefore, no links can be projected.

Antecedent selection quality. If both the anaphor and the antecedent are correctly matched to some target-language mentions but these mentions belong to distinct chains, an error of the third type is incurred. The right-hand side of Table 3 shows the anaphora scores calculated on the same data as used until now, but only on correctly matched mentions. It accounts for around 49% of all coreferential links in Czech and 53% in English.

All F-scores move around 90% and more. The only exception is a category of Czech demonstrative pronouns. The reasons behind the errors related to them are various, including annotators' errors and alignment errors. But they are often caused by relatively free nature of demonstratives, which can refer to nominal groups, predicates, larger segments as well as entities outside the text. The free nature then allows the annotators to mark different (but somehow related) mentions as antecedents, especially when a different syntax structure of the languages encourages it. For instance, in Example 7 both expressions "the exchange" and "volume" are in some sense possible as the antecedent of "it". The same holds for the Czech translation.

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

7 Conclusion

Coreference projection performs poorly in both investigated directions. And since the experiments were undertaken on gold data, it is doubtful that performance with automatic links would be better.

The analysis confirmed the conclusions drawn in the related literature. First, the bottleneck of coreference projections seems to be alignment and mention-matching, incurring mostly recall errors. Second, precision of projection on pronouns is much better than for nominal groups and named entities, which leads us to the belief that shorter mentions are easier to project. However, to confirm it we would need to define span boundaries for tectogrammatical mentions.

Our analysis revealed also some more detailed findings. Reflexive pronouns seem to be very problematic. Not only are they difficult to align, but they neither excel at mention matching. Surprisingly, a relatively high proportion of Czech personal pronouns remain unaligned. The reason for this cannot be clearly generalized from the current corpus and thus should be verified on data that consist of various domains and translation directions.

8 Acknowledgments

References

Mariana S. C. Almeida, Cláudia Pinto, Helena Figueira, Pedro Mendes, and André F. T. Martins. 2015. Aligning Opinions: Cross-Lingual Opinion Mining with Dependencies. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, Volume 1: Long Papers*, pages 408–418, Stroudsburg, PA, USA. The Association for Computer Linguistics.

Amit Bagga and Breck Baldwin. 1998. Algorithms for Scoring Coreference Chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Shoshana Blum-Kulka. 1986. Shifts of Cohesion and Coherence in Translation. In *Interlingual and intercultural communication*, pages 17–35, Tübingen, Germany. Gűnter Narr.

Yulia Grishina. 2017. Combining the Output of Two Coreference Resolution Systems for Two Source Languages to Improve Annotation Projection. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 67–72, Copenhagen, Denmark. Association for Computational Linguistics.

Yulia Grishina and Manfred Stede. 2015. Knowledge-Lean Projection of Coreference Chains across Languages. In *Proceedings of the Eighth Workshop*

on Building and Using Comparable Corpora, pages 14–22, Beijing, China. Association for Computational Linguistics.

Yulia Grishina and Manfred Stede. 2017. Multi-Source Annotation Projection of Coreference Chains: Assessing Strategies and Testing Opportunities. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 41–50, Valencia, Spain. Association for Computational Linguistics.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection Across Parallel Texts. *Natural Language Engineering*, 11(3):311–325.

Mitchell Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Penn Treebank 3.

André F. T. Martins. 2015. Transferring Coreference Resolvers with Posterior Regularization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, Volume 1: Long Papers*, pages 1427–1437, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 169–176, Paris, France. European Language Resources Association.

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

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. 2017. Projection-based Coreference Resolution Using Deep Syntax. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 56–64, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michal Novák and Zdeněk Žabokrtský. 2014. Cross-lingual Coreference Resolution of Pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Franz J. Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maciej Ogrodniczuk. 2013. Translation- and Projection-Based Unsupervised Coreference Resolution for Polish. In *Language Processing and Intelligent Information Systems*, 7912, pages 125–130, Berlin / Heidelberg. Springer.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual Annotation Projection of Semantic Roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

Oana Postolache, Dan Cristea, and Constantin Orǎsan. 2006. Transferring Coreference Chains through Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 889–892, Genoa, Italy. European Language Resources Association.

Altaf Rahman and Vincent Ng. 2012. Translation-based Projection for Multilingual Coreference Resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 968–977, Stroudsburg, PA, USA. Association for Computational Linguistics.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht, Netherlands.

José G. C. de Souza and Constantin Orǎsan. 2011. Can Projected Chains in Parallel Corpora Help Coreference Resolution? In *Proceedings of the 8th International Conference on Anaphora Processing and Applications*, pages 59–69, Berlin, Heidelberg. Springer-Verlag.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan T. McDonald, and Joakim Nivre. 2013. Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. *TACL*, 1:1–12.

Don Tuggener. 2014. Coreference Resolution Evaluation for Higher Level Applications. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pages 231–235. The Association for Computer Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A Model-theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Conference on Message Understanding*, pages 45–52, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexander Wallin and Pierre Nugues. 2017. Coreference Resolution for Swedish and German using Distant Supervision. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 46–55, Gothenburg, Sweden. Association for Computational Linguistics.