M A T E M A T I C K O - F Y Z I K Á L N Í F A K U L T A

P R A H A

COREFERENCE RESOLUTION IN THE PRAGUE DEPENDENCY TREEBANK

Ngụy Giang Linh, Michal Novák, Anna Nedoluzhko

ÚFAL/CKL Technical Report
TR-2011-43

U N I V E R S I T A S C A R O L I N A P R A G E N S I S

Copies of ÚFAL/CKL Technical Reports can be ordered from:

Institute of Formal and Applied Linguistics (ÚFAL MFF UK)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague 1
Czech Republic

or can be obtained via the Web: http://ufal.mff.cuni.cz/techrep

Coreference Resolution in the Prague Dependency Treebank

Ngụy Giang Linh, Michal Novák, Anna Nedoluzhko

This technical report summarizes results obtained during the research on coreference resolution at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, during 2009 - 2011. It contains a brief description of foreign approaches to this topic, a description of manual coreference annotation in the Prague Dependency Treebank 2.0 and an account of possibilities of automatic coreference annotation.

This work has been supported by the grants GAČR 405/09/0729, MŠMT ČR LC536, GAUK 4383/2009, GAUK 4226/2011 and GAČR 201/09/H057.

Contents

Chapter 1

Introduction

In spoken and written language it is commonly observed that the same real-world entity is referred to by a variety of noun phrases. The task of **coreference resolution** is to determine which noun phrases in a text or dialogue refer to the same real-world entity. An accurate coreference resolution is required by many natural language processing applications such as machine translation, information extraction etc.

1.1 Basic Terminology

Natural languages provide speakers with a variety of ways to refer to entities. Two referring expressions that are used to refer to the same real-world entity are said to **corefer**. Reference to an entity that has been previously introduced into the discourse is called **anaphora**. **Anaphor** is a given referring expression and the entity to which it refers is its **antecedent**. The anaphor and its antecedent refer to the same entity in the real world; hence, they are **coreferential** with each other. All expressions in a text or dialogue referring to the same entity form a coreference sequence called **coreferential chain**. A typical coreference resolution system (depicted in Figure 1.1) takes an arbitrary document as input and produces the appropriate coreferential chains as output.

1.2 Coreference Types

There are many varieties of coreference according to the form of the anaphor and antecedent or to their locations. In subsections below we describe coreference types typical for Czech. For a more complete coreference categorization see [Mitkov, 2002].

1.2.1 Pronominal Anaphora

Pronominal anaphora arises when a referring expression is:

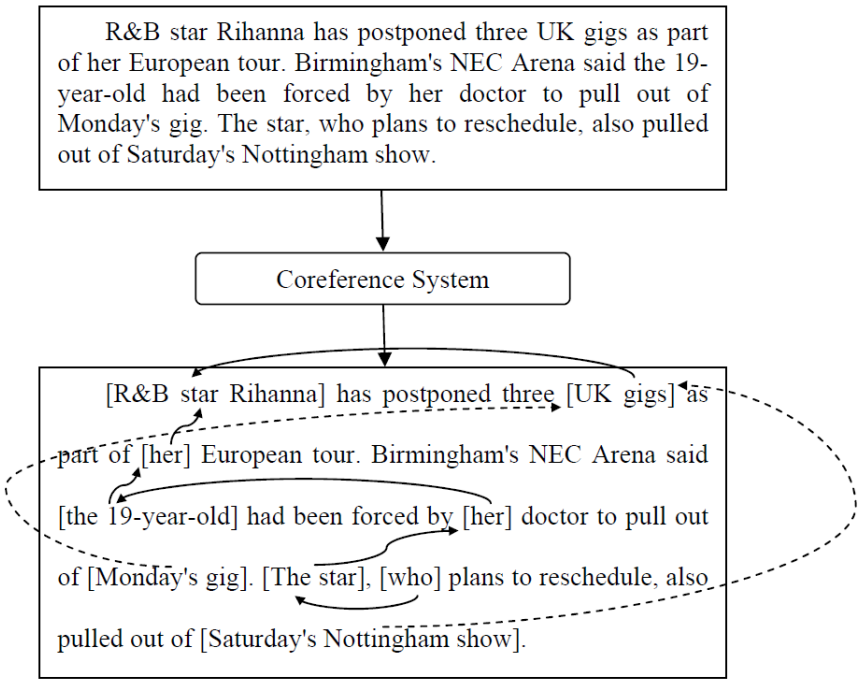

Figure 1.1: Coreference System: A full arrow represents anaphora based on identity; a dashed arrow stands for brigding anaphora, i.e. a reference to the antecedent based on generic knowledge (in our example, a relation of part-whole/element-set).

a personal pronoun :

(1.1) *Karel Schwarzenberg v neděli prohlásil, že pokud [Praha]$_i$ eurozóně nepůjčí, hrozí [jí]$_i$ izolace.*

'Karel Schwarzenberg said on Sunday that if [Prague]$_i$ doesn't give the euro area a loan, [it]$_i$ can be threatened by isolation.'

a possessive pronoun :

(1.2) *Na rozdíl od jiných 130 [Newtonových]$_i$ rukopisů nyní mají zájemci o [jeho]$_i$ práci možnost vidět [jeho]$_i$ takřka kompletní dílo.*

'Unlike other 130 [Newton's]$_i$ manuscripts, now people interested in [his]$_i$ work can see [his]$_i$ almost complete work.'

a reflexive pronoun :

(1.3) *[Dramaturgie]$_i$ [si]$_i$ z půlstoleté Stahuljakovy tvorby jednostranně vybrala skladby z dvacátých let.*

'[The dramaturgy]$_i$ unilaterally chose songs from the twenties of the half century of Stahuljak's works to [itself]$_i$.'

(1.4) *[Sázková kancelář Fortuna]$_i$ přepustila [svůj]$_i$ menšinový podíl výhradně občanům Slovenska.*

'[The betting agency Fortuna] rendered [its.RFLX]$_i$ minority share entirely to citizen of Slovakia.'

a demonstrative pronoun :

(1.5) *Především společenským bonbónkem se stal písňový cyklus op. 4 v podání [ambasadorovy choti]$_i$, [sopranistky Ivanky Stahuljak]$_i$. [Ta]$_i$ zaujala spíše výrazovou stránkou projevu a niternosti uměleckého prožitku než kvalitou či technikou hlasu.*

'Above all, social gems became a song cycle of the op. 4 in rendition of [the ambassador's wife]$_i$, [soprano Ivanka Stahuljak]$_i$. [That.fem.sg]$_i$ captivated more by expressive aspect of speech and interiority of artistic experience than voice quality or technique.'

(1.6) *[Rozprava o podobě reformy veřejných financí bude zahájena ve středu. Všechna jednání proběhnou za zavřenými dveřmi.]$_i$ Lidovým novinám [to]$_i$ sdělil včera ministr financí.*

'[The debate about the form of public finance reform will be opened on Wednesday. All meetings will take place behind closed doors.]$_i$ The Minister of Finance told [that]$_i$ to Lidove noviny yesterday.'

a relative pronoun (or an adverb) :

(1.7) *Do [diskuse]$_i$, [která]$_i$ rozděluje politickou scénu, se v pondělí zapojil [prezident Václav Klaus]$_j$, [který]$_j$ má v úterý osobně přijít vládě vymluvit půjčku zadluženým zemím eurozóny.*

'On Monday [the debate]$_i$, [which]$_i$ divides the political scene, was joined by [President Vaclav Klaus]$_j$, [who]$_j$ has to come in person on Tuesday to talk the government out of giving loan to indebted countries in the euro area.'

(1.8) *Členové družstva ČR a SR se měli sejít v kompletním složení [včera]$_i$, [kdy]$_i$ z turnajů v Gstaadu a Ósace přicestovali Karel Nováček s Petrem Kordou.*

'Team members of the Czech and Slovak Republic should meet in the full composition yesterday, when Karel Novacek and Petr Korda arrived from the tournament in Gstaad and Osaka.'

1.2.2 Zero Anaphora

Zero or null anaphora, ellipsis, occurs when anaphoric expressions are not expressed but nevertheless understood.

Zero pronominal anaphora occurs in case of the most common form of ellipsis, where pronouns are omitted. This phenomenon of ”pronoun-dropping” usually appears in Japanese, Chinese, Spanish, Portuguese and Slavic languages such as Czech and Polish (pro-drop languages).

(1.9) *[Otec]$_i$ vždycky tvrdil, že Ø$_i$ opery nesnáší. Ø$_i$ říkal, že [mu]$_i$ na opeře vadí hlavně ten zpěv.*

‘[Father]$_i$ always said, that (he)$_i$ hated opera. (He)$_i$ said it was the opera singing that primarily annoyed [him]$_i$.’

Another subtype of zero anaphora is **control**. We work with the theory of control present within the dependency-based framework of Functional Generative Description (FGD, [Sgall et al., 1986]), in which control is defined as a relation of a referential dependency between a controller (antecedent - a participant of the main clause) and a controllee (anaphor - empty subject of the nonfinite complement (controlled clause)).

(1.10) *Novelu zákona o malé privatizaci včera [sněmovně]$_i$ doporučil Ø$_i$ schválit rozpočtový výbor.*

‘The budget committee recommended [the Chamber]$_i$ Ø$_i$ to approve the amendment to the small privatization ’

Anaphora also arises in **reciprocity** constructions. In Czech, a reciprocal anaphor can be expressed by the reflexive `se/si` or it can be omitted. The reciprocal anaphor refers to the subject and they fill together the role of both verbal arguments expected on the basis of verbal valency (see [Panevová, 1999], [Panevová, 2007]).

(1.11) *[Sultáni]$_i$ [se]$_i$ vystřídali na trůnu.*

‘[Sultans]$_i$ changed [each other]$_i$ on throne.’

1.2.3 Nominal Anaphora

In nominal anaphora, an anaphor can be any kind of phrases the head of which is a noun, pronoun or other noun-like word. In non-pro-drop languages like English, this class of anaphora covers whole coreferential chains, therefore it has been researched most widely. In our work for Czech, we use this definition to also include zero pronominal anaphora.

(1.12) *[Policejní prezident Petr Lessy]$_i$ se v pondělí znovu postavil proti razantnímu osekávání rozpočtu policie, kvůli kterému by muselo do roku 2014 odejít téměř 10 tisíc policistů. Podle [jejich šéfa]$_i$ by to bylo likvidační, policistů je už teď nedostatek.*

'On Monday [police president Petr Lessy]$_i$ stood again against firm chipping of police budget, due to which nearly 10 thousand policemen would have to leave by 2014. According to [their boss]$_i$ it would be liquidation, the police is already scarce.'

1.2.4 Bridging Anaphora

Bridging anaphora or indirect anaphora is a relation between two elements in which the anaphor indirectly refers to its antecedent on the basis of the reader's common sense inference.

(1.13) *Když se [Take That]$_i$ rozpadla, kritici nedali [Robbie Williamsovi]$_i$ žádnou šanci na úspěch.*

'When [Take That]$_i$ split up, critics didn't give [Robbie Williams]$_i$ any chance of success.'

(1.14) *Po včerejším tréninku mě bolí [celé tělo]$_i$, nejvíc [obě nohy]$_i$.*

After yesterday's training [my whole body]$_i$ hurts, [both legs]$_i$ the most.

1.2.5 Other Types of Coreference

Cataphora refers to an anaphoric relation in which a referring expression refers to the entity mentioned explicitly later in the text.

Exophora or **deixis** arises when the antecedent is not expressed in the discourse but nevertheless understood according to the given context or situation.

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).

1.3 Evaluation Metrics in Coreference Resolution

Precision and recall are two widely used measures for evaluating the quality of results. Precision can be seen as a measure of exactness, whereas recall is a measure of completeness. There are different evaluation metrics for coreference resolution, but we describe only the pairwise one, which we use to evaluate our coreference systems. In the pairwise evaluation, the precision is the number of noun phrase pairs correctly labeled as coreferential (true positives, see Table 1.1) divided by the total number of pairs labeled as coreferential (i.e. the sum of true positives and false positives,

which are pairs incorrectly labeled as coreferential). Recall in this context is defined as the number of true positives divided by the total number of pairs that actually corefer (i.e. the sum of true positives and false negatives, which are pairs which were not labeled as coreferential but should have been).

	Correct classification	
Obtained classification	true positive (TP)	false positive (FP)
	false negative (FN)	true negative (TN)

Table 1.1: Comparison between the given classification of a noun phrase pair and the desired correct classification.

Usually, precision and recall scores are combined into a single measure, the F-measure, which is the weighted harmonic mean of precision and recall.

$$Precision = \frac{TP}{TP + FP} = \frac{\text{number of correctly predicted coreference links}}{\text{number of all predicted links}}$$

$$Recall = \frac{TP}{TP + FN} = \frac{\text{number of correctly predicted coreference links}}{\text{number of all coreference links}}$$

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Chapter 2

Coreference Annotation in Text Corpora

Coreferential and bridging relations between discourse entities are of major importance for establishing and maintaining textual coherence. The ability to automatically resolve these kinds of relations is an important feature of text understanding systems. For both the training as well as the evaluation of these systems, manually annotated corpora are required. That is the reason why several anaphoric annotation schemes have been presented just in the last few years.

2.1 Foreign Coreference Annotation Systems

The MUC scheme [MUC-7, 1998] and its continuation ACE [Doddington et al., 2004] are the best known and most widely used coreference schemes, developed primarily for the information extraction and other NLP tasks. Being applied to rather limited corpora, the MUC is the only existing coreference annotation scheme whose reliability has been systematically tested. Priority is given to preserving high interannotator agreement, so only identity relations for nouns, NPs and pronouns are annotated for coreference. The ACE program is limited to the recognition of seven entity types (person, location etc.), for which identical coreferential relations are annotated.

The MATE project, its extension on the GNOME and VENEX corpora [Poesio, 2004] and the ongoing project of the ARRAU corpus [Poesio and Artstein, 2008] are the most well-known projects where also bridging relations are annotated. Based on MATE, the annotation scheme for coreference in Spanish was developed [Potau, 2008], but bridging relations have not been annotated largescale there.

In PoCoS [Krasavina and Chiarcos, 2007], two layers of coreference annotation schemes were proposed: the Core Scheme is general and reusable, while the Extended Scheme supports a wider range of specific extensions. The Core Scheme is used for annotating some cases of nominal coreference, while non-nominal coreference and bridging relations are annotated as part of the Extended Scheme.

All coreference annotation schemes described above consist of two steps. First, the so called "markables" (the linguistic items between which coreference relations might hold) are (mostly automatically) marked, second, the relation itself is (mostly manually) determined. Markables are specified differently according to the given scheme.

In GNOME, all NPs are treated as markables, including predicative NPs, in MUC all nouns, NPs and pronouns, including 1st and 2nd person pronouns are markables, PoCoS has a sophisticated system of primary and secondary markables. Primary markables are all potential anaphors, they include definite NPs, pronouns and some other anaphoric elements. Secondary markables are e.g. indefinite NPs and are subject to annotation only if they serve as antecedents of a primary markable.

The BBN Pronoun Coreference and Entity Type Corpus [Weischedel and Brunstein, 2005] is a manually annotated one million word Penn Treebank corpus of Wall Street Journal texts. The corpus contains stand-off annotation of pronoun coreference as well as annotation of a variety of entity and numeric types.

Manual annotation of coreference is costly and time-consuming, therefore the PlayCoref project comes up with the idea of using coreference links annotated by game players via internet. This alternative way of the coreference annotation collection is supposed to get a substantially larger volume of annotated data than any expert annotation can ever achieve.

2.2 Coreference Annotation in the Prague Dependency Treebank

2.2.1 Prague Dependency Treebank 2.0

The Prague Dependency Treebank 2.0 [1] (PDT 2.0, [Jan Hajič, et al., 2006]) is a large collection of linguistically annotated data and documentation, based on the theoretical framework of Functional Generative Description. In PDT 2.0, Czech newspaper texts selected from the Czech National Corpus are annotated using a rich annotation scenario divided into three layers:

- **morphological layer** (m-layer), on which a lemma and a positional morphological tag are added to each token (word form or punctuation mark) in each sentence of the source texts,
- **analytical layer** (a-layer), where each sentence is represented as a surface-syntactic dependency tree, in which each node corresponds to one m-layer token; edges correspond either to dependency relations between tokens (such as subject, object, attribute), or to other relations of a non-dependency nature (such as coordination),
- **tectogrammatical layer** (t-layer, see [Mikulová et al., 2005] for details), where each sentence is represented as a complex deep-syntactic dependency tree (tectogrammatical tree, t-tree), in which only autosemantic words have nodes of their own (functional words such

[1] http://ufal.mff.cuni.cz/pdt2.0/

as prepositions or auxiliary verbs are represented by other means); on the other hand, tectogrammatical trees contain also nodes having no counterparts in the surface shape of the sentences, for instance nodes corresponding to 'pro-dropped' subjects. Coreference annotation is considered as one of the components of the t-layer annotation scheme.

PDT 2.0 contains 3,168 newspaper texts annotated at the tectogrammatical level. Altogether, they consist of 49,431 sentences. Coreference has been annotated manually in all data. There are 45,174 coreference links (counting both textual and grammatical ones). In PDT 2.0 following grammatical and textual coreference relations are annotated (see their occurrence frequency in Table 2.1):

- **grammatical coreference** - reflexive pronouns, relative pronouns/adverbs, arguments of verbs of control and reciprocity;
- **textual coreference** - (expressed and zero) 3rd person and possessive pronouns, demonstrative pronouns

Type/Count	`train`	`dtest`	`etest`
Personal pron.	12,913	1,945	2,030
Relative pron.	6,957	948	1,034
Controllees	6,598	874	907
Reflexive pron.	3,381	452	571
Demonstrative pron.	2,582	332	344
Reciprocity pron.	882	110	122
Other	320	35	42
Total	34,983	4,909	5,282

Table 2.1: Distribution of different anaphor types in PDT 2.0.

Figure 2.1 shows a sample t-tree in which coreference links are depicted. They form a coreferential chain corresponding to surface tokens *Novotná – své – jí* [Novotná – her (reflexive pronoun) – her (possessive pronoun)].

As the tectogrammatical structures are highly complex, there can be more than twenty attribute-value pairs associated with the individual nodes. The tree in the Figure 2.1 is displayed in a simplified fashion: the nodes are labeled only with tectogrammatical lemmas, functors, and semantic parts of speech. We present only a brief explanation of these attributes in the following paragraphs.

The first attribute is the tectogrammatical lemma, which stands either for the canonical word form of the word present in the surface sentence form or for the artificial value of a newly created node on the tectogrammatical layer. The (artificial) tectogrammatical lemma `#PersPron` stands

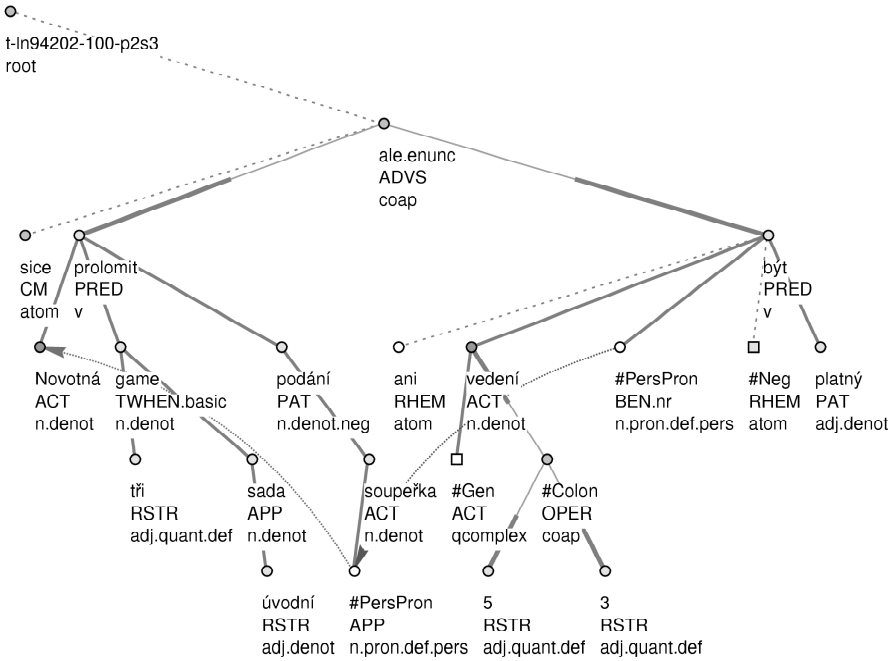

Figure 2.1: Simplified t-tree representing the sentence *Novotná sice prolomila ve třetím gamu úvodní sady podání své soupeřky, ale ani vedení 5:3 jí nebylo platné.* (Lit.: Novotná indeed broke through in the third game of the initial set serve of her opponent, but not the lead 5:3 her was efficient.)

for personal (and possessive) pronouns, be they expressed on the surface (i.e., present in the original sentence) or restored during the annotation of the tectogrammatical tree structure (zero pronouns).

The second attribute is the functor, which stands for the type of the edge leading from the node to its governor; the edge primarily represents a dependency relation (understood as a relation in the underlying structure of the sentence), or some well-specified technical phenomena. Following FGD, the dependency functors are divided into actants (ACT – actor, PAT – patient, ADDR – addressee, etc.) and free modifiers (LOC – location, BEN – benefactor, RHEM – rhematizer, TWHEN – temporal modifier, APP – appurtenance, etc.).

The third attribute displayed below the nodes is the semantic part-of-speech, representing categories of the tectogrammatical layer corresponding to basic onomasiologic categories and are not identical with the 'traditional' parts of speech. The main semantic parts of speech distinguished

in PDT 2.0 are: semantic nouns, semantic adjectives, semantic adverbs and semantic verbs. These basic sets are further subdivided. In the following list we present those subtypes of semantic nouns which most frequently appear as antecedent nodes (clearly, the value of `sempos` is helpful for selecting antecedent 'candidates'):

n.denot – denotative semantic noun,

n.denot.neg – denotative semantic noun with separately represented negation feature,

n.pron.def.demon – demonstrative definite pronominal semantic noun,

n.pron.def.pers – pronominal definite personal semantic noun,

n.pron.indef – indefinite pronominal semantic noun,

n.quant.def – quantification definite semantic noun.

Coreference links are displayed as arrows in the figure, pointing from an anaphor to its antecedent. In the tree editor `tred`[2] used for PDT 2.0 annotation, different arrow colors are used for distinguishing textual and grammatical coreference.

In the PDT 2.0 the data representation for coreferential chains differs from these described in [Kučová et al., 2003] and [Kučová and Hajičová, 2004]. Three completely new attributes are established for each anaphor:

coref_gram.rf – identifier or a list of identifiers of the antecedent(s) related via grammatical coreference

coref_text.rf – identifier or a list of identifiers of the antecedent(s) related via textual coreference

coref_special – values `segm` (segment) and `exoph` (exophora) standing for special types of textual coreference.

In the next stage of coreference annotation, which is being carried out on PDT 2.0 now, the textual coreference is extended to non-pronominal and non-zero NPs, and also to some cases of adjectives, numerals and adverbs. Together with the textual coreference, bridging relations of several types are being annotated. Discourse deixis is annotated separately for references to non-nominal entities and references to a discourse segment of more than one sentence.

In terms of the number of coreference links, PDT 2.0 is one of the largest existing manually annotated resources, which contains not only pronominal anaphora, noun phrase anaphora [3] and bridging anaphora, but also zero anaphora. Another comparably large resource is BBN Pronoun Coreference and Entity Type Corpus.

[2] http://ufal.mff.cuni.cz/~pajas/tred/

[3] We borrow the broadly used term "NP anaphora" even if there are no noun phrases (in the sense of phrase-structure grammar) annotated in the PDT. Where we use the term NP, we actually mean a subtree which has as its head a noun.

2.2.2 Extended Prague Dependency Treebank 2.0

Extended coreferential and bridging referring expressions

Unlike ACE, we do not restrict the annotation to a set of named entities (NE), and annotate all referential entities, also the abstract and generic ones. Thus the coreference annotation in PDT actually captures some kind of pragmatic references to the actual notions.

The extended coreferential and bridging relations are to be marked between elements of the following categories: full NPs (*Prague – the capital of the Czech Republic*), anaphoric adverbs (*the capital of the Czech Republic – there*), numerals (*1999 – this year*), clauses and sentences if coreferring with NPs (*[They tried to teach him to read]$_i$ – [The attempt]$_i$ was not successful*). Similarly to MUC, adjectives are annotated only if they are coreferential with a named entity or a nominal head, so e.g. we annotate pairs as *German – Germany*. Coordinated NPs and appositional structures are also potential markables, in the syntactic structure of the tectogrammatical trees, their roots (conjunctions or punctuation marks) technically serve as coreferring nodes (see [Mikulová et al., 2005]).

Names and other named entities are all subjects to annotation. A substring of a named entity, however, is not to be annotated if it is not a named entity itself. Thus, for the sequence *The Charles University in [Prague]$_i$... [Prague]$_i$ was...*, the two instances of *Prague* are to be marked coreferential; but in *Institute of Nuclear [Research]$_i$ ... nuclear [research]$_j$* the two instances of NP *research* are neither as coreferring nor marked as a bridging relation.

Contrary to MUC and ACE, predicate nominals are not considered to be coreferential with the subject, and neither the coreferential relation between appositional phrases is established.

Extended Textual Coreference

Extended textual coreference is marked between two elements that refer to the same object, notion or activity in the discourse. Each markable can only be the object of no more than one coreferential expression. Some exceptions to this rule for pronominal coreference [Kučová and Hajičová, 2004] are being corrected by the annotation of the extended textual coreference.

Textual coreference is further subclassified into two types: coreference of NPs with specific reference (`coref_text`, type 0) and relations between NPs with generic reference (`coref_text`, type NR). In contrast to other schemes (GNOME, ACE, etc.), in our scheme the feature of genericity is not assigned to all generic NPs. Nevertheless, we assume generic NPs to have other anaphoric properties in discourse, in addition they result in richer ambiguity and are the cause of lower inter-annotator agreement. These were the reasons to separate them into a special category of coreferential relations. Compare the following examples (all English examples are constructed in parallel to the corresponding original Czech ones):

(2.1) '[Mary]$_i$ and John went together to Israel, but [Mary]$_{i:coref\_text:0}$ had to return because of the illness.'

(2.2) '[A lion]$_i$ lives in a forest. I wrote my Ph.D. thesis about [this animal]$_{i:coref\_text:NR}$.'

We do not distinguish between coreference pairs with the same lemmas (*Mary – Mary*) from the cases in which the entities are synonymous, hyponymous/hyperonymous or are just different nominations of any other kind (*Germany – the state*, *Mary – she*, etc.). Using grammatical attributes of the tectogrammatical tree, this kind of information can be easily extracted automatically. Unlike [Potau, 2008], we do not annotate false positive links (lexically identical but noncoreferential NPs) as coreferential.

Special cases of textual coreference. Two special cases of (co)reference are being annotated in PDT.

First, the textual coreference covers the cases of endophoric **references to discourse segment of more than one sentence**, including also the cases, when the antecedent is understood by inferencing from a broader cotext. The pronominal coreference relations being already annotated in PDT 2.0, we add the links in which the anaphor is expressed by a full NP or an adverb.

(2.3) *Celní unie bude sice existovat na papíře ještě dalších dvanáct měsíců, ale v praxi dostanou vzájemné vztahy punc tvrdosti mezinárodního obchodu. Poroste administrativa. Jistotu [v tomto směru]$_{segm}$ dávají nejnovější kroky vlády SR.*

The custom union will formally function for twelve more months, but in fact the relations will be of a kind of international trade. The bureaucracy will go up. The latest steps of the Slovak government confirm [this direction]$_{segm}$.'

This kind of relation does not have (unlike [Recasens et al., 2007]) explicitly marked antecedent, it just shows the fact that the given anaphoric NP corefers with some discourse antecedent of more than one sentence. We consider this decision to be provisional and we plan to complete it later.

Second, a specifically marked **link for exophora** denotes that the referent is 'out' of the context, it is known only from the actual situation. In the same way as for segments, the new nominal and adverbial links are being added.

Bridging Relations

Bridging relations [Johnson-Laird and Wason, 1977] hold between two elements in which the second element is interpreted by an inferential process ('bridge') on the basis of the first one.

Unlike [Recasens et al., 2007], bridging relations in PDT are annotated only between nominal expressions, no verbs are considered as anaphors. Each node can only be an antecedent/anaphor for no more than one type of bridging relations.

Given that the marking of bridging relations is very useful for information extraction, question answering and other NLP tasks, we decided to annotate them in PDT. However, this is a very complicated and time-consuming task, which up to now did not show high enough evaluation

results. Also the sets of bridging relations vary in different annotation schemes (see the rich variety of types in [Johnson-Laird and Wason, 1977], seven types in MATE, and their reduction to three types (element, subset and poss in GNOME and VENEX; part-of, set membership and thematic in [Recasens et al., 2007], and part-of, set membership, and a converse relation in ARRAU).

In our project, we annotate two basic types that are widely agreed upon, and add four other types, which frequently occurred in the pilot annotation experiments and seem to be relatively reliably identifiable. The five subtypes of bridging relations in PDT are:

- **part-of** (prototypical example *room – ceiling*): This relation has two directions – the type PART_WHOLE is used for the case when the antecedent of the anaphoric NP corresponds to the whole of which the anaphor is a part (and WHOLE_PART for the opposite).
- **set subset/element of the set** (prototypical example participants - one of participants/some participants): This relation is two-directional with the types SUB_SET and SET_SUB.

In some cases, the distinction between *part-of* and *set subset* groups is quite problematic, so that the only reason to decide for the type of a bridging relation is the countability of corresponding nouns.

(2.4) *Revidoval [text Prezidentské adresy]*$_i$*. [Poslední věta]*$_{i:WHOLE\_PART/SET\_SUB}$*, kterou v životě napsal, zněla ...*

'He edited [the text of President's address]$_i$. [The last sentence]$_{i:WHOLE\_PART/SET\_SUB}$, which he wrote in his life, was...'

For the time being, the instruction for a resolution of such type of ambiguities is to annotate type PART only in clear cases of nonseparable parts.

- **object – individual function on this object** (prototypical example *government – prime minister*): This relation is two-directional with types P_FUNCT for the sequence object – function and FUNCT_P for the opposite.
- **coherence relevant discourse opposites** (type CONTRAST)

(2.5) '[People]$_i$ don't chew, it's [cows]$_{i:CONTRAST}$ who chew.'

The CONTRAST relation is not really bridging relation in a restricted sense, it could be rather labeled rhetorical or something like that. However, this kind of semantic dependence has a similar influence on the text cohesion as bridging ones. In addition, it supplements the similar kind of information in the topic-focus articulation annotation, where contrast topic is marked, and the currently annotated contrast on the discourse level [Mladová et al., 2008].

- **noncospecifying explicit anaphoric relation**: The anaphor is marked with a demonstrative, bridging type ANAF is used.

(2.6) *"[Duha]$_i$?" Kněz přiložil prst k [tomu slovu]$_{i:ANAF}$, aby nezapomněl, kde skončil.*

"'[Rainbow]$_i$?" The priest put the finger on [this word]$_{i:ANAF}$, so that he didn't forget, where he stopped.'

- **further underspecified group REST**

This type is used for capturing bridging references - potential candidates for a new group of bridging relations (e.g. *location – resident*, relations between relatives (*mother – son*, etc.), event - argument (*listening – listener*) and some other relations). The last type is not marked as a special group for its relatively rare occurrence in our corpus (as we do not mark verbs as bridging entities). If needed, this relation can be relatively easily extracted from the annotated data.

The participation on the text cohesion is considered to be important, so in ambiguous cases, those relations are annotated that are important for the text cohesion.

2.2.3 Prague Czech-English Dependency Treebank 2.0

The Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0) is a manually parsed Czech-English parallel corpus sized over 1.2 million running words in almost 50,000 sentences for each language. The English part contains the entire Penn Treebank-Wall Street Journal (WSJ) Section (Linguistic Data Consortium, 1999). The Czech part comprises Czech translations of all the Penn Treebank-WSJ texts. The corpus is 1:1 sentence-aligned.

The manual coreference annotation in PCEDT 2.0 captures the grammatical coreference and pronominal textual coreference in 65,598 coreference links in the Czech part and 63,736 in the English part. The pronominal anaphora annotation in the English part comes from the BBN Pronoun Coreference and Entity Type Corpus.

Chapter 3

Coreference Resolution in Foreign Approaches

This chapter outlines some of the methods that have been successfully used in coreference resolution. In early machine learning approaches, one of the most commonly applied methods is classification in which every pair of an anaphor and its potential antecedent candidate is identified as coreferential or not. However, by treating each pair separately, this technique loses valuable information from other candidates and in the end it gives lower results than ranking technique, in which the entire candidate set is considered at once. Finally, we introduce unsupervised methods, the advantage of which is that there is no requirement for enormous amounts of annotated training data for most domains and languages.

3.1 Decision Tree Algorithm

Decision tree algorithm uses a decision tree as a classifier model. In the tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications (depicted in Figure 3.1). Applying a decision tree algorithm for coreference resolution requires a set of features describing pairs of noun phrases and recasting the coreference problem as a classification task (e.g. [Aone and Bennett, 1995], [McCarthy and Lehnert, 1995], [Soon et al., 2001]). A noun phrase coreference system described by [Ng and Cardie, 2002a] extends the Soon et al. corpus-based approach.

Firstly, Ng and Cardie build a noun phrase coreference classifier using the C4.5 decision tree induction system. For a non-pronominal noun phrase, the closest non-pronominal preceding antecedent is selected to generate the positive training example. For pronouns, the closest preceding antecedent is selected. After training, texts are processed from left to right. Each noun phrase encountered is compared in turn to each preceding noun phrase from right to left. For each pair the coreference classifier returns a number between 0 and 1. Noun phrase pairs with class values

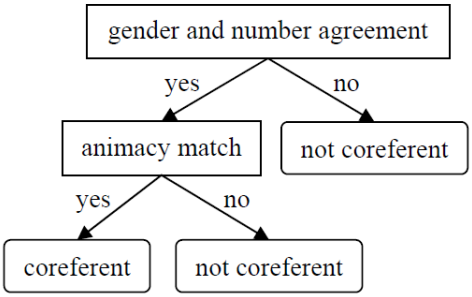

Figure 3.1: Simplified decision tree for coreference resolution.

above 0.5 are considered COREFERENT; otherwise the pair is considered NOT COREFERENT. The noun phrase with the highest coreference likelihood value from among preceding NPs with coreference class values above 0.5 is selected as the antecedent. The process terminates as soon as the antecedent is found or the beginning of the text is reached.

In the Ng and Cardie's coreference system a set of 53 features was proposed. The features were not derived empirically from the corpus, but were based on common-sense knowledge and linguistic intuitions regarding coreference. Surprisingly, the results using the full feature set are significantly low when compared with the results with a manual feature selection, with an eye toward eliminating low precision rules for common noun resolution F-measure of 70.4% on the MUC-6 coreference data sets and 63.4% on MUC-7.

3.2 A Twin-Candidate Model

The main idea of a twin-candidate model of [Yang et al., 2008] is to treat anaphora resolution as a preference classification problem. Firstly, the model learns a binary classifier that judges the preference between competing candidates of a given anaphor. Secondly, each candidate is compared with every other candidate by a preference classifier that can determine which one is preferred to be the antecedent. The final antecedent is identified based on the classified preference relationships among the candidates. Evaluating on the ACE data sets, Yang et al.'s twin-candidate model achieves the highest accuracy by 78.7% by using SVM for the first classifier and Round Robin for the second.

3.3 Specialized Models and Ranking

Denis and Baldridge's work [Denis and Baldridge, 2008] is based on the idea that training separate models that specialize in different types of anaphoric expressions and using a ranking loss function can perform better in comparison with standard joint classification approaches.

Specialized ranker models are created and evaluated on the ACE corpus for: (i) third person pronouns 82.2%, (ii) speech pronouns 66.9%, (iii) proper names 83.5%, (iv) definite descriptions 66.5%, (v) other types of phrases 63.6%.

3.4 Algorithm Based on the Bell Tree

[Luo et al., 2004] use the Bell tree to model the process of partitioning mentions into entities. A mention is defined as a referring expression, which can be all kinds of noun phrases, and the collection of mentions referring to the same object form an entity (by another name an equivalence class, used in the Cardie and Wagstaff's work [Cardie and Wagstaf, 1999]).

First, they traverse mentions in a document from beginning to end. The root node consists of a partial entity containing the first mention in the document. In each step of the algorithm, one mention is added by either linking to each of existing entities, or starting a new entity. A new layer of nodes is created to represent all possible coreference outcomes by adding one mention. The number of tree leaves is the number of possible coreference outcomes and it equals the Bell number [Bell, 1934].

The Bell Number $B(n)$ is the number of ways of partitioning n distinguishable objects (i.e., mentions) into non-empty disjoint subsets (i.e., entities).

$$B(n) = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}$$

Since the Bell number increases rapidly as the number of mentions increases, pruning is necessary. Thus, instead of finding the best leaf node Luo et al. look for the best path from the root to leaves in the Bell tree. The algorithm uses maximum entropy model [Berger et al., 1996] to rank paths and prunes any children with an insufficient score.

In the maximum entropy model a set of basic and composite features is selected. Composite features are generated by taking conjunction of basic features. Testing the algorithm on the MUC6 data Luo et al.'s system has 85.7% F-measure when using the official MUC scorer [Vilain et al., 1995a].

3.5 Clustering Approach

Cardie and Wagstaff's [Cardie and Wagstaf, 1999] unsupervised corpus-based clustering approach to the coreference task stems from the observation that each group of coreferent noun phrases defines an equivalence class (depicted in Figure 3.2). They start at the end of the document and compare each noun phrase to all preceding noun phrases. If the distance between two noun phrases is less than the clustering radius threshold r and their coreference equivalence classes are compatible, then the classes are merged. The distance between two noun phrases is measured by a feature's weight and incompatibility function for each feature from the NP feature set. The NP feature set consists of word, head noun, position, pronoun type, article, words-substring, appositive, number, proper name, semantic class, gender and animacy. The incompatibility function returns a value between 0 and 1.

$$dist(NP_i, NP_j) = \sum_{f \in F} w_f * incompatibility_f(NP_i, NP_j)$$

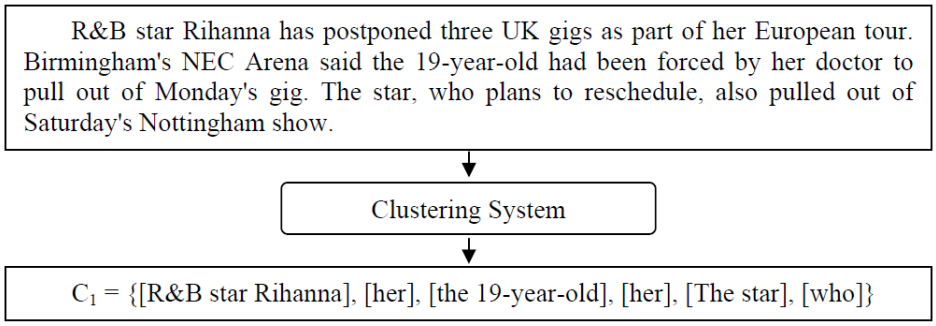

Figure 3.2: Coreference equivalence class in the sample text.

If two noun phrases do not match in number/proper names/class/gender/animacy feature, the distance between them gets a value of ∞ representing the incompatibility. Conversely, the appositive and words-substring terms with a weight of ∞ force coreference with compatible values.

In an evaluation on the MUC-6 coreference resolution corpus, Cardie and Wagstaff's clustering approach achieves the best F-measure of 53.6% with r = 4.

3.6 Nonparametric Bayesian Approach

[Haghighi and Klein, 2007] present an unsupervised, nonparametric Bayesian model that captures both within- and cross-document coreference. At the top, a hierarchical Dirichlet process captures

cross-document entity sharing. While at the bottom, a sequential salience model captures within-document sequential structure. They used Gibbs sampling and experiment performing on MUC-6 gave 70.3% MUC F1 measure.

3.7 Expectation Maximization Works

[Charniak and Elsner, 2009] propose an expectation-maximization algorithm for personal pronoun anaphora that learns virtually all of its parameters. The presented work is interesting in two ways. First, it is one of the few approaches that effectively use EM for NLP tasks. Secondly, their system is available on the web. In comparison with other unsupervised anaphora resolution systems [Cherry and Bergsma, 2005, Kehler et al., 2004, Haghighi and Klein, 2007], the Charniak and Elsner's classifies non-anaphoric pronouns jointly, handles first, second and third person pronouns as well as possessive and reflexive pronouns, and learns gender without an external database. The performance of the evaluated system on the dataset annotated by [Ge et al., 1998] is 68.6%.

Chapter 4

Coreference Resolution in Czech

4.1 Previous Work

[Kučová et al., 2003] presented a coreference annotation scheme for PDT. Within the annotation, a list of hand-written rules was created in order to resolve relative, reflexive and control coreference. They achieved a precision of 87.8%.

[Kučová and Žabokrtský, 2005] proposed a set of filters for personal pronominal anaphora resolution. The list of candidates was built from the preceding and the same sentence as the personal pronoun. After applying each filter, improbable candidates were cut off. If there was more than one candidate left at the end, the nearest one to the anaphor was chosen as its antecedent. The reported final success rate was 60.4 % (counted simply as the number of correctly predicted links divided by the number of pronoun anaphors in the test data section).

Some experiments with using C4.5 top-bottom decision trees or hand-written rules for all grammatical and pronominal textual coreference are described in [Nguỵ, 2006].

Another rule-based approach to pronominal textual coreference was presented in Nguỵ and Žabokrtský [Nguỵ and Žabokrtský, 2007]. Their rules are related to preferences and constraints. All antecedent candidates for the given anaphor, which have been filtered by gender and number agreement, are assigned a positive or negative score. The F-measure of their system is 74.2%.

4.2 Coreference Resolution for Third Person and Possessive Pronouns

In the following section we describe two works on coreference resolution for third person and possessive pronouns. One tries to automatically detect zero personal pronouns. The other builds two machine learning systems to resolve the antecedent identification for manually annotated overt and zero pronouns. It should be said that these two works are not yet joined into one system.

4.2.1 Anaphor Detection

In Czech, it is natural to drop out personal pronouns in the subject position of the clause. An overt subject pronoun indicates an emphasis of the speaker. In this section we discuss the case of an unexpressed subject identification problem, because an unexpressed implicit subject in the third person form is often an anaphor that refers to an entity already mentioned in the text.

In a subjectless finite verb clause we distinguish the following four types of unexpressed subjects:

Implicit subject : The subject is omitted in the surface text but can be understood from the verb morphological information; most often it stands for an entity already mentioned in the text or can be deictic.

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

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

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

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

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

Null subject : The 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.

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

We used the maximum entropy method to train a model for unexpressed subject classification and chose the data of the PDT 2.0 for the training and testing procedures. However, the corpus selection does not suit the task and we will discuss it later.

Resolution method

Maximum entropy was first introduced to Natural Language Processing (NLP) area by Berger et al. [Berger et al., 1996]. Since then, the maximum entropy principle has been used widely in NLP, e.g. for tagging, parsing, named entity recognition and machine translation. Maximum entropy models have the following form:

$$p(y|x) = \frac{1}{Z(x)} exp \sum_i \lambda_i f_i(x, y)$$

where f_i is a feature function, λ_i is its weight, and $Z(x)$ is the normalizing factor.

For our task, we chose a maximum entropy classifier, an implementation of Laye Suen[1] , a machine learning tool that takes data items and place them into one of k classes. In addition, it also gives probability distributions over classifications. Our approach can be described in the following steps:

1. In a training set, extract features from each finite verb without an overt subject;
2. Train a MaxEnt classifier with them;
3. Test the MaxEnt model on a test set;

Data description

At the tectogrammatical layer of PDT 2.0, the meaning of the sentence is represented as a dependency tree structure. In addition to nodes corresponding to surface tokens, there are newly established nodes the tectogrammatical lemma of which is an artificial t-lemma substitute beginning with #. Our focused unexpressed subjects can be found at t-layer among nodes with t-lemma `#PersPron`, `#Gen` and `#Unsp`; except null subjects, which were not reconstructed at t-layer. These t-lemma substitutes have the following meanings:

`#PersPron` t-lemma substitutes are assigned to:

- personal and possessive pronouns present in the surface sentence;
- zero pronouns representing the implicit subject;
- textual ellipsis - obligatory arguments of the governing verb / noun;

`#Gen` t-lemma substitutes are used for:

- grammatical ellipsis of an obligatory argument - general argument;
- zero pronouns representing the general subject;

[1] A Perl module `AI::MaxEntropy`, see http://search.cpan.org/perldoc?AI::MaxEntropy

`#Unsp` t-lemma substitutes stand for:

- grammatical ellipsis of an obligatory argument - unspecified Actor;
- zero pronouns representing the unspecified subject;

Feature extraction

Our maximum entropy classifier was trained on the basis of feature vectors for each finite verb (predicate) having no overt subject depending on it. The following features were used:

Categorial features : t-lemma, form, tense, gender, number, person, and:

- adverbial form – an adverb in the case of an 'adverbial' predicate (*to be + an adverb*)
- nominal form – a nominal part in the case of a nominal predicate

Binary features :

- `has_actor` – the considered predicate has an overt Actor
- `is_reflexive` – the predicate is reflexive
- `is_passive` – the predicate is a passive verb
- `has_o-ending` – the predicate is a finite verb ending with *o*
- `is_to-be-infin` – the predicate is in the construction of 'to be + infinitive'
- `has_dep-clause` – there is a dependent clause hanging on the verb

Concatenated features :

- `reflexive_o-ending` – concatenation of `is_reflexive` and `has_o-ending`
- `passive_o-ending` – concatenation of `is_passive` and `has_o-ending`
- `reflexive_person_number_gender` – concatenation of `is_reflexive`, person, number and gender
- `passive_person_number_gender` – concatenation of `is_passive`, person, number and gender

The feature selection relies on characteristics of each unexpressed subject type. A general subject often comes along with a third person singular reflexive verb or a third person singular passive verb. A reflexive verb can be easily recognized by a reflexive particle. A third person singular passive verb and a past tense third person singular reflexive verb always end with *o*. The case of a subject expressed by a dependent clause can be detected by the `has_dep-clause` feature. An adverbial form can indicate a null subject, e.g. *Je polojasno* ('It is somewhat cloudy').

Data problems

In PDT 2.0 we have to face several problems. The most crucial problem is the absence of the explicit annotation of unexpressed subjects we are interested in. In Figure 4.1 and Figure 4.2, we illustrate ambiguous cases, in which two nodes with `#PersPron` and `#Gen` appear.

We tried to solve the problem of missing manual unexpressed subject annotation by proposing some rules listed in Algorithm 1.

Algorithm 1: Manual unexpressed subject annotation.

if *verb has #Unsp among children* **then**
    It is the case of an unspecified subject.
else if *verb has generated #PersPron* **and** *#Gen.ACT among children* **then**
    if *verb has o-ending* **or** *is to-be-infin* **or** *is rflx_pass_by_active_present_3sg* **then**
        It is the case of a general subject.
    else
        It is the case of an implicit subject.
else if *verb has generated #PersPron.ACT among children* **then**
    It is the case of an implicit subject.
else if *verb has #Gen.ACT among children* **then**
    It is the case of a general subject.
else if *verb has generated #PersPron.ACT among children* **and** *(is passive* **or** *rflx_pass_not_active_present_3sg) with no o-ending* **then**
    It is the case of an implicit subject.
else
    It is the case of a null subject.

Another problem with the PDT 2.0 data is the absence of the manual annotation of person, number and gender. This information is very important for us because it indicates a general / null subject by a third person singular neuter / animate form or an unspecified subject by a third person plural animate form.

We have no rules that guarantee a 100% correct resolution for the identification of unexpressed subjects on annotated data of the PDT 2.0. In addition, we rely on the genre of the corpus, where proverbs with general subjects do not often occur, and suppose all cases with third person singular animate active verb to be an implicit subject; whereas all cases with third person singular neutrum passive or reflexive verb to be a general subject. We expect that the occurrence of singular neuter implicit subject is sporadic as well.

Baselines

Baselines for automatic identification of unexpressed subjects are described in Algorithms 2, 3, 4 and 5. Each of them was run separately.

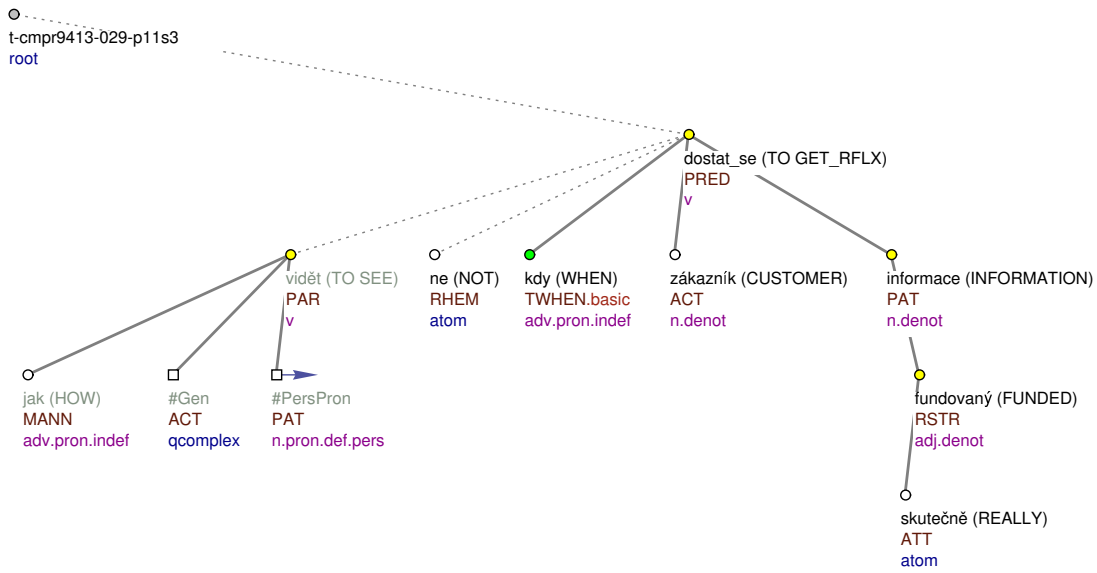

Figure 4.1: A simplified t-tree representing the sentence *Jak je vidět, ne vždy se zákazníkovi dostane skutečně fundovaných informací.* (Lit. How it's seen, not always RFLX customer gets really funded information.) In this case, the node with `#Gen` is considered to be the unexpressed general subject.

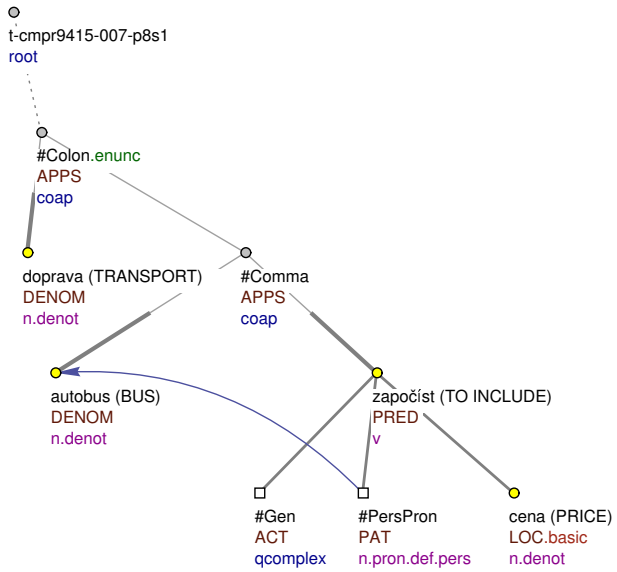

Figure 4.2: A simplified t-tree representing the sentence *Doprava: Autobus, je započten v ceně.* (Lit. Transport: Bus, is included in price.) In this case, the node with `#PersPron` is the unexpressed implicit subject.

Algorithm 2: Baseline for implicit subject identification.

if *a clause contains a finite verb* **then**
if *the verb has neither overt subject nor Actor depending on it* **and** *it has no o-ending* **and** *it is not a reflexive passive verb* **and** *it is not a passive verb with o-ending* **and** *its t-lemma is not an impersonal verb* **then**
There is an implicit subject.

Algorithm 3: Baseline for general subject identification.

if *a clause contains a finite verb* **then**
if *the verb has neither overt subject nor Actor depending on it* **and** *(it has o-ending* **or** *has a 'to be + infinitive' construction* **or** *it is a reflexive passive verb having an active present tense third person singular form)* **then**
There is a general subject.

Algorithm 4: Baseline for unspecified subject identification.

if *a clause contains a finite verb* **then**
if *the verb has no overt subject depending on it* **and** *has a third person animate plural form* **and** *(there is no preceding finite verb* **or** *the preceding finite verb has not a third person animate plural form* **or** *it has not a dependent animate plural noun with functor ACT/PAT/ADDR)* **then**
There is an unspecified subject.

Algorithm 5: Baseline for null subject identification.

if *a clause contains a finite verb* **then**
if *the verb has neither overt subject nor Actor depending on it* **then**
There is a null subject.

Evaluation and Discussion

If the problem of missing manual unexpressed subject annotation is considered to be 100% successfully resolved by proposed hand-written rules, then we obtain the results given in Table 4.1.

The poor result of unspecified subject identification can be explained for its rare occurrences in the data, the problem of missing manual person, gender and number annotation and the fact that it requires knowledge of a potential antecedent existence. If there is an antecedent to which the unexpressed subject can refer, then it is a case of an implicit subject; otherwise an unspecified subject. An anaphora resolution might help to improve this result.

	P	R	F
Implicit Baseline	95.4%	98.4%	96.9%
Implicit MaxEnt	90.6%	99.4%	94.8%
General Baseline	24.9%	87.2%	38.7%
General MaxEnt	96.7%	74.4%	84.1%
Unspecified Baseline	4.55%	3.45%	3.92%
Unspecified MaxEnt	0%	0%	0%
Null Baseline	98%	85.7%	91.5%
Null MaxEnt	82.5%	29.7%	43.7%

Table 4.1: Results for the unexpressed subject identification.

The result of null subject identification might be higher by adding a sophisticated list of impersonal / weather verbs / constructions as well. In general a deeper error analysis should bring overall improvements and explain the doubt of better baseline results.

4.2.2 Antecedent Identification

In this section we compare two Machine Learning approaches to the task of automatic antecedent identification for 3rd person and possessive pronouns: a conventional classification system based on C5.0 decision trees, and a novel perceptron-based ranker. The perceptron system achieves f-score 79.43% on recognizing coreference of personal and possessive pronouns, which clearly outperforms the classifier and which is the best result reported on PDT 2.0 data set so far.

Training data preparation

The training phase of both presented AR systems can be outlined as follows:

1. detect nodes which are anaphors,
2. for each anaphor a_i, collect the set of antecedent candidates $\mathrm{Cand}(a_i)$,
3. for each anaphor a_i, divide the set of candidates into positive instances (true antecedents) and negative instances,
4. for each pair of an anaphor a_i and an antecedent candidate $c_j \in \mathrm{Cand}(a_i)$, compute the feature vector $\Phi(c, a_i)$,

5. given the anaphors, their sets of antecedent candidates (with related feature vectors), and the division into positive and negative candidates, train the system for identifying the true antecedents among the candidates.

Steps 1-4 can be seen as training data preprocessing, and are very similar for both systems. System-specific details will be further described.

In the presented work, only third person personal (and possessive) pronouns are considered, be they expressed on the surface or reconstructed. We treat as anaphors all tectogrammatical nodes with lemma `#PersPron` and third person stored in the `gram/person` grammateme. More than 98 % of such nodes have their antecedents (in the sense of textual coreference) marked in the training data. Therefore we decided to rely only on this highly precise rule when detecting anaphors.

In both systems, the predicted antecedent of a given anaphor a_i is selected from an easy-to-compute set of antecedent candidates denoted as $\mathrm{Cand}(a_i)$. We limit the set of candidates to semantic nouns which are located either in the same sentence before the anaphor, or in the preceding sentence. Table 4.2 shows that if we disregard cataphoric and longer anaphoric links, we loose a chance for correct answer with only 6 % of anaphors.

Antecedent location	Percnt.
Previous sentence	37 %
Same sentence, preceding the anaphor	57 %
Same sentence, following the anaphor	5 %
Other	1 %

Table 4.2: Location of antecedents with respect to anaphors in the training section of PDT 2.0.

If the true antecedent of a_i is not present in $\mathrm{Cand}(\mathrm{a_i})$, no training instance is generated. If it is present, the sets of negative and positive instances are generated based on the anaphor. This preprocessing step differs for the two systems, because the classifier can be easily provided with more than one positive instance per anaphor, whereas the ranker can not.

In the classification-based system, all candidates belonging to the coreferential chain are marked as positive instances in the training data. The remaining candidates are marked as negative instances.

In the ranking-based system, the coreferential chain is followed from the anaphor to the nearest antecedent which itself is not an anaphor in grammatical coreference.[2] The first such node is put on the top of the training rank list, as it should be predicted as the winner (E.g., the nearest antecedent of the zero personal pronoun *he* in the Example A.1 is the relative pronoun *who*, however, it is

[2]Grammatical anaphors are skipped because they usually do not provide sufficient information (e.g., reflexive pronouns provide almost no cues at all). The classification approach does not require such adaptation – it is more robust against such lack of information as it treats the whole chain as positive instances.

a grammatical anaphor, so its antecedent *Brien* is chosen as the winner instead). All remaining (negative) candidates are added to the list, without any special ordering.

Feature extraction

Our model makes use of a wide range of features that are obtained not only from all three levels of PDT 2.0 but also from the Czech National Corpus and the EuroWordNet. Each training or testing instance is represented by a feature vector. The features describe the anaphor, its antecedent candidate and their relationship, as well as their contexts. All features are listed in Table A.1 in the Appendix.

When designing the feature set on personal pronouns, we take into account the fact that Czech personal pronouns stand for persons, animals and things, therefore they agree with their antecedents in many attributes and functions. Further we use the knowledge from the Lappin and Leass's algorithm [Lappin and Leass, 1994], the Mitkov's robust, knowledge-poor approach [Mitkov, 2002], and the theory of topic-focus articulation [Kučová et al., 2005]. We want to take utmost advantage of information from the antecedent's and anaphor's node on all three levels as well.

Distance: Numeric features capturing the distance between the anaphor and the candidate, measured by the number of sentences, clauses, tree nodes and candidates between them.

Morphological agreement: Categorial features created from the values of tectogrammatical gender and number[3] and from selected morphological categories from the positional tag[4] of the anaphor and of the candidate. In addition, there are features indicating the strict agreement between these pairs and features formed by concatenating the pair of values of the given attribute in the two nodes (e.g., `masc_neut`).

Agreement in dependency functions: Categorial features created from the values of tectogrammatical functor and analytical functor (with surface-syntactic values such as `Sb`, `Pred`, `Obj`) of the anaphor and of the candidate, their agreement and joint feature. There are two more features indicating whether the candidate/anaphor is an actant and whether the candidate/anaphor is a subject on the tectogrammatical level.[5]

[3] Sometimes gender and number are unknown, but we can identify the gender and number of e.g. relative or reflexive pronouns on the tectogrammatical level thanks to their antecedent.

[4] A positional tag from the morphological level is a string of 15 characters. Every positions encodes one morphological category using one character.

[5] A subject on the tectogrammatical level can be a node with the analytical functor `Sb` or with the tectogrammatical functor `Actor` in a clause without a subject.

Context: Categorial features describing the context of the anaphor and of the candidate:

- parent – tectogrammatical functor and the semantic POS of the effective parent[6] of the anaphor and the candidate, their agreement and joint feature; a feature indicating the agreement of both parents' tectogrammatical lemma and their joint feature; a joint feature of the pair of the tectogrammatical lemma of the candidate and the effective parent's lemma of the anaphor; and a feature indicating whether the candidate and the anaphor are siblings.[7]
- coordination – a feature that indicates whether the candidate is a member of a coordination and a feature indicating whether the anaphor is a possessive pronoun and is in the coordination with the candidate
- collocation – a feature indicating whether the candidate has appeared in the same collocation as the anaphor within the text[8] and a feature that indicates the collocation assumed from the Czech National Corpus.[9]
- boundness – features assigned on the basis of contextual boundness (available in the tectogrammatical trees) {contextually bound, contrastively contextually bound, or contextually non-bound}[10] for the anaphor and the candidate; their agreement and joint feature.
- frequency – 1 if the candidate is a denotative semantic noun and occurs more than once within the text; otherwise 0.

Semantics: Semantically oriented feature that indicates whether the candidate is a person name for the present and a set of 63 binary ontological attributes obtained from the EuroWordNet.[11] These attributes determine the positive or negative relation between the candidate's lemma and the semantic concepts.

Classifier-based system

Our classification approach uses C5.0, a successor of C4.5 [Quinlan, 1993], which is probably the most widely used program for inducing decision trees. Decision trees are used in many AR sys-

[6] The "true governor" in terms of dependency relations.

[7] Both have the same effective parent.

[8] If the anaphor's effective parent is a verb and the candidate is a denotative semantic noun and has appeared as a child of the same verb and has had the same functor as the anaphor.

[9] The probability of the candidate being a subject preceding the verb, which is the effective parent of the anaphor.

[10] Contextual boundness is a property of an expression (be it expressed or absent in the surface structure of the sentence) which determines whether the speaker (author) uses the expression as given (for the recipient), i.e. uniquely determined by the context.

[11] The Top Ontology used in EuroWordNet (EWN) contains the (structured) set of 63 basic semantic concepts like Place, Time, Human, Group, Living, etc. For the majority of English synsets (set of synonyms, the basic unit of EWN), the appropriate subset of these concepts are listed. Using the Inter Lingual Index that links the synsets of different languages, the set of relevant concepts can be found also for Czech lemmas.

tems such as [Aone and Bennett, 1995], [McCarthy and Lehnert, 1995], [Soon et al., 2001], and [Ng and Cardie, 2002a].

Our classifier-based system takes as input a set of feature vectors as previously described and their classifications (1 – true antecedent, 0 – non-antecedent) and produces a decision tree that is further used for classifying new pairs of candidate and anaphor.

The classifier antecedent selection algorithm works as follows. For each anaphor a_i, feature vectors $\Phi(c, a_i)$ are computed for all candidates $c \in \mathrm{Cand}(a_i)$ and passed to the trained decision tree. The candidate classified as positive is returned as the predicted antecedent. If there are more candidates classified as positive, the nearest one is chosen.

If no candidate is classified as positive, a system of handwritten fallback rules can be used. The fallback rules are the same rules as those used in the baseline system presented later.

Ranker-based system

In the ranker-based AR system, every training example is a pair (a_i, y_i), where a_i is the anaphoric expression and y_i is the true antecedent. Using the candidate extraction function Cand, we aim to rank the candidates so that the true antecedent would always be the first candidate on the list. The ranking is modeled by a linear model of the previously described features. According to the model, the antecedent $\hat{y}_i$ for an anaphoric expression a_i is found as:

$$\hat{y}_i = \underset{c \in \mathrm{Cand}(a_i)}{\mathrm{argmax}} \; \Phi(c, a_i) \cdot \overrightarrow{w}$$

The weights $\overrightarrow{w}$ of the linear model are trained using a modification of the averaged perceptron algorithm [Collins, 2002]. This is averaged perceptron learning with a modified loss function adapted to the ranking scenario. The loss function is tailored to the task of correctly ranking the true antecedent, the ranking of other candidates is irrelevant. The algorithm (without averaging the parameters) is listed as Algorithm 6. Note that the training instances where $y_i \notin \mathrm{Cand}(a_i)$ were excluded from the training.

Antecedent selection algorithm using a ranker: For each third person pronoun create a feature vector from the pronoun and the semantic noun preceding the pronoun and is in the same sentence or in the previous sentence. Use the trained ranking features weight model to get out the candidate's total weight. The candidate with the highest features weight is identified as the antecedent.

Baseline system

We have made some baseline rules for the task of AR and tested them on the PDT 2.0 evaluation test data. Their results are reported in Table 4.3. Baseline rules are following: For each third person pronoun, consider all semantic nouns which precede the pronoun and are not further than the previous sentence, and:

Algorithm 6: Modified perceptron algorithm for ranking. Φ is the feature extraction function, a_i is the anaphoric expression, y_i is the true antecedent.

input : N training examples (a_i, y_i),
number of iterations T
init : $\overrightarrow{w} \leftarrow \overrightarrow{0}$;
for $t \leftarrow 1$ **to** T, $i \leftarrow 1$ **to** N **do**
$\hat{y}_i \leftarrow \text{argmax}_{c \in \text{Cand}(a_i)} \Phi(c, a_i) \cdot \overrightarrow{w}$;
if $\hat{y}_i \neq y_i$ **then**
$\overrightarrow{w} = \overrightarrow{w} + \Phi(y_i, a_i) - \Phi(\hat{y}_i, a_i)$;
end
end
output: weights $\overrightarrow{w}$

- select the nearest one as its antecedent (BASE 1),
- select the nearest one which is a clause subject (BASE 2),
- select the nearest one which agrees in gender and number (BASE 3),
- select the nearest one which agrees in gender and number; if there is no such noun, choose the nearest clause subject; if no clause subject was found, choose the nearest noun (BASE 3+2+1).

Experimental results and discussion

Scores for all three systems (baseline, clasifier with and without fallback, ranker) are given in Table 4.3. Our baseline system based on the combination of three rules (BASE 3+2+1) reports results superior to the ones of the rule-based system described in [Kučová and Žabokrtský, 2005].

An interesting point of the classifier-based system lies in the comparison with the rule-based system of [Nguy and Žabokrtský, 2007]. Without the rule-based fallback (CLASS), the classifier falls behind the Nguy and Žabokrtský's system (74.2%), while it gives better results with the fallback (CLASS+3+2+1).

Overall, the ranker-based system (RANK) significantly outperforms all other AR systems for Czech with the f-score of 79.43%. Comparing with the model for third person pronouns of [Denis and Baldridge, 2008], which reports the f-score of 82.2%, our ranker is not so far behind. It is important to say that our system relies on manually annotated information and we solve the task of anaphora resolution for third person pronouns on the tectogrammatical level of the PDT 2.0. That means these pronouns are not only those expressed on the surface, but also artificially added (reconstructed) into the structure according to the principles of FGD.

Rule	P	R	F
BASE 1	17.82%	18.00%	17.90%
BASE 2	41.69%	42.06%	41.88%
BASE 3	59.00%	59.50%	59.24%
BASE 3+2+1	62.55%	63.03%	62.79%
CLASS	69.9%	70.44%	70.17%
CLASS+3+2+1	76.02%	76.60%	76.30%
RANK	79.13%	79.74%	**79.43%**

Table 4.3: Precision (P), Recall (R) and F-measure (F) results for the presented AR systems.

4.3 Coreference Resolution for Control

Anaphora resolution is widely studied for its important role in machine translation (MT). We believe that control as a subtype of anaphora can be helpful in MT as well. Consider the following English sentences and their translations into Czech:

(4.5) *Jan$_i$ řekl Marii$_j$, aby Ø$_j$ přišla.*
John told Mary, so that (she) came.
John$_i$ told Mary$_j$ [Ø$_j$ to come].

(4.6) *Marie$_i$ nesouhlasila, že Ø$_i$ přijde.*
Mary didn't agree, that (she) comes.
Mary$_i$ did not agree [Ø$_i$ to come].

(4.7) *Marie$_i$ nesnáší, když Jan$_j$ kouří.*
Mary hates, when John smokes.
Mary$_i$ hates John$_j$ [Ø$_j$ smoking].

The mentioned examples show that the controlled clause can be expressed in one language by an infinitive verb or a gerund verb, whereas in another language, it can be expressed only by a finite verb.

The terms: verb of control (control verb, governing verb), controller (C-er), controllee (C-ee)[12], are known from Chomsky's framework of Government and Binding [Chomsky, 1981]. In this work, we use Panevová's conception of Czech control [Panevová, 1996], in which control is understood in a broader way.

[12]In Example 4.5, the control verb is *told*, the dependent verb is *to come*, the controller is *Mary$_j$*, and the controllee is the covert argument *Ø$_j$*.

[Panevová, 1996] divides control into two groups: constructions with an infinitive and nominalized constructions. The infinitive group is further divided into subgroups according to the syntactic function of the infinitive and the argument type of the controller. The nominalized group consists of only subgroups according to the argument type of the controller with the nominalized verb with the function Patient.

[Panevová et al., 2002] also presents another classification of control constructions: a combination of control verb and dependent verb both of which can be nominalized. An example of a control construction that can be expressed in all mentioned categories is: 1. slíbit napsat dopis (to promise to write a letter), 2. slib napsat dopis (a promise to write a letter), 3. slíbit napsání dopisu (to promise writing of a letter), 4. slib napsání dopisu (a promise of writing of a letter).

In [Kučová et al., 2003] and [Mikulová et al., 2007], the control classification was extended by a new type of control - quasi-control. **Quasi-control** can be found within a complex (multi-word) predicate [Cinková and Kolářová-Řezníčková, 2004], where its verbal part and nominal part share some of their valency modifications. This sharing is called quasi-control.

(4.8) *Jan*$_{i:ACT}$ *poskytl* *Marii*$_{j:ADDR}$ *[Ø*$_{i:ACT}$ *ochranu* *Ø*$_{j:PAT}$*].*
John provided Mary protection .
John$_{i:ACT}$ provided [Ø$_{i:ACT}$ protection Ø$_{j:PAT}$] for Mary$_{j:ADDR}$.

In Example 4.8, *to provide protection* is a complex predicate formed by a semantically empty verb *to provide* and a noun carrying the main lexical meaning of the entire phrase *protection*[13]. The omitted argument Actor of the noun *protection* refers to the verb's Actor *John* and the noun's non-expressed Patient refers to the verb's Addressee *Mary*.

At the tectogrammatical layer of PDT 2.0, controllees are reconstructed as t-nodes with t-lemma `#Cor` and `#QCor` (quasi-controllees). See the example in Figure 4.3 (*lze zabránit* – it is possible to prevent, *vyjádřili přesvědčení* – expressed conviction).

Related Work

There are many types of anaphora which have been a focus of research in recent years. There belong studies of nominal and pronominal anaphora ([Charniak and Elsner, 2009], [Denis and Baldridge, 2008], [Yang et al., 2008]), bridging (indirect) anaphora ([Poesio et al., 2004a], [Vieira et al., 2006]), and zero anaphora ([Kong and Zhou, 2010], [Iida and Poesio, 2011]). Control as a subtype of zero anaphora was discussed and analysed in [Kučová et al., 2003] and [Nguy, 2006].

[Kučová et al., 2003] provided a rule set for some of control types: if the parent of an infinitive is a verb, then it is a control verb and the controllee refers to one of the control verb's arguments according to the list of control verbs. The list of control verbs was taken from the valency lexicon of Czech verbs VALLEX 1.0 and it includes only three types of control verbs: control verbs with

[13] Its synonymous one-word predicate is 'to protect'.

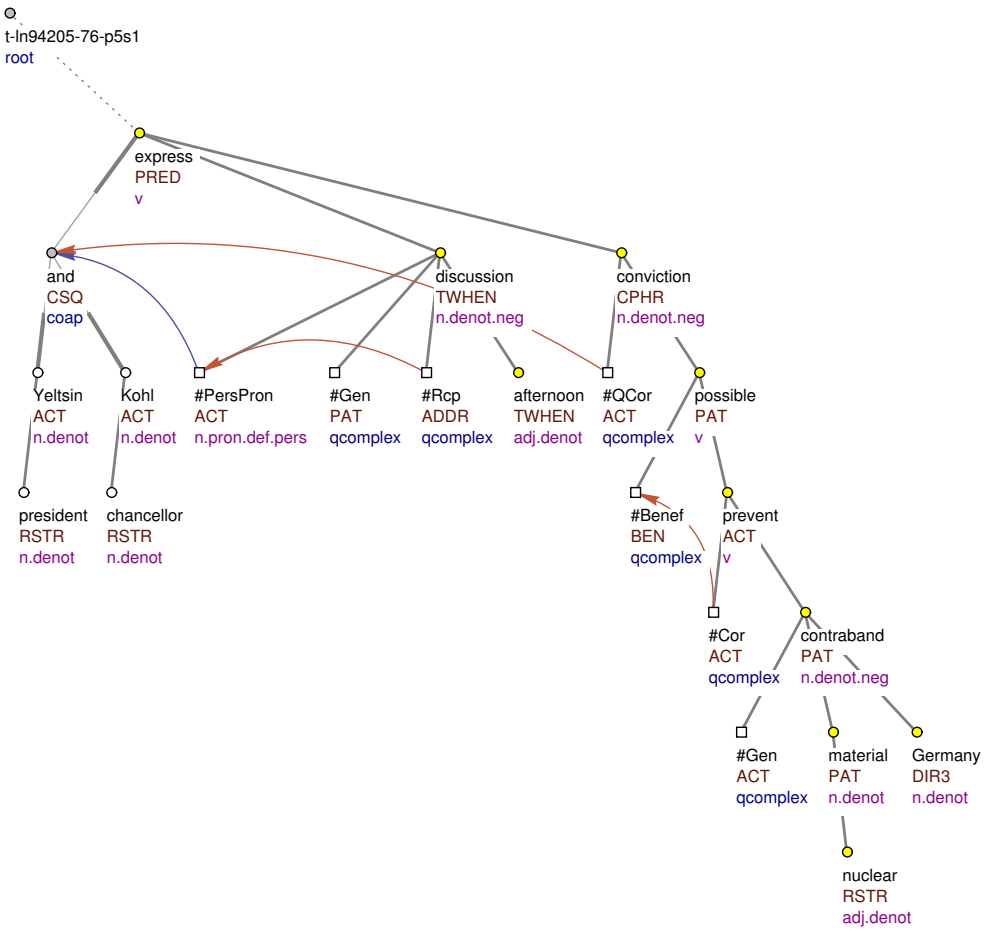

Figure 4.3: Simplified translated t-tree representing the sentence *Prezident Jelcin a kancléř Kohl vyjádřili po odpoledních jednáních přesvědčení, že lze zabránit pašování jaderného materiálu do Německa.* (Lit.: President Yeltsin and chancellor Kohl expressed after afternoon discussions the conviction, that it is possible to prevent from contraband of nuclear material to Germany.)

Actor / Addressee / Patient controller.[14] The reported success rate of the rules was the following: `ControlRuleACT` 69.93%; `ControlRuleADDR` 88.64% and `ControlRulePAT` 33.33%.

[Nguy, 2006] implemented a machine learning approach for the control coreference resolution,

[14] E.g. *doporučit*:ADDR - to urge someone$_i$:ADDR [Ø$_i$ to do something]; *snažit se*:ACT - someone$_i$:ACT to try [Ø$_i$ to do something]; *poslat*:PAT - to send someone$_i$:PAT [Ø$_i$ to do something]

but the features given for training a decision tree were gained mainly from the list of control verbs extended by verbal nominalizations. First a list of antecedent candidates was created. The list includes effective children of the controllee's effective grandparent (except the controllee's effective parent); in cases of constructions *to be resolved / able to do* effective children of controllees's great-grandparent; in cases of constructions *It's possible / necessary to do* effective children of nodes with t-lemma *možný / nutný / třeba*. Then, features of candidates were extracted. The features set was small, containing the candidate's t-lemma, functor, an option of only one possible candidate and the agreement of the candidate's and controllee's anaphor with the grandparent's category. Grandparent's categories are lists of control verbs and deverbal nouns. In addition to them there are also ambiguous control verb lists, i.e. verbs with controllers of different functors or where controller's and controllee's functors differ. The agreement of the candidate's and controllee's anaphor with the grandparent's category is then detected by 18 rules. Using the described features, Nguỵ trained a decision tree to decide whether a pair of controllee and antecedent candidate are coreferential. The success rate of her approach is 91.53%.

Control Resolution

Our control coreference resolution task consists of two subtasks: first we have to identify anaphors, in our case the controllees; after that the antecedents, in our case the controllers have to be detected. The resolution for the first subtask is based on the list of t-lemmas of the controllees' effective parent. The second subtask is resolved by using a perceptron-based ranker inspired by [Collins, 2002].

The controllee identification process relies on the creation of a list of dependent verbs (deverbal nouns) for controllees and quasi-controllees from the training data. The list contains pairs of a dependent verb (noun) lemma and a controllee's functor. There are two independent procedures for identifying controllees and quasi-controllees. The procedure for controllees works as follows: for each infinitive, reconstruct a controllee with the functor, which either was found from the extracted list by the infinitive's lemma or was filled with ACT.

In the case of quasi-controllees, the following simple rule was used: for each node with the functor CPHR[15] and a lemma from the extracted list, reconstruct one or more quasi-controllees with different functors according to the list.[16]

For the controller detection we use a simple scoring function: the optimal weight vector of which is estimated by averaged perceptron learning modified for ranking [Nguỵ et al., 2009]. The ranker is trained on the basis of feature vectors for a controllee and its possible antecedents. For every controllee a set of feature vectors containing only one positive instance and negative instances is formed. The positive instance includes features obtained from the controllee and its controller, whereas the negative ones are from the controllee and the non-coreferent phrase.

We consider three possible positions of the controller with respect to the controllee (Figure 4.4):

[15] CPHR is the functor of the nominal part of a complex predicate.

[16] See the Example 4.8, in which two quasi-controllees occur: one with ACT and another with PAT.

1. the controller is the controllee's *uncle* (the most frequent case)
2. the controller is the controllee's *cousin* (in cases of control constructions *It's possible / necessary to do*)
3. the controller is a sibling of the controllees' effective grandparent (in cases of complex control construction[17])

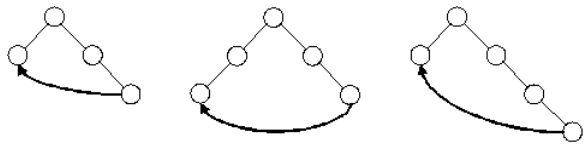

Figure 4.4: The tree representation of possible controllers' positions

The training features can be unary and related either to the controllee or to the candidate for the controller or to the controllee's effective parents (control verb and dependent verb), or they can be concatenated to represent the more complex relations between the controllee, the controller, the (complex) control verb (noun) and the dependent verb (noun). Altogether 30 features are used:

- Candidate (i): t-lemma, functor, tree position according to the controllee, semantic POS[18] (sempos), candidate's effective parent (ipar)'s t-lemma
- Controllee (j): t-lemma, functor
- Controllee's effective parent (jpar): t-lemma (lemma), functor (fun), sempos
- Controllee's effective grandparent (jpar2): t-lemma, functor, sempos
- Controllee's effective great-grandparent (jpar3): t-lemma, functor, sempos
- `Concatenate(ipar_lemma, i_lemma)`: concatenation of the t-lemma of the candidate's effective parent and the candidate's t-lemma
- `Concatenate(ipar_lemma, i_fun), Concatenate(jpar_lemma, i_fun, j_fun)`
- `Concatenate(jpar2_lemma, ipar_lemma), Concatenate(jpar2_lemma, i_fun)`,
- `Concatenate(jpar2_sempos, i_fun), Concatenate(jpar2_lemma, i_fun, j_fun)`
- `Concatenate(jpar2_lemma, jpar_lemma, i_fun, j_fun)`,
- `Concatenate(jpar2_lemma, jpar_sempos, i_fun)`,

[17] A complex control construction is a construction of a complex control verb (predicate) + a dependent verb.
[18] Semantic parts of speech correspond to the basic onomasiological categories.

- `Concatenate(jpar2_lemma, jpar_sempos, i_fun, j_fun)`,
- `Concatenate(jpar3_lemma, jpar2_lemma, i_fun, j_fun)`,
- `Concatenate(jpar3_lemma, jpar2_lemma, jpar_lemma, i_fun, j_fun)`
- `Concatenate(jpar2_lemma, ipar_lemma, jpar_lemma, i_fun, j_fun)`
- `Concatenate(jpar2_lemma, ipar_lemma, i_lemma, jpar_lemma, j_lemma, i_fun, j_fun)`

Evaluation and Discussion

We applied the following baseline rule for controller detection: for each controllee, select its *uncle* with functor `ACT` as its controller. The scores of rules for the controllee and quasi-controllee identification and the baseline rule and ranker for controller detection are given in Table 4.4.

	P	R	F
Cor.Ident.Rule	83.381%	86.222%	84.778%
QCor.Ident.Rule	86.219%	85.915%	86.067%
Coref.Baseline	56.065%	57.351%	56.701%
Coref.Ranker	82.161%	84.046%	83.093%

Table 4.4: Results for the control resolution.

The errors of controllee (Cor) identification arise in the following cases: dependent verb is nominalized (14.525%); Cor was not annotated; Cor was annotated with `#PersPron` or `#Gen` instead. The problem with quasi-controllee (QCor) identification was the recognition of its functor. If the correct recognition of QCor's functor is not in the task, then the f-measure is 96.075%.

The success rate of the automatic control coreference resolution depends on the previous subtask, the controllee identification. If the control coreference ranker is tested on golden trees (with manually annotated controllees), then it achieves the f-measure of 96.246% and outperforms the system of [Nguỵ, 2006]. The errors of the ranker occur when the controller is a verb or an adjective; or the controller is in another position than those given in Figure 4.4.

4.4 Coreference Resolution for Reciprocity

Syntactic reciprocity is an operation on the valency frames of verbs in which two verbal arguments are put into a symmetric relation as is illustrated by Example 4.9.

(4.9) *Jan a Marie se líbali.* = *Jan líbal Marii a (zároveň) Marie líbala Jana.*
Jan and Marie REFL kissed. = Jan kissed Marie and (simultaneously) Marie kissed Jan.

The primary means for syntactic reciprocity in Czech is the expression *se/si* combined with a coordination of subjects or with a subject expressed by a noun in plural (or a noun with a collective and similar meanings), where these noun phrases fill the role of both verbal arguments expected on the basis of verbal valency [Panevová, 1999, Panevová, 2007].

Syntactic reciprocity occurs also with the (deverbal) nouns and adjectives; compare:

(4.10) *boj znepřátelených stran mezi sebou*
fight of enemy sides between each other

(4.11) *lidé bojující mezi sebou navzájem*
people fighting among themselves each other

However, we do not take these cases into consideration in the present stage of research. [19]

At the tectogrammatical level of the Prague Dependency Treebank 2.0, syntactic reciprocity is represented by a newly established node with the `#Rcp` lemma that is inserted to the position of an unexpressed reciprocalized valency argument. The relation between the newly established node and the node in the expressed reciprocalized position is indicated as a relation of grammatical coreference.

Hand-written rules - baseline

Our heuristic procedure for identifying reciprocity occurrences works as follows:

1. First a list of all verbs, where reciprocity occurs, is created from training data.
2. The list is pruned: all words that appear less than twice, verbs with no *se/si* in the lemma are eliminated.
3. For all finite verbs that have a lemma from the list: If the current verb has no child with the preposition *s* [with] and one of the following conditions is true:
   (a) There is a reciprocity expression among verb's children (*navzájem, vzájemně* [each other]).
   (b) The subject of the current clause has a plural 'meaning':

[19] We developed our systems only on verbs, because experiments on nouns proved to be quite problematic. Reciprocity occurrences among adjectives in training data were rare.

(i) The subject is plural.

(ii) The subject is in a coordination.

(iii) The subject is a number or a quantitative noun (e.g. skupina, počet [group, number]).

(iv) The subject represents a human group (e.g. parlament, koalice [parliament, coalition]).[20]

(c) There is a prepositional phrase with the preposition *mezi* [between] among verb's children.

Then it is a reciprocity instance.

Improved hand-written rules

During the error analysis of hand-written rules described above, we have figured out that the verb list can be divided into different subgroups with specific attributes. The modified rules are:

For all finite verbs: If one of the following conditions is true:

1. There is a reciprocity expression among verb's children (*navzájem*, *vzájemně* [each other]).
2. The verb belongs to the `without s` verb list and has no child with preposition *s* (e.g. *dohodnout se, potkat se, hádat se* [agree on, meet, argue]).
3. The verb belongs to the `mezi+PAT` verb list and has a *mezi* prepositional phrase among verb's patients (e.g. *rozlišit* [distinguish]).
4. The verb belongs to the `plural PAT` verb list and has a plural patient (e.g. *sjednotit, sdružit* [unify, combine]).
5. The verb belongs to the pruned basic reciprocity verb list and one of the following conditions is true:

   (a) There is an expression *spolu* [together] among verb's children and the verb has *se/si* in the lemma .

   (b) There is a prepositional phrase with the preposition *mezi* [between] among verb's children.

   (c) There is a reflexive pronoun *se/si* among verb's children.

   (d) The subject of the current clause has a plural meaning.

   Then it is a reciprocity instance.

[20] We have created a list of words representing a human group from the EuroWordNet.

Maximum entropy classifier

For each finite verb we have created a feature vector with the following features:

- verb's lemma, form, tense, gender, number, person, sub-POS
- is passive?
- has a *s*-prepositional phrase among children?
- has a reflexive pronoun among children?
- has a *mezi*-prepositional phrase among Patients?
- has a reciprocity expression among children?
- has a subject with a plural meaning?
- has a Patient with a plural meaning?
- and concatenated features consisting of the verb's lemma and one `has a` condition

All instances are classified as `RCP`, if it is a reciprocity case, otherwise as `NONE`. Then they are used as an input for maximum entropy classifier training. We chose the implementation of Laye Suen.

Evaluation and Discussion

Using standard metrics, we have obtained results in Table 4.5.

	P	R	F
Baseline	75.76%	50%	60.24%
Rules	87.88%	58%	69.88%
MaxEnt	88%	44%	58,67%

Table 4.5: Results for the reciprocity resolution.

The slight different precisions from rule-based and MaxEnt-based approaches can be explained by the fact that reciprocity is a grammatical coreference. Therefore, a rule based method can give as high scores as a machine learning based method. We believe that the final results can be improved by expansion of the reciprocal verb lists.

4.5 Coreference Resolution for Noun Phrases

In this section, we make use of partial results coming from the annotation of extended anaphoric relations. Thus another motivation for our research was to help annotators to decide on coreference links with automatic pre-annotation of the data.

A substantial amount of newly annotated data is represented by so called noun phrase (NP) coreference, by which we mean coreference relations when the head of an expression in the later context – anaphor is a noun. This work focuses only on this type of coreference relations.

In this work almost all of the proposed features comes from a gold standard annotation. This decision is acceptable, if the coreference resolution system serves as an aid for annotators. However, if it becomes a part of end-to-end Natural Language Processing system, these features will have to be replaced by their counterparts obtained from morphological and syntactical analysis.

4.5.1 Extracted features

Features the resolver works with can be divided into the following categories:

Grammatical: These features are extracted from m-layer and consist of morphological tags of the anaphor and the antecedent, agreement in number, gender and negation. In addition, the t-layer supplies semantic functions of dependency relations, information about the presence of a determiner 'tento' ('this') and also a technical feature of being an apposition member.

Distance: How far the antecedent lies from its anaphor is a key attribute in coreference resolution. We measure it by a word and sentence distance.

Lexical: The most important component for lexical features is a lemma. We utilized features which indicate whether lemmas of the anaphor and the antecedent candidate are equal, particularly the ranking feature based on this property.[21]

We incorporated a dictionary of synonyms from a translation model extracted from the Czech-English Parallel Corpus [Bojar and Žabokrtský, 2009]. This dictionary served as a basis of synonymy feature.

Looking at the data, we noted that the entities which are frequent in a document are more likely to appear again. Hence we introduced a ranking feature denoting the number of occurrences of the particular word in the text.

Another set of lexical features relates to named entities. We introduced a simple feature indicating whether the first letter of the lemma is upper-cased. Apart from this, we exploited the information about possible named entity types stored on the m-layer of PDT. However, for future work, we see a possible improvement in complying the findings of [Denis and Baldridge, 2008] and training a special model for coreference with a proper noun anaphor.

[21] Ranking features assign positive integers to candidates, which meet some condition (e.g. lemma equality), in a way that the antecedent candidate closest to the anaphor obtains 1, the second closest one gets 2, etc. If the condition does not hold, the feature is undefined.

All features that we have introduced so far are describing only heads of either anaphor or antecedent candidates. They ignore nodes depending on the noun which is the head of the given NP. Therefore we suggested several tree features which involve all nodes belonging to the NP subtree. For instance, we included a ranking feature indicating the equality of whole phrases. We also designed features that compare the number of dependent nodes of both participants (if their head lemmas are identical), or the number of dependent nodes that are common for them.

It is necessary to emphasize that except for synonymy approximation, all features originate from PDT annotation which is manual gold standard.

From the list of weights, the learning method assigned to features, we noticed that some rarely distributed features obtained relatively high weights. For this reason we decided to incorporate feature pruning in this work. The extent to which features are cut off is determined by a parameter σ. For each multi-value feature we sorted its values by the number of occurrences and merged those least frequent values which in sum account for the proportion of at most σ.

4.5.2 Data preparation for machine learning

Annotation of extended anaphoric relations in PDT [Nedoluzhko et al., 2009] is an ongoing project, which aims to enrich PDT with remaining coreference and bridging relations. The data resulting from this project are not yet published, since the process of annotation is not completed yet (extended anaphoric relations are planned to be a part of the next version of PDT).

Whereas in corpora MUC-7 [MUC-7, 1998] and ACE [NIST, 2007], which are extensively used for coreference resolution systems for English, the coreference is annotated on the surface level between NP chunks of words, in PDT it is labeled on the t-layer between heads of subtrees (see Figure 4.5). An advantage of its annotation on the t-layer is in the presence of surface-dropped words and availability of rich linguistic features, with many of them being related to semantics. This provides more information to decide on coreference links.

Although PDT is already divided into training, development and evaluation set, it is not completely covered with NP coreference annotation. Therefore, we had to make our own partitioning of available data. The number of instances and the proportion of coreferential links in the data is sketched in Table 4.6.

	train				dev		eval	
	complete		reduced					
all	98,053		16,384		25,784		21,467	
coreferential	13,790	14.1%	2,694	16.4%	3,781	14.7%	3,148	14.7%

Table 4.6: Number of NP coreference links in data sets used during experiments. Reduced train set represents the data the final model was built from.

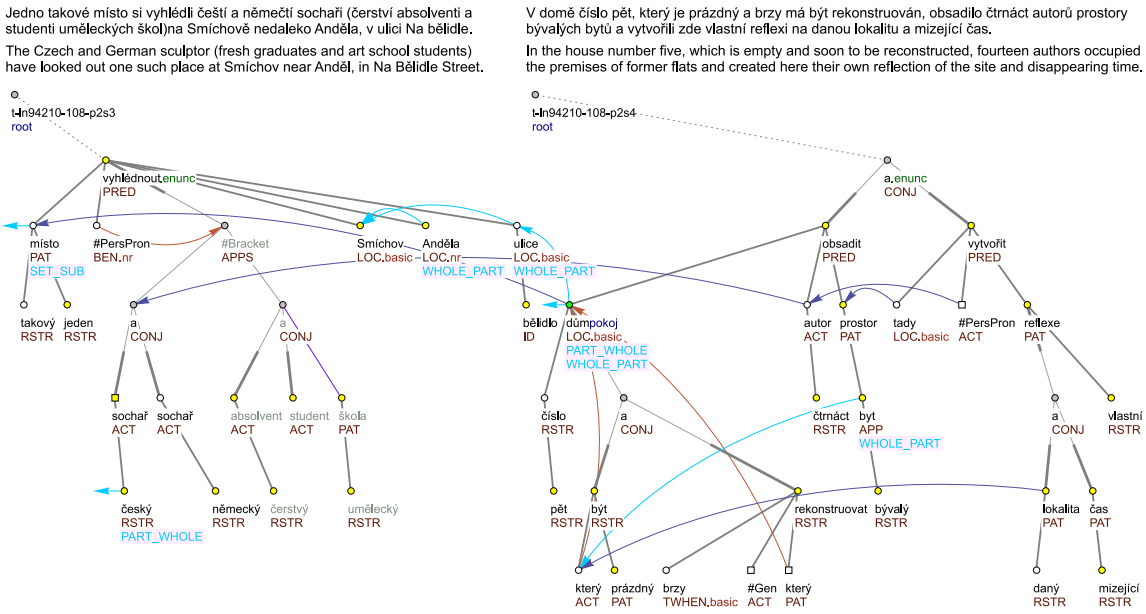

Figure 4.5: Example of a tectogrammatical representation of two sentences interlinked with various types of anaphora.

As it is the dominating practice, we treat recognition of individual coreference links as separated task instances. One instance consists of an anaphor candidate a and a set of its antecedent candidates c_i, out of which exactly one antecedent should be chosen by a Machine Learning technique. For this purpose, a rich set of features is provided for each pair $\langle a, c_i \rangle$. Following [Rahman and Ng, 2009], we join anaphoricity determination and antecedent selection into a single step. For this purpose, a is artificially included into the set of antecedent candidates. If a is non-coreferential, then a is supposed to be chosen from the antecedent candidate set, which is interpreted as an absence of any coreference link leading from the given anaphor candidate.

Since we are interested merely in NP coreference, we constrained anaphors to be subtrees with a noun head. Because pronouns do not carry a sufficient amount of information to be matched with an NP anaphor, we restricted antecedent heads to be nouns as well.[22] After such filtering noun-to-pronoun links are omitted. Hence, if the head of the closest true antecedent is not a noun, we follow the coreferential chain in order to find the noun antecedent. If such a node is found, it is marked as a true antecedent, otherwise the anaphor candidate is assigned to be non-anaphoric.

Selecting the proper window size determines how many antecedent candidates will be under consideration. To avoid the computational complexity we decided to collect candidates for training

[22] Noun phrases account for 72% of antecedents.

from the sentence where the anaphor lies[23] and previous 10 sentences. Such choice covers 97% of antecedents. For the testing data there is no need for such a restriction so we use a much larger window: 200 previous sentences.

4.5.3 Training and resolving

Data, preprocessed in the way described above, served as an input for modeling by means of various machine learning techniques. We decided to compare two ranking approaches based on different learning methods – maximum entropy (ME) and perceptron. Although in previous works it has been already shown that rankers are more suitable for coreference resolution than classifiers, we wanted to confirm that a performance drop of classifiers appears also for our specific task of Czech NP coreference resolution. In the following we briefly describe the learning methods that we incorporated.

Maximum entropy (ME) classifier

Having pairs of an anaphor and an antecedent candidate $\langle a, c_i \rangle$, classifiers tackle each pair separately. Every such pair carry a label, whether it is coreferential (COREF) or not. Coreference modeling is conceived as a learning how likely it is for the pair, described by a given feature vector f_j, that a class COREF is assigned to it. These probabilities are modeled by maximum entropy and in the stage of resolution calculated for every anaphor a and corresponding candidates c_i with the following formula:

$$P\left(\text{COREF}|\left\langle a, c_i \right\rangle\right) = \frac{\exp\left(\sum_{j=1}^{n} \lambda_j f_j\left(\left\langle a, c_i \right\rangle, \text{COREF}\right)\right)}{\sum_c \exp\left(\sum_{j=1}^{n} \lambda_j f_j\left(\left\langle a, c_i \right\rangle, c\right)\right)}$$

Among the candidates, whose probability of being coreferential is greater than 0.5, the one closest to the anaphor is picked as the antecedent (closest-first strategy [Soon et al., 2001]). For maximum entropy modeling we employed a Perl library from CPAN `AI::MaxEntropy`, specifically the L-BFGS algorithm [Dong C. Liu and Jorge Nocedal, 1989] for estimating parameters.

Maximum entropy ranker

In contrast to the classifier, a ranker takes into account all candidates at once. In this case, the maximum entropy model itself includes a competition between individual candidates, thus there is no need for an additional step to single out an antecedent, as it is in the case of classification. That

[23] I.e. those words that precede the anaphor.

candidate is denoted as an antecedent for which the following probability is maximum:

$$P\left(c_i|a\right) = \frac{\exp\left(\sum_{j=1}^{n} \lambda_j f_j\left(a, c_i\right)\right)}{\sum_k \exp\left(\sum_{j=1}^{n} \lambda_j f_j\left(a, c_i\right)\right)}$$

We used an implementation of maximum entropy ranker from the Toolkit for Advanced Discriminative Modeling[24] [Malouf, 2002], which was already employed for English pronominal coreference resolution in [Denis and Baldridge, 2007a]. Parameters were estimated with a limited memory variable metric algorithm, closely resembling the L-BFGS algorithm, which we adopted for the classifier.

Perceptron ranker

This method follows the ranking scenario as in the previous case. Nonetheless, instead of maximum entropy, it provides a modeling by a perceptron. In order to pick an antecedent, perceptron model does not work with probabilities, though maximizing of dot product of weights and a feature vector remains the same as in the case of ME ranker.

The main difference lies in the algorithm used for estimating parameters. We reused the perceptron ranker, which successfully served as a modeling method for the system for Czech pronominal coreference resolution [Nguỵ et al., 2009]. Parameters were estimated using an averaged perceptron algorithm [Collins, 2002] with a modified loss function tailored to the ranking approach.

4.5.4 Evaluation and model analysis

During development experiments we discovered several facts. Although available training data contained almost 100,000 instances, we noticed in the preliminary tests that the ME as well as perceptron ranking models built just from 16,384 instances perform superior to models trained on full number of instances. Due to better performance and also in order to compare learning methods on the same data, we adopted this training subset for creation of all computational models involved in final evaluation tests.

Moreover, training a model with the maximum entropy classifier turned out to be much more time-consuming than with the other methods. This time complexity led us to omit all additional experiments on this model except for the final evaluation, having left the pruning parameter σ equal to that used with the ME ranker.

Obviously, we had to find proper values of the pruning parameter σ before we proceeded to the final evaluation. The tuning was performed on the development set. Figure 4.6 shows the highest F-scores for the ME ranker (44.11%) and the perceptron ranker (44.52%) achieved by models pruned with $\sigma = 0.09$ and $\sigma = 0.15$, respectively. These values were used for final tests on the evaluation set.

[24] http://tadm.sourceforge.net/

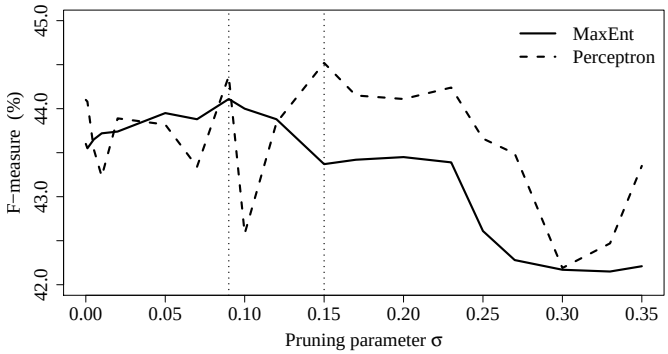

Figure 4.6: Values of F-score on the development data while changing the pruning parameter σ.

Method	Precision	Recall	F-Measure
MaxEnt classifier	57.30%	33.54%	42.32%
MaxEnt ranker	58.55%	35.58%	44.26%
Perceptron ranker	42.39%	46.54%	44.37%
Baseline	26.29%	60.01%	36.56%
Inter-annotator agreement	—	—	68.00%

Table 4.7: Performance of trained models compared with a baseline and inter-annotator agreement.

We assessed the quality of the proposed NP coreference resolution system on the evaluation set described in Section 4.5.2. As a baseline we set the result of a simple resolver, which for each anaphor candidate picks as its antecedent the closest candidate from the window with a lemma equal to the anaphor‘s lemma. If there is none, it is non-coreferential. We specified the upper bound as an inter-annotator agreement measured in [Nedoluzhko et al., 2009] on the subset from extended PDT similar to that we used. Performance of various models compared to lower and upper bound can be seen in Table 4.7.

All three machine learning approaches outperformed the baseline. The ranking approach proved to be more suitable for the task of coreference resolution than the classification one. There is no significant difference between F-values of the two ranking approaches. However, if the coreference resolution system is to be used as an aid for annotators, high values of precision are preferred. From this point of view, maximum entropy ranker performs better than perceptron ranker.

Except for the final evaluation we were interested how models deal with quantitative and qualitative changes. Since the annotation of the data we exploited is not finished, findings on the former can give us information, whether it is worth going on in the annotation process. The latter will

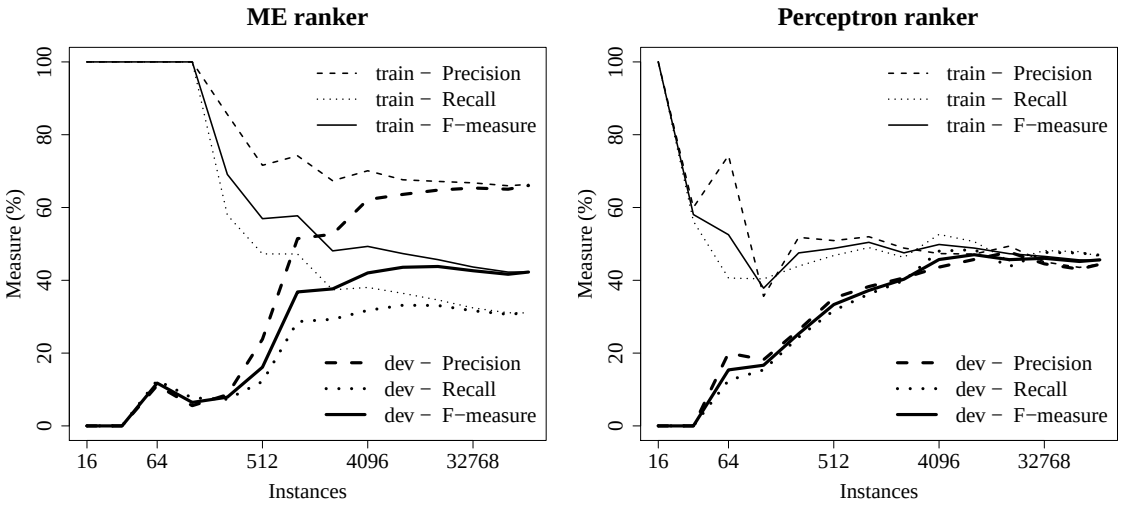

Figure 4.7: Learning curves show how the ranking models perform on the training and development set with various sizes of training data.

elaborate on how valuable are the novel features which exploit a tree structure of sentences in PDT.

To show the impact of changes in quantity we examined how model accuracy was changing, when built from different amounts of data. Sizes of the training data ranged along the logarithmic scale from 2^4 to the full size of training set.[25] These models were tested on the data, whose size accounted for $1/8$ of the training data size and the size of the complete development data for limited and full training sets, respectively. Furthermore, we carried out testing of models on the training data they were created from.

Resulting learning curves of the ME and perceptron rankers depicted in Figure 4.7 show averaged values after performing 9-fold cross validation.[26] Looking at the graph, we can observe three trends. The first is a convergence of success rate performed on seen and unseen data. Second, with amount of the training data growing over 5000 instances the quality of the computational model remains more or less the same. Lastly, while two learning approaches we investigated exhibit comparable F-scores, precision and recall behaves in a different way. ME ranker achieves about 25% better values of precision than recall. Conversely, these statistics are bound around the same value for perceptron ranker.

[25] It corresponds to less than 2^{17} as we can see in Table 4.6.

[26] N-fold cross validation requires the testing segments to be mutually disjoint for every two folds. In our case, this holds except for the full data, where we allowed over-lapping. The reason is simple arithmetic that for $n = 9$ this condition cannot be fulfilled.

To show qualitative influence of tree features we tweaked the final model by adding or leaving them out. If a feature was present in the final model, its removal would negatively affect the result. On the other hand, potential inclusion of a feature omitted from the final model would not improve the score. We analyzed the differences in F-score between the final and tweaked model.

In Table 4.8 we can see which features were included into and which excluded from the final model. We observe that influence of these features is up to 0.75%. The most valuable features are those, which capture an equality of the anaphor's and antecedent candidate's lemmas (`desc_self_equal_rank` and `desc_counts_equal`).

Final feature set		44.11%
Included		
`desc_self_equal_rank`	ranking feature of `desc_self_equal`	+0.74%
`desc_counts_equal`	equality of numbers of dependent nodes for identical lemmas	+0.40%
`anaph_this_attr`	is the determiner 'tento' a descendant of the anaphor head	+0.29%
`both_functors`	concatenation of semantic functions	+0.28%
`anaph_functor`	semantic function of the anaphor	+0.04%
`ante_functor`	semantic function of the antecedent	+0.03%
Excluded		
`desc_self_equal`	equality of whole NPs	0.00%
`desc_counts_zero`	`desc_counts_equal` with zero dependent nodes	-0.05%
`common_desc_lemmas_count`	number of words in common between NPs	-0.17%

Table 4.8: List of tree features and their influence on the final model.

Chapter 5

Conclusion

In this report we summarized results of our research on coreference resolution based on the Prague Dependency Treebank. We experimented with different techniques for different subtasks of coreference resolution: anaphoric person pronoun detection, pronominal anaphora resolution and coreference of deletions - the cases of control and reciprocity.

We developed a scheme for annotation of extended textual coreference and bridging relations. We carried out first experiments on manually annotated data with noun phrase anaphora, in which different machine learning methods were used.

In the future, we plan to re-run the experiments using data annotated by automatic tools (all needed tools are available in the TectoMT software framework [Žabokrtský et al., 2008]) instead of golden data set. We hope the integrated part of coreference resolution system will lead to a real improvement in machine translation.

Besides C5.0 and perceptron, we want to use also other classifiers (especially Support Vector Machine, which is often employed in AR experiments, e.g. by [Ng, 2005] and [Yang et al., 2006]), and extend the feature set. Both of these steps are expected to positively influence the AR system performance.

Finally, we would like to apply our AR system on English data of the Prague Czech-English Dependency Treebank. It will be interesting to see how coreference resolution for these two languages differs.

Bibliography

[Agirre et al., 2009] Agirre, E., Alfonseca, E., Hall, K., Kravalová, J., Pasca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *HLT-NAACL*, pages 19–27.

[Aone and Bennett, 1995] Aone, G. and Bennett, S. W. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. In *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 122–129.

[Bagga and Baldwin, 1998] Bagga, A. and Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *COLING-ACL*, pages 79–85.

[Bell, 1934] Bell, E. (1934). Exponential numbers. *The American Mathematical Monthly*, 41(7):411–419.

[Berger et al., 1996] Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22:39–71.

[Bojar and Žabokrtský, 2009] Bojar, O. and Žabokrtský, Z. (2009). CzEng 0.9, Building a Large Czech-English Automatic Parallel Treebank. *The Prague Bulletin of Mathematical Linguistics*, (92):63–83.

[Bojar et al., 2009] Bojar, O., Žabokrtský, Z., Janíček, M., Klimeš, V., Kravalová, J., Mareček, D., Novák, V., Popel, M., and Ptáček, J. (2009). Czeng 0.9.

[Cardie and Wagstaf, 1999] Cardie, C. and Wagstaf, K. (1999). Noun phrase coreference as clustering. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC1999)*, pages 82–89, College Park, Maryland, USA.

[Charniak and Elsner, 2009] Charniak, E. and Elsner, M. (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. Association for Computational Linguistics.

[Cherry and Bergsma, 2005] Cherry, C. and Bergsma, S. (2005). An expectation maximization approach to pronoun resolution. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL2005)*, pages 88–95, Ann Arbor, Michigan, USA.

[Chomsky, 1981] Chomsky, N. (1981). *Lectures on Government and Binding*, volume 9. Foris.

[Cinková and Kolářová-Řezníčková, 2004] Cinková, S. and Kolářová-Řezníčková, V. (2004). Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank. In *Korpusy a korpusová lingvistika v zahraničí a na Slovensku*.

[CNC, 2005] CNC (2005). Czech national corpus – SYN2005.

[Collins, 2002] Collins, M. (2002). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of EMNLP*, volume 10, pages 1–8.

[Denis and Baldridge, 2007a] Denis, P. and Baldridge, J. (2007a). A Ranking Approach to Pronoun Resolution. In *IJCAI*, pages 1588–1593.

[Denis and Baldridge, 2007b] Denis, P. and Baldridge, J. (2007b). Joint determination of anaphoricity and coreference resolution using integer programming. In *HLT-NAACL*, pages 236–243.

[Denis and Baldridge, 2008] Denis, P. and Baldridge, J. (2008). Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP2008)*, pages 660–669, Honolulu, Hawaii, USA.

[Doddington et al., 2004] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The automatic content extraction (ace) program — tasks, data, and evaluation. *Evaluation*, pages 837–840.

[Dong C. Liu and Jorge Nocedal, 1989] Dong C. Liu and Jorge Nocedal (1989). On the Limited Memory BFGS Method for Large Scale Optimization. *Mathematical Programming*, 45:503–528.

[Ge et al., 1998] Ge, N., Hale, J., and Charniak, E. (1998). A statistical approach to anaphora resolution. In *Proceedings of the 6th Workshop on Very Large Corpora (WVLC-6)*, pages 161–170, Montreal, Quebec, Canada.

[Haghighi and Klein, 2007] Haghighi, A. and Klein, D. (2007). Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855, Prague, Czech Republic. Association for Computational Linguistics.

[Haghighi and Klein, 2009] Haghighi, A. and Klein, D. (2009). Simple Coreference Resolution with Rich Syntactic and Semantic Features. In *EMNLP*, pages 1152–1161.

[Haghighi and Klein, 2010] Haghighi, A. and Klein, D. (2010). Coreference Resolution in a Modular, Entity-Centered Model. In *HLT-NAACL*, pages 385–393.

[Hana et al., 2005] Hana, J., Zeman, D., Hajič, J., Hanová, H., Hladká, B., and Jeřábek, E. (2005). Manual for morphological annotation, revision for the Prague Dependency Treebank 2.0. Technical Report TR-2005-27, ÚFAL MFF UK, Praha, Czechia.

[Iida and Poesio, 2011] Iida, R. and Poesio, M. (2011). A cross-lingual ilp solution to zero anaphora resolution. In *ACL*, pages 804–813.

[Jan Hajič, et al., 2006] Jan Hajič, et al. (2006). Prague dependency treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.

[Johnson-Laird and Wason, 1977] Johnson-Laird, P. N. and Wason, P. C. (1977). *Thinking: Readings in Cognitive Science*. Cambridge University Press, New York, NY, USA.

[Kehler et al., 2004] Kehler, A., Appelt, D., Taylor, L., and Simma, A. (2004). Competitive self-trained pronoun interpretation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Kong and Zhou, 2010] Kong, F. and Zhou, G. (2010). A tree kernel-based unified framework for chinese zero anaphora resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 882–891, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Krasavina and Chiarcos, 2007] Krasavina, O. and Chiarcos, C. (2007). Pocos potsdam coreference scheme. In *LAW '07 Proceedings of the Linguistic Annotation Workshop*.

[Kučová and Hajičová, 2004] Kučová, L. and Hajičová, E. (2004). Coreferential relations in the Prague Dependency Treebank. In *Proceedings of DAARC2004*, pages 97–102.

[Kučová et al., 2003] Kučová, L., Kolářová, V., Žabokrtský, Z., Pajas, P., and Čulo, O. (2003). Anotování koreference v pražském závislostním korpusu. Technical Report TR-2003-19, ÚFAL MFF UK, Prague, Prague.

[Kučová et al., 2005] Kučová, L., Veselá, K., Hajičová, E., and Havelka, J. (2005). Topic-focus articulation and anaphoric relations: A corpus based probe. In Heusinger, K. and Umbach, C., editors, *Proceedings of Discourse Domains and Information Structure workshop*, pages 37–46, Edinburgh, Scotland, UK, Aug. 8-12.

[Kučová and Žabokrtský, 2005] Kučová, L. and Žabokrtský, Z. (2005). Anaphora in Czech: Large Data and Experiments with Automatic Anaphora. In *LNCS/Lecture Notes in Artificial Intelligence/Proceedings of Text, Speech and Dialogue*. Springer Verlag Heidelberg.

[Lappin and Leass, 1994] Lappin, S. and Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

[Luo, 2005] Luo, X. (2005). On coreference resolution performance metrics. In *HLT/EMNLP*.

[Luo et al., 2004] Luo, X., Ittycheriah, A., Jing, H., Kambhatla, A., and Roukos, S. (2004). A Mention-Synchronous Coreference Resolution Algorithm Based on the Bell Tree. In *Proc. of the ACL*, pages 135–142.

[Malouf, 2002] Malouf, R. (2002). A Comparison of Algorithms for Maximum Entropy Parameter Estimation. In *Proceedings of the 6th conference on Natural language learning - Volume 20*, COLING-02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

[McCarthy and Lehnert, 1995] McCarthy, J. F. and Lehnert, W. G. (1995). Using decision trees for coreference resolution. In *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1050–1055.

[Mikulová et al., 2007] Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Ševčíková, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., and Žabokrtský, Z. (2007). Annotation on the tectogrammatical level in the prague dependency treebank. Technical Report 3.1, ÚFAL, Charles University.

[Mikulová et al., 2005] Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z., and Kučová, L. (2005). Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka (t-layer annotation guidelines). Technical Report TR-2005-28, ÚFAL MFF UK, Prague, Prague.

[Mitkov, 2002] Mitkov, R. (2002). *Anaphora Resolution*. Longman, London.

[Mladová et al., 2008] Mladová, L., Zikánová, Š., and Hajičová, E. (2008). From sentence to discourse: Building an annotation scheme for discourse based on prague dependency treebank. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 1–7.

[MUC-6, 1995] MUC-6 (1995). Coreference task definition. In *Proceedings of the Sixth Message Understanding Conference*, San Francisco, CA. Morgan Kaufmann.

[MUC-7, 1998] MUC-7 (1998). Coreference Task Definition. In *Proceedings of the Seventh Message Understanding Conference*, San Francisco, CA. Morgan Kaufmann.

[Nedoluzhko, 2009] Nedoluzhko, A. (2009). *Zpracování rozšířené textové koreference a asociační anafory na tektogramatické rovině v Pražském závislostním korpusu*. PhD thesis, MFF UK, Praha, Czech Republic. In Czech.

[Nedoluzhko et al., 2009] Nedoluzhko, A., Mírovský, J., Ocelák, R., and Pergler, J. (2009). Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*.

[Němčík, 2006] Němčík, V. (2006). Anaphora resolution. Master's thesis, Faculty of Informatics, Masaryk University.

[Ng, 2005] Ng, V. (2005). Supervised ranking for pronoun resolution: Some recent improvements. In *AAAI*, pages 1081–1086.

[Ng, 2008] Ng, V. (2008). Unsupervised models for coreference resolution. In *EMNLP*, pages 640–649.

[Ng, 2009] Ng, V. (2009). Graph-cut-based anaphoricity determination for coreference resolution. In *HLT-NAACL*, pages 575–583.

[Ng, 2010] Ng, V. (2010). Supervised Noun Phrase Coreference Research: The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden. Association for Computational Linguistics.

[Ng and Cardie, 2002a] Ng, V. and Cardie, C. (2002a). Improving machine learning approaches to coreference resolution. In *Proceedings of the ACL*, pages 104–111.

[Ng and Cardie, 2002b] Ng, V. and Cardie, C. (2002b). Improving machine learning approaches to coreference resolution. In *ACL*, pages 104–111.

[Nguy, 2006] Nguy, G. L. (2006). Proposal of a set of rules for anaphora resolution in czech. Master's thesis, Faculty of Mathematics and Physics, Charles University.

[Nguy et al., 2009] Nguy, G. L., Novák, V., and Žabokrtský, Z. (2009). Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech. In *Proceedings of the SIGDIAL 2009 Conference*, pages 276–285, London, UK. ACL.

[Nguy and Žabokrtský, 2007] Nguy, G. L. and Žabokrtský, Z. (2007). Rule-based approach to pronominal anaphora resolution applied on the prague dependency treebank 2.0 data. In *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007)*, pages 77–81.

[NIST, 2007] NIST (2007). ACE Evaluation Plan. Technical report.

[Novák, 2010] Novák, M. (2010). Machine learning approach to anaphora resolution. Master's thesis, MFF UK, Prague, Czech Republic. In English.

[Němčík, 2006] Němčík, V. (2006). Anaphora resolution. Master's thesis, FI MU, Brno, Czech Republic. In English.

[Och et al., 1999] Och, F. J., Tillmann, C., Ney, H., and Informatik, L. F. (1999). Improved alignment models for statistical machine translation. In *University of Maryland, College Park, MD*, pages 20–28.

[Pajas, 2010] Pajas, P. (2010). *The Prague Markup Language (version 1.1)*.

[Pajas and Štěpánek, 2006] Pajas, P. and Štěpánek, J. (2006). XML-based representation of multilayered annotation in the PDT 2.0. In *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, pages 40–47.

[Pala and Všianský, 2000] Pala, K. and Všianský, J. (2000). *Slovník českých synonym*. Nakladatelství Lidové noviny, Prague, Czech Republic.

[Panevová, 1991] Panevová, J. (1991). Koreference gramatická nebo textová? In *Etudes de linguistique romane et slave*.

[Panevová, 1996] Panevová, J. (1996). *More Remarks on Control*, volume 2, pages 101–120. J.Benjamins Publ. House, Amsterdam - Philadelphia.

[Panevová, 1999] Panevová, J. (1999). Česká reciproční zájmena a slovesná valence. *Slovo a slovesnost*, 60:269–275.

[Panevová, 2007] Panevová, J. (2007). Znovu o reciprocitě. *Slovo a slovesnost*, 68(2):91–100.

[Panevová et al., 2002] Panevová, J., Kolářová-Řezníčková, V., and Urešová, Z. (2002). The theory of control applied to the prague dependency treebank (pdt). In Frank, R., editor, *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6)*, pages 175–180, Venezia, Italy. Universita di Venezia.

[Poesio, 2004] Poesio, M. (2004). The MATE/GNOME proposals for anaphoric annotation, revisited. In *In Michael Strube and Candy Sidner (editors), Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162.

[Poesio and Artstein, 2008] Poesio, M. and Artstein, R. (2008). *Anaphoric Annotation in the ARRAU Corpus*, pages 1170–1174. European Language Resources Association (ELRA).

[Poesio et al., 2002] Poesio, M., Ishikawa, T., im Walde, S. S., and Vieira, R. (2002). Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC)*, pages 1220–1224.

[Poesio et al., 2004a] Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004a). Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Poesio et al., 2004b] Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004b). Learning to resolve bridging references. In *ACL*, pages 143–150.

[Potau, 2008] Potau, M. R. (2008). Towards coreference resolution for catalan and spanish. Master's thesis, Universitat de Barcelona.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Rahman and Ng, 2009] Rahman, A. and Ng, V. (2009). Supervised models for coreference resolution. In *EMNLP*, pages 968–977.

[Recasens et al., 2007] Recasens, M., Martí, M. A., and Taulé, M. (2007). Text as scene: Discourse deixis and bridging relations. *Procesamiento del Lenguaje Natural*, (39):205–212.

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

[Soon et al., 2001] Soon, W. M., Ng, H. T., and Lim, C. Y. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544.

[Stoyanov et al., 2009] Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. (2009). Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664, Suntec, Singapore. Association for Computational Linguistics.

[Venables et al., 2002] Venables, W. N., Ripley, B. D., and Venables, W. N. (2002). *Modern applied statistics with S*. Springer, New York, 4th ed edition.

[Vieira et al., 2006] Vieira, R., Bick, E., Coelho, J., Muller, V., Collovini, S., Souza, J., and Rino, L. (2006). Semantic tagging for resolution of indirect anaphora. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, SigDIAL '06, pages 76–79, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Vilain et al., 1995a] Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995a). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, MUC6 '95, pages 45–52, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Vilain et al., 1995b] Vilain, M. B., Burger, J. D., Aberdeen, J. S., Connolly, D., and Hirschman, L. (1995b). A model-theoretic coreference scoring scheme. In *MUC*, pages 45–52.

[Vossen, 1998] Vossen, P., editor (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.

[Žabokrtský et al., 2008] Žabokrtský, Z., Ptáček, J., and Pajas, P. (2008). TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the 3rd Workshop on Statistical Machine Translation, ACL*.

[Weischedel and Brunstein, 2005] Weischedel, R. and Brunstein, A. (2005). BBN Pronoun Coreference and Entity Type Corpus. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2005T33, Philadelphia.

[Yang et al., 2006] Yang, X., Su, J., and Tan, C. L. (2006). Kernel-based pronoun resolution with structured syntactic knowledge. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL2006)*, pages 41–48, Sydney, Australia.

[Yang et al., 2008] Yang, X., Su, J., and Tan, C. L. (2008). A twin-candidate model for learning-based anaphora resolution. *Comput. Linguist.*, 34(3):327–356.

[Yang et al., 2003] Yang, X., Zhou, G., Su, J., and Tan, C. L. (2003). Coreference resolution using competition learning approach. In *ACL*, pages 176–183.

[Žabokrtský et al., 2008] Žabokrtský, Z., Ptáček, J., and Pajas, P. (2008). TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170.

[Zhou and Kong, 2009] Zhou, G. and Kong, F. (2009). Global learning of noun phrase anaphoricity in coreference resolution via label propagation. In *EMNLP*, pages 978–986.

Appendix A

Examples of Coreference Resolution

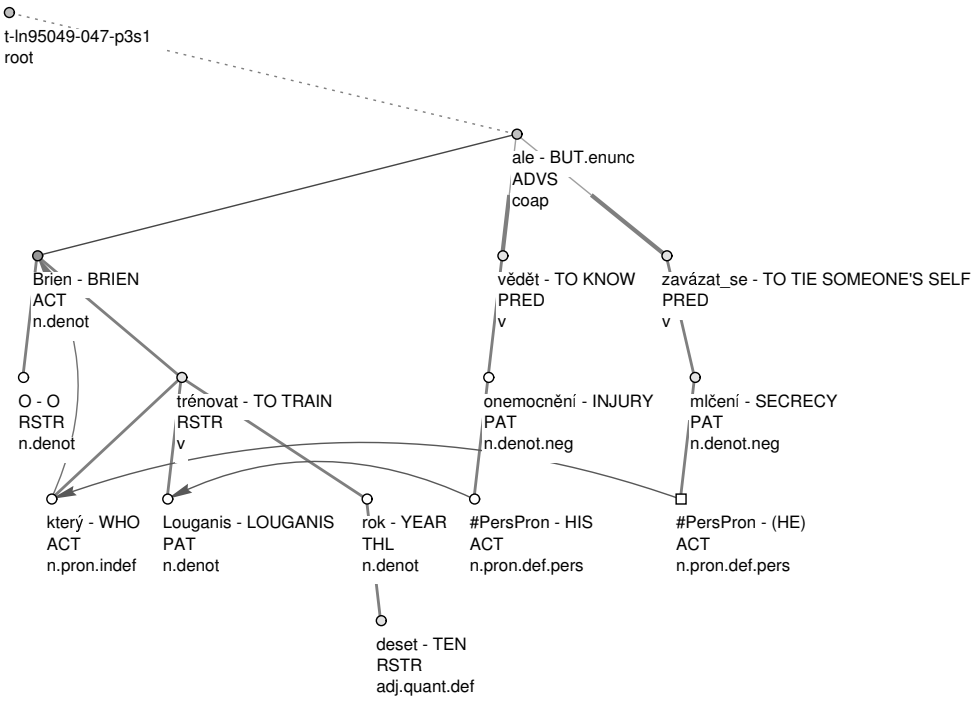

Figure A.1: Simplified tectogrammatical tree representing the sentence *O'Brien, který Louganise trénoval deset let, o jeho onemocnění věděl, ale zavázal se mlčením.* (Lit.: O'Brien, who Louganis trained for ten years, about his injury knew, but (he) tied himself to secrecy.) Note two coreferential chains {Brien, who, (he)} and {Louganis, his}.

Distance	
sent_dist	sentence distance between c and a_i
clause_dist	clause distance between c and a_i
node_dist	tree node distance between c and a_i
cand_ord	mention distance between c and a_i
Morphological Agreement	
gender	t-gender of c and a_i, agreement, joint
number	t-number of c and a_i, agreement, joint
apos	m-POS of c and a_i, agreement, joint
asubpos	detailed POS of c and a_i, agreement, joint
agen	m-gender of c and a_i, agreement, joint
anum	m-number of c and a_i, agreement, joint
acase	m-case of c and a_i, agreement, joint
apossgen	m-possessor's gender of c and a_i, agreement, joint
apossnum	m-possessor's number of c and a_i, agreement, joint
apers	m-person of c and a_i, agreement, joint
Functional Agreement	
afun	a-functor of c and a_i, agreement, joint
fun	t-functor of c and a_i, agreement, joint
act	c/a_i is an actant, agreement
subj	c/a_i is a subject, agreement
Context	
par_fun	t-functor of the parent of c and a_i, agreement, joint
par_pos	t-POS of the parent of c and a_i, agreement, joint
par_lemma	agreement between the parent's lemma of c and a_i, joint
clem_aparlem	joint between the lemma of c and the parent's lemma of a_i
c_coord	c is a member of a coordination
app_coord	c and a_i are in coordination & a_i is a possessive pronoun
sibl	c and a_i are siblings
coll	c and a_i have the same collocation
cnk_coll	c and a_i have the same CNC collocation
tfa	contextual boundness of c and a_i, agreement, joint
c_freq	c is a frequent word
Semantics	
cand_pers	c is a person name
cand_ewn	semantic position of c's lemma within the EuroWordNet Top Ontology

Table A.1: Features used in the pronominal anaphora resolution.

Feature	Value(s)	Weight
join_gen	anim_nr	44.87
join_gen	nr_anim	40.91
app_coord	1 or -1	39.58
gen_agree	1 or -1	37.65
cand_asubpos	D	30.98
join_num	nr_sg	29.55
num_agree	1 or -1	26.72
Gas	1 or -1	24.63
sent_dist	0	20.31
Natural	1 or -1	17.55
Animal	1 or -1	8.21
cand_pers	1 or -1	5.00
subj_agree	1 or -1	2.41
Human	1 or -1	2.30
Object	1 or -1	-7.95
join_gen	inan_anim	-27.83
join_num	pl_sg	-31.14
join_gen	nr_nr	-32.00
sibl	1 or -1	-56.30

Table A.2: Some feature weights estimated by the perceptron.

The ÚFAL/CKL Technical Report Series

ÚFAL

ÚFAL (Ústav formální a aplikované lingvistiky; http://ufal.mff.cuni.cz) is the Institute of Formal and Applied linguistics, at the Faculty of Mathematics and Physics of Charles University, Prague, Czech Republic. The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later the Faculty of Mathematics and Physics. Together with the "sister" Institute of Theoretical and Computational Linguistics (Faculty of Arts) we aim at the development of teaching programs and research in the domain of theoretical and computational linguistics at the respective Faculties, collaborating closely with other departments such as the Institute of the Czech National Corpus at the Faculty of Philosophy and the Department of Computer Science at the Faculty of Mathematics and Physics.

CKL

As of 1 June 2000 the Center for Computational Linguistics (Centrum komputační lingvistiky; http://ckl.mff.cuni.cz) was established as one of the centers of excellence within the governmental program for support of research in the Czech Republic. The center is attached to the Faculty of Mathematics and Physics of Charles University in Prague.

Technical Reports

The ÚFAL/CKL technical report series has been established with the aim of disseminate topical results of research currently pursued by members, cooperators, or visitors of the Institute. The technical reports published in this Series are results of the research carried out in the research projects supported by the Grant Agency of the Czech Republic, GAČR 405/96/K214 ("Komplexní program"), GAČR 405/96/0198 (Treebank project), grant of the Ministry of Education of the Czech Republic VS 96151, and project of the Ministry of Education of the Czech Republic LN00A063 (Center for Computational Linguistics). Since November 1996, the following reports have been published.

ÚFAL TR-1996-01 Eva Hajičová, *The Past and Present of Computational Linguistics at Charles University*
Jan Hajič and Barbora Hladká, *Probabilistic and Rule-Based Tagging of an Inflective Language – A Comparison*

ÚFAL TR-1997-02 Vladislav Kuboň, Tomáš Holan and Martin Plátek, *A Grammar-Checker for Czech*

ÚFAL TR-1997-03 Alla Bémová at al., Anotace na analytické rovině, *Návod pro anotátory (in Czech)*

ÚFAL TR-1997-04 Jan Hajič and Barbora Hladká, *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structural Tagset*

ÚFAL TR-1998-05 Geert-Jan M. Kruijff, *Basic Dependency-Based Logical Grammar*

ÚFAL TR-1999-06 Vladislav Kuboň, *A Robust Parser for Czech*

ÚFAL TR-1999-07 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (in Czech)*

ÚFAL TR-2000-08 Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek, *On Complexity of Word Order*

ÚFAL/CKL TR-2000-09 Eva Hajičová, Jarmila Panevová and Petr Sgall, *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-10 Zdeněk Žabokrtský, *Automatic Functor Assignment in the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-11 Markéta Straňáková, *Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování*

ÚFAL/CKL TR-2001-12 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (III. verze)*

ÚFAL/CKL TR-2002-13 Pavel Pecina and Martin Holub, *Sémanticky signifikantní kolokace*

ÚFAL/CKL TR-2002-14 Jiří Hana, Hana Hanová, *Manual for Morphological Annotation*

ÚFAL/CKL TR-2002-15 Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarská and Vendula Benešová, *Tektogramaticky anotovaný valenční slovník českých sloves*

ÚFAL/CKL TR-2002-16 Radu Gramatovici and Martin Plátek, *D-trivial Dependency Grammars with Global Word-Order Restrictions*

ÚFAL/CKL TR-2003-17 Pavel Květoň, *Language for Grammatical Rules*

ÚFAL/CKL TR-2003-18 Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, Václava Benešová, *Valency Lexicon of Czech Verbs VALLEX 1.0*

ÚFAL/CKL TR-2003-19 Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo, *Anotování koreference v Pražském závislostním korpusu*

ÚFAL/CKL TR-2003-20 Kateřina Veselá, Jiří Havelka, *Anotování aktuálního členění věty v Pražském závislostním korpusu*

ÚFAL/CKL TR-2004-21 Silvie Cinková, *Manuál pro tektogramatickou anotaci angličtiny*

ÚFAL/CKL TR-2004-22 Daniel Zeman, *Neprojektivity v Pražském závislostním korpusu (PDT)*

ÚFAL/CKL TR-2004-23 Jan Hajič a kol., *Anotace na analytické rovině, návod pro anotátory*

ÚFAL/CKL TR-2004-24 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2004-25 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *The Prague Dependency Treebank, Annotation on tectogrammatical level*

ÚFAL/CKL TR-2004-26 Martin Holub, Jiří Diviš, Jan Pávek, Pavel Pecina, Jiří Semecký, *Topics of Texts. Annotation, Automatic Searching and Indexing*

ÚFAL/CKL TR-2005-27 Jiří Hana, Daniel Zeman, *Manual for Morphological Annotation (Revision for PDT 2.0)*

ÚFAL/CKL TR-2005-28 Marie Mikulová a kol., *Pražský závislostní korpus (The Prague Dependency Treebank) Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2005-29 Petr Pajas, Jan Štěpánek, *A Generic XML-Based Format for Structured Linguistic Annotation and Its application to the Prague Dependency Treebank 2.0*

ÚFAL/CKL TR-2006-30 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Annotation manual)*

ÚFAL/CKL TR-2006-31 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Anotace na tektogramatické rovině Pražského závislostního korpusu (Referenční příručka)*

ÚFAL/CKL TR-2006-32 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall,Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Reference book)*

ÚFAL/CKL TR-2006-33 Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*

ÚFAL/CKL TR-2006-34 Markéta Lopatková, Zdeněk Žabokrtský, Václava Benešová (in cooperation with Karolína Skwarska, Klára Hrstková, Michaela Nová, Eduard Bejček, Miroslav Tichý) *Valency Lexicon of Czech Verbs. VALLEX 2.0*

ÚFAL/CKL TR-2006-35 Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský, *Annotation of English on the tectogrammatical level*

ÚFAL/CKL TR-2007-36 Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza, *Zpracování pojmenovaných entit v českých textech*

ÚFAL/CKL TR-2008-37 Silvie Cinková, Marie Mikulová, *Spontaneous speech reconstruction for the syntactic and semantic analysis of the NAP corpus*

ÚFAL/CKL TR-2008-38 Marie Mikulová, *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory*

ÚFAL/CKL TR-2008-39 Zdeněk Žabokrtský, Ondřej Bojar, *TectoMT, Developer's Guide*

ÚFAL/CKL TR-2008-40 Lucie Mladová, *Diskurzní vztahy v češtině a jejich zachycení v Pražském závislostním korpusu 2.0*

ÚFAL/CKL TR-2009-41 Marie Mikulová, *Pokyny k překladu určené překladatelům, revizorům a korektorům textů z Wall Street Journal pro projekt PCEDT*

ÚFAL/CKL TR-2011-42 Loganathan Ramasamy, Zdeněk Žabokrtský, *Tamil Dependency Treebank (TamilTB) - 0.1 Annotation Manual*

ÚFAL/CKL TR-2011-43 Nguy Giang Linh, Michal Novák, Anna Nedoluzhko, Coreference Resolution in the Prague Dependency Treebank