Using Czech-English Parallel Corpora in Automatic Identification of *It*

Kateřina Veselovská, Ngụy Giang Linh, Michal Novák

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

Abstract

In this paper we have two goals. First, we want to present a part of the annotation scheme of the recently released Prague Czech-English Dependency Treebank 2.0 related to the annotation of personal pronoun *it* on the tectogrammatical layer of sentence representation. Second, we introduce experiments with the automatic identification of English personal pronoun *it* and its Czech counterpart. We design sets of tree-oriented rules and on the English side we combine them with the state-of-the-art statistical system that altogether results in an improvement of the identification. Furthermore, we design and successfully apply rules, which exploit information from the other language.

Keywords: personal pronoun *it*, pleonastic *it*, automatic identification, parallel corpus, coreference resolution

1. Introduction

In the majority of cases in English, the pronoun *it* illustrates nominal anaphora, tending to refer back to another noun phrase in the text. These cases have been surveyed as a part of anaphora resolution research and described e.g. in (Mitkov, 2002) or (Kučová et al., 2003). However, in a minor but still large enough class of cases, the pronoun *it* is used in exceptional ways that fail to demonstrate strict nominal anaphora and can be used without referring to any specific entity. In the present study we investigate mainly these occurrences.

Needless to say that the identification of pronouns to nominal expressions constitutes an important component of the process of coreference resolution, which has been found to be crucial in the fields of information extraction (Hirschman, 1997), machine translation (Peral et al., 1999), and automatic summarization (Harabagiu and Maiorano, 1999).

The English personal pronoun *it* can be translated into Czech as a demonstrative pronoun *to (this / that)* or a personal pronoun in singular *on / ona / ono (he / she / it)*, since English third person singular pronouns are distinguished according to animacy and gender, whereas Czech third person singular pronouns are used to identify grammatical gender only.

(1) Vezmu si **to**.
I will take RFLX **it**.
'I will take **it**.'

(2) (**Ono**) Je těžké v době krize sehnat práci.
(**It**) is difficult in times of crisis to get job.
'**It** is difficult in times of crisis to get a job.'

(3) Společnost Faulding uvedla, že (**ona**) vlastní
Company Faulding said, that (**she**) owns
33 % akcií společnosti Moleculon.
33% of voting stock of company Moleculon.
'Faulding said **it** owns 33% of Moleculon's voting stock.'

The Czech demonstrative pronoun *to* is usually used to refer back to a substantial section of a text, hence in this work we have decided to focus on the third person singular pronouns as the equivalents of the English *it* only. As mentioned before, the automatic identification of personal pronouns (coreferential or not) in English as well as in Czech plays an important role in coreference resolution.

In the present paper, the occurrences of personal pronoun *it* are identified using a parallel Czech-English dependency data collected in the Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0) (Hajič et al., 2011). The English part of PCEDT 2.0 contains the entire Penn Treebank-Wall Street Journal Section (Marcus et al., 1999). The Czech part consists of Czech translations of all of the Penn Treebank-WSJ texts. The corpus is 1:1 sentence-aligned. PCEDT 2.0 is a collection of linguistically annotated tree structures which is based on the theoretical framework of Functional Generative Description (FGD) (Sgall et al., 1967; Sgall, 1969). The annotation scheme of the PCEDT 2.0 consists of three layers: morphological, analytical and tectogrammatical. In the present study, we will mostly pursue the tectogrammatical layer (i.e. underlying structure).

The goal of this work is to use the benefits of the manually annotated parallel data in PCEDT 2.0 to construct a tool to determine anaphoricity of *it* or its Czech counterpart, even on the automatically analyzed data. Furthermore, our long-term objective is to improve the coreference resolution using bilingual parallel data not only from PCEDT 2.0, but also from much larger parallel corpus CzEng 1.0 (Bojar et al., 2011).

This paper is organized as follows. The English *it* and its Czech equivalent classification is described in Section 2. Section 3. provides a brief survey of related work. Section 4. presents the data we use for our system development. Description of the experiments for English and Czech is given in Section 5. and Section 6. Section 7. follows with the use of the parallel data. In Section 8., conclusions and ideas for future work are presented.

2. Theoretical Background

There have been several uses of *it* in English identified in the literature (Quirk et al., 1985; Sinclair, 1995; Swan, 1995). In FGD, we distinguish five basic types of personal pronoun *it* according to their function. They are described by the examples below:

1. The **anaphoric *it*** refers to a preceding noun denoting an inanimate entity or a not personalized animal.

   (4) I bought a new hat but my husband did not like *it*.

2. The **anticipatory *it*** anticipates on a part of the sentence which appears later in subject as well as in object position:

   (5) *It* is no good bothering about *it*.

   (6) *It* is feared that the ship was wrecked.

3. The **deictic *it*** belongs to deictic personal pronouns in general. It is used for deixis out of the language. The deictic pronoun as well as the copula verb must be in morphological agreement with the entity *it* refers to. The need of number agreement is typical of the deictic *it*.

   (7) Is *it* your suitcase (over there)?

4. The **exclamative *it*** is also used in deictic contexts but it refers to a situation implicitly known in the discourse rather than immediately to the given entity:

   (8) (Knock knock knock...) "*It*' s me, open the door!"

5. The **prop *it*** has little or no semantic content. It occurs in clauses which do not require any subject. It is typically clauses signifying time, atmospheric conditions and distance where the copula verb to be is regarded:

   (9) *It* is not far to New York.

   (10) *It* is 5 o'clock.

   (11) *It* is our wedding anniversary next month.

   (12) *It* is Sunday.

In Czech, it is natural to drop out personal pronouns in subject position of the clause. An overt subject pronoun indicates an emphasis of the speaker. Nevertheless the unexpressed subject pronoun can be understood from the verb morphological information thanks to its morpheme that identifies person, number and in some cases also gender.[1] In Nguy and Ševčíková (2011) four types of unexpressed subjects are distinguished:

1. The **implicit subject** most often stands for an entity already mentioned in the text or can be deictic.

   (13) Jana$_i$ ráda peče. Dnes Ø$_i$
   Jane gladly bakes. Today (she)
   upekla jablečný koláč.
   baked$_{3.SG.FEM}$ apple pie.
   'Jane likes to bake. Today she has baked an apple-pie.'

2. The **general subject** does not refer to any concrete entity; it has a general meaning, so it can be omitted in the surface structure.

   (14) S rizikem se Ø počítá.
   With risk RFLX (one) counts$_{3.SG}$.
   'Risk is counted in. (One counts risk in.)'

3. The **unspecified subject** denotes an entity more or less known from the context which is however not explicitly referred to.

   (15) Ø Hlásili to v rádiu.
   (They) Announced$_{3.PL.ANIM}$ it on radio.
   'It was announced on radio. (They announced it on radio.)'

4. The **null subject** does not refer to any entity in the real world. It is neither phonetically realized, nor can be lexically retrieved. In this case the predicate is an impersonal (weather) verb.

   (16) Zítra Ø bude oblačno.
   Tomorrow (it) will$_{3.SG}$ cloudy.
   'Tomorrow it will be cloudy.'

For the coreference resolution purpose, the personal pronoun distinction is simplified to **referential** and **non-referential**. As shown in (Evans, 2001; Nguy and Ševčíková, 2011), the automatic identification of other types has a poor accuracy because of its low occurrence.
The non-referential *it* is also referred to as **non-anaphoric** (Mitkov, 2002), **pleonastic** (Lappin and Leass, 1994) or **prop it** (Quirk et al., 1985).
We adopted the categorization from the PCEDT 2.0 annotation, which is as follows:

anaphoric – English anaphoric and anticipatory *it* and its equivalent Czech anaphoric unexpressed implicit third person singular subject.

non-anaphoric – English deictic and exclamative *it* and Czech deictic unexpressed implicit third person singular subject.

pleonastic – English prop *it* and Czech unexpressed general and null subject.

3. Related Work

Pleonastic pronouns have been resolved in a number of research on anaphora resolution. Lappin and Leass (1994)'s and Denber (1998)'s algorithm is based on pattern recognition, e.g. 'It is {a modal adjective} that'. Paice and Husk (1987)'s approach improves the pattern-matching process

[1] Gender is recognizable in past participle form of verbs only.

by constraints. As an illustration, a pronoun *it* is identified as non-referential if it occurs in the sequence 'it ... that'.
Evans (2001) proposes a machine learning based system for the automatic classification of *it*, which attempts to classify *it* for different usages such as nominal anaphoric, clause anaphoric, idiomatic, pleonastic and others. However, the system reports a high accuracy only on classifying pleonastic and nominal anaphoric *it*. The reason is simple, the features used in the training process are most appropriate for classification of pleonastic instances, and other types of *it* occur quite rare.
In recent years the study of pleonastic *it* identification has shifted toward different machine learning methods such as using support vector machines in (Litrán et al., 2004) or using a Bayesian network in (Hammami et al., 2010). Charniak and Elsner (2009) detect non-referential *it* in a unsupervised generative model. The detection of non-referential pronouns using counts from web-scale N-gram data is described in (Bergsma and Yarowsky, 2011).
For a task related to ours, a parallel corpus is used in (Camargo de Souza and Orăsan, 2011). Camargo de Souza and Orasan present a coreference resolution system for Portuguese trained on an English-Portuguese parallel corpus. The noun phrase coreference chains are identified thanks to the projected English coreference chains, which have been obtained from an English coreference resolver. Mitkov and Barbu (2002) develop a bilingual pronoun resolution system for English and French using an English-French parallel corpus, which benefits from the gender distinction of *it* in French and from the performance of the English algorithm.

4. Annotated Data

PCEDT 2.0 contains 2312 documents annotated at the tectogrammatical layer of Czech and English. Altogether, they consist of 49 208 pairs of sentences. Personal pronoun *it* has been annotated manually in all this data, independently in Czech and English part of the corpus, with the automatic word-alignment done afterwards (Mareček et al., 2008), including the alignment between nodes of the tectogrammatical layer.

4.1. Layers of Annotation

The PCEDT 2.0 annotation consists of multiple linguistically motivated layers:
The **m-layer** (morphological layer) captures the surface form of the sentence with words automatically part-of-speech tagged and lemmatized.
The **a-layer** (analytical layer) represents the surface syntax (a parse). The syntactic dependencies are provided with labels that carry the usual syntactic information; e.g. 'subject', 'attribute' or 'predicate complement'. Figure 1 presents the visualization of an analytical sentence representation.
The **t-layer** (tectogrammatical layer) is a linguistic representation that combines syntax and, to a certain extent, semantics, in the form of semantic labeling, coreference resolution[2] and argument structure description based on a valency lexicon. This representation draws on the framework of the Functional Generative Description.
The **p-layer** (phrase-structure layer) contains the original Penn Treebank annotation.

[2]Within the theoretical framework of FGD, coreference is divided into two subtypes: grammatical and textual (Panevová, 1991). **Grammatical coreference** occurs if the antecedent can be identified using grammatical rules and sentence syntactic structure (e.g. reflexive pronouns usually refer to the subject of the clause), whereas **textual coreference** is more context-based (e.g. personal pronouns).

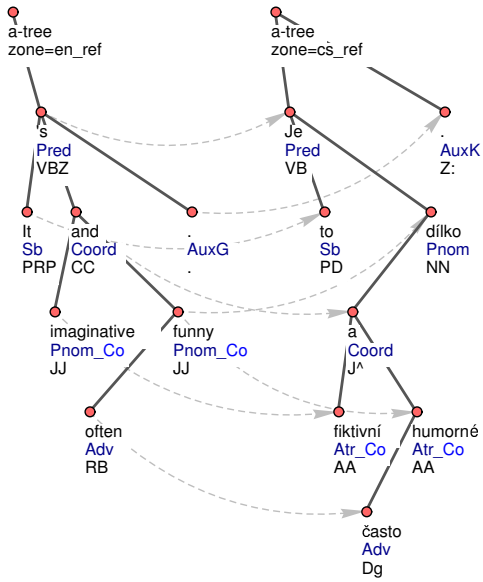

Figure 1: An example of parallel Czech-English a-trees representing sentences *It's imaginative and often funny* and *Je to fiktivní a často humorné dílko*.

4.2. Fully Automatic Annotation

In our study we use both manually annotated PCEDT 2.0 data and the same data automatically analyzed within the Treex framework (Žabokrtský, 2011).
Treex is a multi-purpose open-source framework for developing Natural Language Processing applications, which provides a wide range of integrated modules, such as tools for sentence segmentation, tokenization, morphological analysis, part-of-speech tagging (Spoustová et al., 2007), shallow and deep syntax parsing (McDonald et al., 2005), named entity recognition, anaphora resolution and others.
For our development we have the tokenized plain text from the PCEDT 2.0 of both languages as an input. Then we apply all possible tools in Treex to get them annotated at all layers. After that we used the automatic alignment tool. An example of the final alignment of Czech gold and automatic and English gold and automatic data at t-layer is shown on Figure 2.

4.3. Quantitative Properties

Thanks to the PCEDT 2.0 features mentioned in previous section we could easily distinguish three basic types of *it* in our corpora:

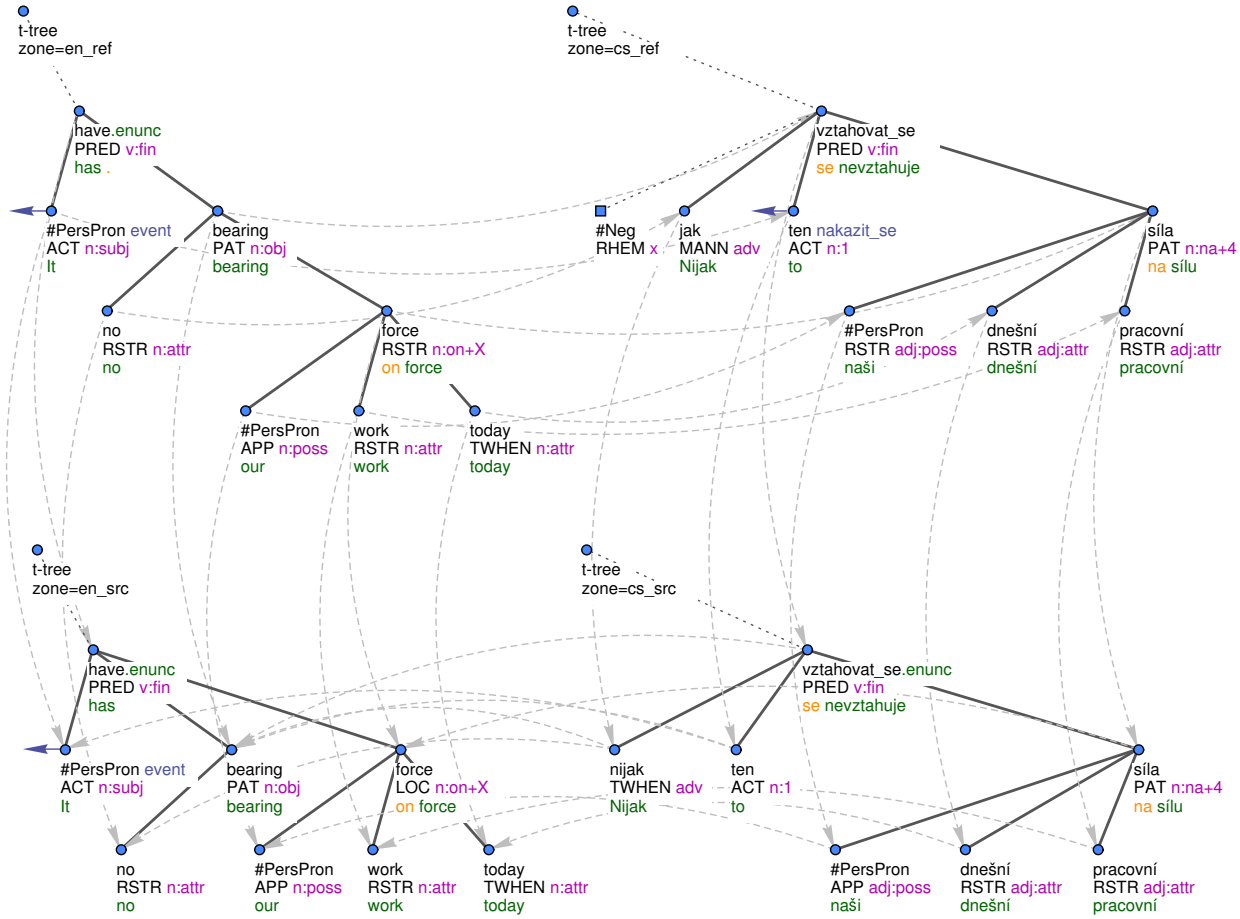

Figure 2: An example of gold parallel Czech-English t-trees aligned with automatic ones ([left to right, top to bottom]: English gold tree, Czech gold tree, English automatic tree, Czech automatic tree) representing sentences *It has no bearing on our work force today* and *Nijak se to nevztahuje na naši dnešní pracovní sílu*.

anaphoric – having a t-lemma substitute `#PersPron` (artificial t-lemma for overt and unexpressed personal pronoun[3]), an a-lemma `it` and a link pointing to its antecedent.

non-anaphoric – having a t-lemma substitute `#PersPron` and an a-lemma `it`, but not having a link pointing to its antecedent.

pleonastic – not having its own t-node on a tectogrammatical layer.

Their Czech equivalents are as follows:

anaphoric – a generated node representing third person singular pronoun having a t-lemma substitute `#PersPron` and a link pointing to its antecedent.

non-anaphoric – a generated node representing third person singular pronoun having a t-lemma substitute `#PersPron`, but not having a link pointing to its antecedent.

pleonastic – a generated node having a t-lemma substitute `#Gen` (artificial t-lemma for grammatical ellipsis of an obligatory argument - general argument) or not having its own t-node on a tectogrammatical layer.

Table 1 shows occurrence frequencies of anaphoric, non-anaphoric and pleonastic pronoun *it* on the English side and its counterparts on the Czech side of the PCEDT 2.0 subsets, we used for experimenting (see the following section).

	Dev data		Eval data	
	English	Czech	English	Czech
anaphoric	2053	4599	1932	3954
non-anaphoric	652	19	425	16
pleonastic	396	349	393	293

Table 1: Personal pronoun *it* number in PCEDT 2.0

We detected 911 occurrences of English anaphoric *it*, which has a Czech equivalent as a demonstrative pronoun *that (to)*; 3085 English non-pleonastic *it* having an equivalent Czech personal pronoun; 11 English pleonastic *it* that has a Czech pleonastic equivalent and 10 Czech pleonastic *it* with an English pleonastic corresponding node; 81 English and

[3] `#PersPron` also stands for textual ellipsis - obligatory arguments of a governing verb / noun.

21 Czech anaphoric *it* that refers to a clause or a sequences of sentences.

4.4. Experimental Data Subsets

In the experiments we used sections 00 – 10 of PCEDT 2.0 as a development data and sections 11 – 19 for final evaluation of proposed methods. The development data were not only aimed to be an inspiration for rules' design but their English side was used for training the bunch of parameters, as well (see Section 5.2.).

5. Resolution in English

For the English part of our work we have developed some hand-written rules on gold data. On automatically analyzed data we have integrated the state-of-the-art system NADA and used it as our baseline. Then we have applied and extended the rules to improve it.

5.1. Experiments on Gold Data

The rules applied on gold data are based on the grammatical, surface and deep syntactic information. Therefore, they are able to detect the pleonastic *it* but they hardly capture non-anaphoric *it*, which commonly requires the wider context or out-of-text information.
Thanks to the tectogrammatical tree structure, the pleonastic *it* identification on gold data is quite simple. In contrast to the Czech task, we do not limit ourselves to the *it*-subjects only, because the corresponding Czech *ono / to*-object is always referential, whereas the English one can be also pleonastic. The proposed algorithm is as follows:
For all personal pronouns *it* having a verb as its parent, **if** one of the following conditions is true:

1. The verb is active and has a predicate of a subordinate subject clause annotated as its Actor.

2. The verb is passive and has a predicate of a subordinate subject clause annotated as its Patient.

3. The verb's lemma is *make* and got a predicate of a subordinate subject clause annotated as its Patient. It is the case of *make it (easy / hard/ etc.) to*.

Then it is a pleonastic instance.

5.2. Experiments on Automatically Analyzed Data

The results of resolving pleonastic *it* on gold data are quite high, but that is only a motivation to improve the deep syntactic parser. Therefore, we have experimented with the NADA system and some other rules on automatically analyzed data.

Rule-based system
Because of the unreliability of automatically annotated actants, we have to change the rules used on gold data. The approach works as follows:
For all personal pronouns *it*, **if** *it* has a verb as its parent **and** one of the following conditions is true:

1. The verb's lemma is *be / become/ make / take* and has an infinitive among its children.

   (17) *It* doesn't take much to provoke an intense debate.

2. The verb's lemma is *be* and there are a subject complement expressed as a predicate nominative or a predicate adjective and a subordinate clause.

   (18) *It* is easy to see why the ancient art is on the ropes.

   (19) *It*'s a shame their meeting never took place.

3. The verb is an active cognitive verb (*appear / follow / matter / mean / seem*) or a passive cognitive verb (believe / expect / note / recommend / say / think) and has a subordinate clause.

   (20) Before the sun sets on the '80s, *it* seems nothing will be left unhocked.

   (21) *It* can be said that the trend of financial improvement has been firmly set.

Then it is a pleonastic instance.
The condition 1 and 2 are further modified to prevent error cases, where *it* has been misannotated to be a child of other node than the verb in condition 1 or the subordinate clause is a subtree of the subject complement instead of the main predicate in condition 2.

NADA system
The NADA system (Bergsma and Yarowsky, 2011) is the state-of-the-art tool for anaphoricity determination of English *it*. Following the lexical and web count features, every occurrence of *it* is assigned a probability of being referential with a previously-mentioned entity. After having set the decision boundary (by default, it is 0.5), the occurrences can be binary classified as anaphoric and non-anaphoric.
The indisputable advantage of NADA is that the input does not have to be linguistically pre-processed at all, it accepts a surface text. Moreover, no linguistic analysis is being performed inside the tool. It makes NADA very simple and quick. On the other hand, if the rich linguistic annotation is available, it cannot exploit it.
As this software is freely available, we were able to integrate it into the Treex framework and combine the tree-oriented rules with the estimates produced by NADA.

Combination of NADA and rules
By combination of the statistical system working on a surface level and tree oriented hand-crafted rules we aimed to extract the best from both approaches. We decided to make a linear interpolation of the features, which consisted of every single rule in the previous approach, their disjunction and quantized values of NADA probability estimates. The parameters have been learnt from the development data using a maximum entropy classifier.[4]

[4]We employed the Perl module `AI::MaxEntropy`

5.3. Evaluation

As we stated in Section 5.2., NADA is a binary classifier distinguishing between anaphoric *it* and the other types. Since PCEDT 2.0 differentiate between 3 types of *it*, in order to successfully combine NADA with the designed rules two of these classes must be merged into one. We conducted experiments with 2 of 3 possible binarizations. The one with a merged class of anaphoric and non-anaphoric was left out as our central target is to be able to distinguish between these two classes.

The binarization with a joint class of non-anaphoric and pleonastic (NON-ANAPH+PLEO) as a class of positive instances accords with the way NADA was meant to be used. The overall results assessed in terms of accuracy as well as precision, recall and F-score measured on the positive class can be seen in Table 2.

NADA alone achieves a score similar to accuracy of 86% reported in (Bergsma and Yarowsky, 2011).[5] In comparison, relying just on the designed rules cannot compete with NADA, suffering mostly from a low coverage of the rules, reflected in a low value of recall. Even on the gold data the rules perform slightly worse mostly because they were tuned to describe just pleonastic occurrences. Combination of the statistical system and rules seemed to be promising. However, we register only a slight improvement of the success rate compared to NADA used separately.

The classes of anaphoric and non-anaphoric (mostly deictic and referring to a larger segment) *it* are alike in terms of referring to something, opposed to its pleonastic usage. Moreover, we constructed the rules to fit the class of pleonastic occurrences mainly, which suggests a better score than in case of the above-mentioned binarization. Following experiments are carried out with pleonastic *it* (PLEO) being a positive class.

The score of NADA alone in this configuration is surprisingly better, even though it was not supposed to be evaluated in this way. The values of precision and recall on a positive class changed, apparently due to changes in the distribution between positive and negative instances. As opposed to the previous configuration, the pure rule-based system outperforms NADA in accuracy here, also reaching a higher precision, which can be justified by the fact that the rules were tailored to recognize the pleonastic occurrences. The combination of both approaches results in the best accuracy of almost 90%, outperforming both of the components if used alone.

6. Resolution in Czech

Because of the Czech phenomena of subject absence, we attempt to identify the instances of predicates, to which a personal pronoun will be generated as a substitution of the unexpressed subject. First we apply hand-written rules on gold data, secondly the same rules in automatic data. Then the rules are improved and added by information from English automatic data (see Section 7.).

[5] Recall that NADA does not require any linguistic annotation, so it achieves the same score for the manually as well as the automatically analyzed data.

6.1. Experiments on Gold Data

Our heuristic procedure for identifying unexpressed implicit subject occurrences (anaphoric and non-anaphoric *it*) is based on constraints. We eliminate cases, where it is an overt subject, an unexpressed general subject or null subject. The procedure works as follows:

For all third person singular verbs, **if** all of the following conditions are true:

1. There is no overt subject, that is:
   (a) There is no overt subject represented by a word.
   (b) There is no subject subordinate clause.
2. There is no unexpressed general subject, that is:
   (a) The verb is not a part of the phrase *Je vidět / slyšet / cítit ((It) is seen / heard / felt)*.
   (b) The verb is not a part of the phrase *Lze / Je možné / Je nutné ((One) can / (It) is possible / (One) needs)*.
   (c) The verb is not a reflexive passive, because a third personal singular reflexive passim often determines a general subject.
   (d) The verb has no an *-o* ending, because the *-o* ending indicates a third personal neuter verb and it seems, a third personal neuter verb often implicates an instance of a general subject.
3. There is no null subject, that is:
   (a) The verb is not an impersonal (weather) verb *jednat se / pršet / zdát se / dařit se / oteplovat se / ochladit se / stát se / záležet (be about / rain / seem / do well / get warmer / get colder / happen / depend)*.
   (b) The verb is not a part of the phrase *Jde o ((It) is about)*.

Then there will be added a generated personal pronoun.

6.2. Experiments on Automatically Analyzed Data

The algorithm for anaphoric and non-anaphoric *it* identification on automatically analyzed data is extended by adding conditions to prevent errors that appear in the automatic annotation.

For all third person singular verbs, **if** all of the following conditions are true:

1. There is no overt subject, that is:
   (a) There is no overt subject represented by a word – *unchanged*.
   (b) There is no subject subordinate clause. The same condition on gold data was true, when the head of the subordinate clause was a finite verb having functor Actor. The new condition was true for finite verbs having functor Actor or Patient, because of the functor misannotation.

	NON-ANAPH+PLEO				PLEO			
	A	P	R	F	A	P	R	F
EN: Majority class	70.30	–	–	–	85.75	–	–	–
EN: Rules-gold	83.76	99.31	39.15	56.16	94.67	90.31	68.68	78.03
EN: Rules-autom	76.31	73.24	31.90	44.44	87.54	56.90	51.66	54.16
EN: NADA	83.86	81.10	59.51	68.65	86.19	51.00	78.01	61.68
EN: NADA + Rules-autom	84.44	78.61	65.40	71.40	89.83	71.88	47.06	56.88

Table 2: The results of evaluation of all tested systems, including two types of evaluation (NON-ANAPH+PLEO and PLEO). Quality of the systems was measured on the Evaluation data in terms of accuracy (A), precision (P), recall (R) and F-score (F). Majority class system corresponds to assigning a majority class to all candidates.

(c) If the verb is active, then it has no Actor among its children. This condition prevents errors in automatic subject annotation in the Czech part, where the overt subject was misannotated as other part-of-speech.

(d) If the verb is passive, then it has no Patient among its children (subject error prevention).

2. There is no unexpressed general subject – *unchanged*.

3. There is no null subject – *unchanged*.

Then there will be added a generated personal pronoun.

6.3. Evaluation

Contrary to the English task, where all personal pronouns *it* are presented on the surface sentence and we attempt to identify occurrences to be hidden on the tectogrammatical layer, the Czech target is detecting dropped third person singular pronouns in the subject position in order to express it on the tectogrammatical layer.
We use the binary classification of unexpressed third pronominal singular subject:

- referential – anaphoric and non-anaphoric dropped pronoun in the subject position having a generated node and being a child of the predicate.
- non-referential – pleonastic pronoun not being expressed either on the surface sentence or on the tectogrammatical layer.

There is another difference between the English task and the Czech task. Whereas a non-pleonastic pronoun for the English part means an anaphoric or non-anaphoric *it* only, a non-pleonastic pronoun for Czech is an anaphoric or non-anaphoric *he / she / it*. The reason lies on the gender differentiation of non-animal nouns and the use of gender differentiated pronouns to refer to them in Czech.
The rules on Czech data were implemented to suit the task: looking for a referential/implicit unexpressed subject and generating a tectogrammatical node for it. The scores of both systems are shown in Table 3.
Applying the rules on automatically analyzed data gives a perceptibly lower result than the rules on gold data. It is not surprising because on automatically analyzed data the overt subject is often misannotated as an object or other part-of-speech and vice versa. The subject subordinate clause is not straightforwardly recognizable, too.

7. Exploiting the Parallel Corpus

In the experiments so far, the proposed rules have employed just that language side of the corpus, which they were constructed for. We attempted to exploit the parallel nature of the PCEDT 2.0 corpus by designing rules that look also at the other side.
In general, information from the English side of automatically analyzed trees tends to be more reliable than the one from the Czech side. Particularly, it confirmed to be true for English rules, which used the Czech data. Such rules had no effect when they were combined with other rules for English.
On the other hand, in the opposite direction we designed the following rules:
For all third person singular verbs, **if** all of the following conditions is true:

1. The corresponding English verb has no non-pronominal subject. This condition prevents errors in automatic subject annotation in the Czech part, where the overt subject was misannotated as other part-of-speech.

2. There may be an unexpressed implicit subject, that is one of the following conditions is true:

   (a) Conditions 1 – 3 on automatically analyzed data are true.

   (b) The corresponding English verb has a *he / she* subject. This condition helps to detect cases, where the Czech conditions wrongly identified the existence of an overt subject. See error examples below:

   (22) Na noc se vrací do opuštěné
   At night RFLX returns to condemned
   budovy, kterou nazývá domovem.
   building, which **calls** home$_{ACT.error}$.
   'At night he returns to the condemned building **he calls** home.'

   (23) Banka First Union, říká,
   Bank$_{Sb-of-says.error}$ First Union, **says**,
   má nyní balíčky pro sedm skupin
   has now packages for seven groups
   zákazníků.
   of customers.

'First Union, **he says**, now has packages for seven customer groups.'

Then there will be added a generated personal pronoun.

These turned out to substantially contribute on the final quality of the whole rule-based system thanks to the information about English corresponding personal pronouns *he / she* that are expressed on the surface sentence and subjects, because the subject of an English clause can be also detected easier. Table 3 shows that if we include these inter-language rules, the accuracy increases by almost 3.5% absolute.

	ANAPH+NON-ANAPH			
	A	P	R	F
CZ: Majority class	86.58	–	–	–
CZ: Rules-gold	98.79	92.89	98.39	95.56
CZ: Rules-autom	87.68	52.97	73.34	61.51
CZ: Rules-autom+EN	91.08	64.20	75.87	69.55

Table 3: The results of evaluation of rule-based systems for Czech. Configuration "Rules-autom+EN" shows an impact of adding rules that use the English side

8. Conclusion

In this paper we have presented the annotation of personal pronoun *it* in the recently released Prague Czech-English Dependency Treebank 2.0. We have analyzed its occurrences in both languages and developed rule-based approaches to automatically identify the Czech and English *it* types. On the English side we also combined these tree-oriented rules with the statistical state-of-the-art system for this task, which improved the success rate on resolution of pleonastic occurrences.

Furthermore, we successfully exploited the parallel nature of the PCEDT 2.0 corpus and employed the English data in the task of Czech *it* identification.

In the future work, we plan to develop new rules and integrate machine learning methods in a greater extent. In addition, we would like to apply such system along with a coreference resolver to the much larger automatically analyzed parallel corpus CzEng 1.0 (Bojar et al., 2011). We hope the self-training on larger data together with a richer rule-/feature-set to increase the quality of coreference resolution.

9. Acknowledgments

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013). This work has been supported by the Czech Science Foundation under the contract 201/09/H057 and by the grant GAUK 4226/2011. The authors would like to thank prof. Eva Hajičová, assoc. prof. Zdeněk Žabokrtský and the anonymous reviewers for their valuable comments and suggestions to improve the paper.

10. References

Shane Bergsma and David Yarowsky. 2011. NADA: A robust system for non-referential pronoun detection.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2011. Czeng 1.0.

José Guilherme Camargo de Souza and Constantin Orăsan. 2011. Can Projected Chains in Parallel Corpora Help Coreference Resolution? In *Anaphora Processing and Applications*, pages 59–69. Springer.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 148–156, Athens, Greece, March. Association for Computational Linguistics.

Michael Denber. 1998. Automatic Resolution of Anaphora in English. Technical report, Eastman Kodak Co, Imaging Science Division.

Richard Evans. 2001. Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing*, 16(1):45 – 57.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2011. Prague Czech-English Dependency Treebank 2.0.

Souha Mezghani Hammami, Rahma Sallemi, and Lamia Hadrich Belguith. 2010. A bayesian classifier for the identification of non-referential pronouns in arabic. In *In Proceedings of the 7th International Conference on Informatics and Systems- INFOS 2010*, pages 1–6.

Sandra M. Harabagiu and Steven J. Maiorano. 1999. Knowledge-Lean Coreference Resolution and its Relation to Textual Cohesion and Coherence. In *The Relation of Discourse/Dialog Structure and Reference*.

Lynette Hirschman. 1997. MUC-7 Coreference Task Definition.

Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, and Oliver Čulo. 2003. Anotování koreference v pražském závislostním korpusu. Technical Report TR-2003-19, ÚFAL MFF UK, Prague, Prague.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Comput. Linguist.*, 20(4):535–561, dec.

José Carlos Clemente Litrán, Kenji Satou, and Kentaro Torisawa. 2004. Improving the identification of non-anaphoric it using support vector machines. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, JNLPBA '04, pages 58–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

David Mareček, Zdeněk Žabokrtský, and Václav Novák. 2008. Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In *Proceedings of the Twelfth EAMT Conference*, pages 102–111.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver.

Ruslan Mitkov and Catalina Barbu. 2002. Using bilingual corpora to improve pronoun resolution. *Languages in Contrast*, 4(2):201–211.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, London.

Giang Linh Ngụy and Magda Ševčíková. 2011. Unstated Subject Identification in Czech. In *WDS'11 Proceedings of Contributed Papers, Part I*, pages 149–154.

Chris D. Paice and Gareth D. Husk. 1987. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun *it*. *Computer Speech and Language*, 2.

Jarmila Panevová. 1991. Koreference gramatická nebo textová? In *Etudes de linguistique romane et slave*. Krakow.

Jesús Peral, Manuel Palomar, and Antonio Ferrández. 1999. Coreference-oriented interlingual slot structure machine translation. In *In Proceedings of the ACL Workshop Coreference and its Applications*, pages 69–76.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

Petr Sgall, Alla Goralčíková, Eva Hajičová, and Ladislav Nebeský. 1967. *Generativní popis jazyka a česká deklinace*. Prague:Academia.

Petr Sgall. 1969. *A Functional approach to syntax in generative description of language*. American Elsevier Pub. Co.

John M. Sinclair. 1995. *English Grammar*. Harper Collins Publisher, UK.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, ACL '07, pages 67–74, Stroudsburg, PA. Association for Computational Linguistics.

Michael Swan. 1995. *Practical English Usage*. Oxford University Press, UK.

Zdeněk Žabokrtský. 2011. Treex – an open-source framework for natural language processing. In *Information Technologies – Applications and Theory*, pages 7–14.