How Dependency Trees and Tectogrammatics Help Annotating Coreference and Bridging Relations in Prague Dependency Treebank

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

Jiří Mírovský
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic
mirovsky@ufal.mff.cuni.cz

Abstract

In this paper, we explore the benefits of dependency trees and tectogrammatical structure used in the Prague Dependency Treebank for annotating language phenomena that cross the sentence boundary, namely coreference and bridging relations. We present the benefits of dependency trees such as the detailed processing of ellipses, syntactic decisions for coordination and apposition structures that make possible coding coreference relations in cases that are not so easy when annotating on the raw texts. We introduce the coreference decision for non-referring constructions and present some tectogrammatical features that are useful for annotation of coreference.

1 Introduction

The dependency syntax is one of the most influential linguistic theories. However, its benefits are mainly explored for research of linguistic phenomena that do not cross the sentence boundary and may be illustrated within a single dependency tree. In this paper, we will explore how dependency trees and tectogrammatical structure can help in the annotation of coreference and bridging relations in the Prague Dependency Treebank.

The Prague Dependency Treebank (henceforth PDT, Hajič et al. 2006) is a large collection of linguistically annotated data and documentation. In PDT 2.0, Czech newspaper texts are annotated on three layers: morphological, syntactic and complex semantic (tectogrammatical). In addition to syntax, the tectogrammatical layer includes the annotation of topic-focus articulation, discourse relations[1], coreference links and bridging relations. Benefits of tectogrammatics in the annotation of discourse structure were examined in Mírovský et al. (2012), we will focus on coreference and bridging relations.

When we say that we take certain advantages from the tectogrammatical layer, we should realize that the advantages of two kinds are possible: we can take advantage from the dependency structure itself, independently of the PDT conception, and we can use the information included in the tectogrammatical layer as a specific contribution of the Prague Dependency Treebank. In this paper, when we describe benefits that can be obtained for coreference and bridging annotation using the dependency structure, our examples are syntactically analyzed using the PDT tectogrammatical annotation strategy and are seriously influenced by this approach. However, we suppose that in principle any dependency analyzer would be able to solve these problems in a similar way.

2 The Coreference and Bridging Relations in PDT

There are three types of relations annotated in PDT: (a) grammatical coreference (coreference of relative and reflexive pronouns, arguments of verbs of control, arguments in constructions with reciprocity and verbal complements), (b) pronominal and nominal textual coreference (including zero anaphora), which is further specified into coreference of specific (type SPEC) and generic (type GEN) noun phrases, and (c) bridging relations, which mark some associative semantic relations between non-coreferential entities. The following types of bridging relations are distinguished: PART-OF (e.g. room - ceiling), SUBSET (students - some students) and FUNCT (state - president)

[1] The annotation of discourse and bridging relations is a later addition to the data of PDT, see http://ufal.mff.cuni.cz/discourse/

Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pages 244–251,
Prague, August 27–30, 2013. © 2013 Charles University in Prague, Matfyzpress, Prague, Czech Republic

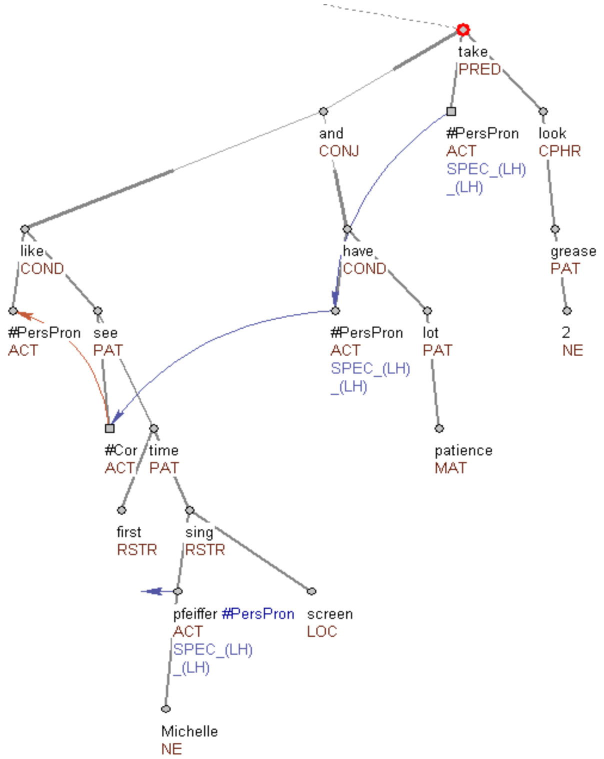

Figure 1: *If you'd like* {#Cor.ACT} *to see the first time Michelle Pfeiffer sang on screen, and you have a lot of patience,* #PersPron *take a look at Grease 2.*

traditional relations, CONTRAST for coherence relevant discourse opposites (this year - last year), ANAF for explicitly anaphoric relations without coreference (rainbow - that word) and the further underspecified group REST[2].

Coreference relations are marked between the whole subtrees of the antecedent/anaphoric expressions that are the subject to annotation.

3 The Annotation Tool

The primary format of PDT 2.0 is called PML. It is an abstract XML-based format designed for annotation of treebanks. For editing and processing data in PML format, a fully customizable tree editor TrEd has been implemented (Pajas and Štěpánek 2008). For the coreference and bridging annotation, a special extension was used, included into the system as a module.

Technically, the coreference extension module of TrEd allows annotation both on raw texts and on dependency trees. However, annotation on dependency trees is more comfortable as it gives more visual information about the function of the annotated noun phrases in the sentence structure, about being in the governing or dependent position in the coreferring expression, about being a part of an appositional or a coordinative construction and so on.

4 Benefits of Dependency Trees and Tectogrammatics

One of the technical advantages of gold dependency trees is the automatic extraction of elements to be annotated for coreference, including so called minimal markables (MIN-IDs) that are always the governing expression of full span markable expressions. Of course, it does not solve the problem for coreference resolution systems, but it makes the manual annotation easier, reducing it to a single step of coding coreferential links between already identified markables.

4.1 Syntactic zeros

In so-called 'pro-drop' languages such as most Romance languages, Japanese, Greek, most Slavonic languages, etc., a phonetic realization is not required for anaphoric references in contexts where they are syntactically or pragmatically inferable. The problem of syntactic zeros is the

[2] For a detailed classification of coreference and bridging relations used in PDT, see e.g. Nedoluzhko et al. (2011).

subject of research in many different linguistic theories (see e.g. the elaboration of different aspects in generative grammar (Roberts 1997), Prague dependency grammar (Panevová 1986, Růžička 1999), Moscow structuralismus (Meľčuk 1974), etc.). There is not much theoretical disagreement concerning such elements (at least in case of zero anaphoric pronouns and control constructions), but they raise a lot of problems with annotation of their relations to coreferential expressions and automatic resolution of these relations. Thus, only a few coreference annotation projects reconstruct the ellided expressions and annotate them for coreference (see e.g. Xue et al. 2005, Pradhan et al. 2007). However, for example, with tools such as MMAX (Müller and Strube 2001), the only option is to have 'verbal markables' as done e.g., in VENEX (Poesio et al. 2004) and LiveMemories (Rodriguez 2010) annotation (i.e., annotate a relation between the immediately following verbal element and the antecedent). In AnCora (Recasens and Martí 2010), zero subjects were added as extra 'empty' tokens, and these were used for annotating coreference.

Consequent annotation of such arguments is better possible with annotation tools that use as a base layer a full syntactic annotation or an argument structure. As TrEd is one of such tools, its benefits are used for reconstructing ellided expressions, their coreference relations being further consistently annotated. Zero arguments are reconstructed using the PDT Valency Lexicon VALLEX (Hajič et al. 2003), which for each autosemantic, valency-capable word provides its valency information.

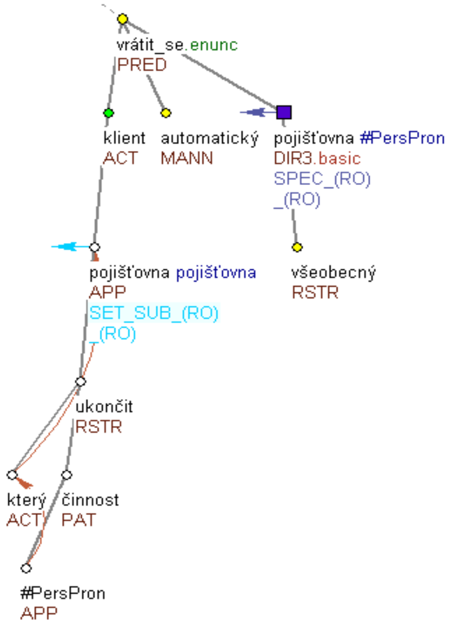

Figure 2: *Klienti pojišťoven, které ukončí svou činnost, se automaticky vrátí k Všeobecné*. (=lit. *Clients of insurance companies which shut down will automatically return to the General {one}*.)

According to the detailed classification of ellipses introduced in Mikulová (2011), PDT uses a rich variety of newly established nodes occupying positions of all kinds of modifications. The classification of these nodes corresponds to the ability of different types of newly established nodes to take part in coreference relations. Newly established nodes that are subjects to coreference annotation are the following:

- **#PersPron**. This lemma is assigned to nodes representing personal or possessive pronouns. It applies both to newly established nodes and to those present at the surface level. In most cases, nodes with #PersPron lemma, especially those representing personal pronouns in the third person, are connected with their antecedents by coreference relations (the rare exceptions are mostly generic uses of pronouns used once in the text without further reference). Cf. Fig. 1.
- **#Cor**. This lemma is assigned to newly established nodes representing the (usually inexpressible) controllee in control constructions. These nodes are always connected by a grammatical coreference link with its controller, cf. coreference of unexpressed actor of the verb in Fig. 1.
- **#QCor**. This lemma is assigned to newly established nodes representing a (usually inexpressible) valency modification in constructions with so-called quasi-control. This case can be found with multi-word predicates the dependent part of which is a noun with valency requirements, cf. *He offered Jan* {#QCor} *protection*. The valency of the verb *offer* as well as the modification of the noun *protection* has the same referent *Jan*. This shared modification can only be present once at the surface level (it is impossible to say: **He offered Jan protection of Jan*). These nodes are always connected by a grammatical coreference link with its controller.
- **#Rcp**. This lemma is assigned to newly established nodes representing participants that are left out as a result of reciprocation. There is always a grammatical coreference

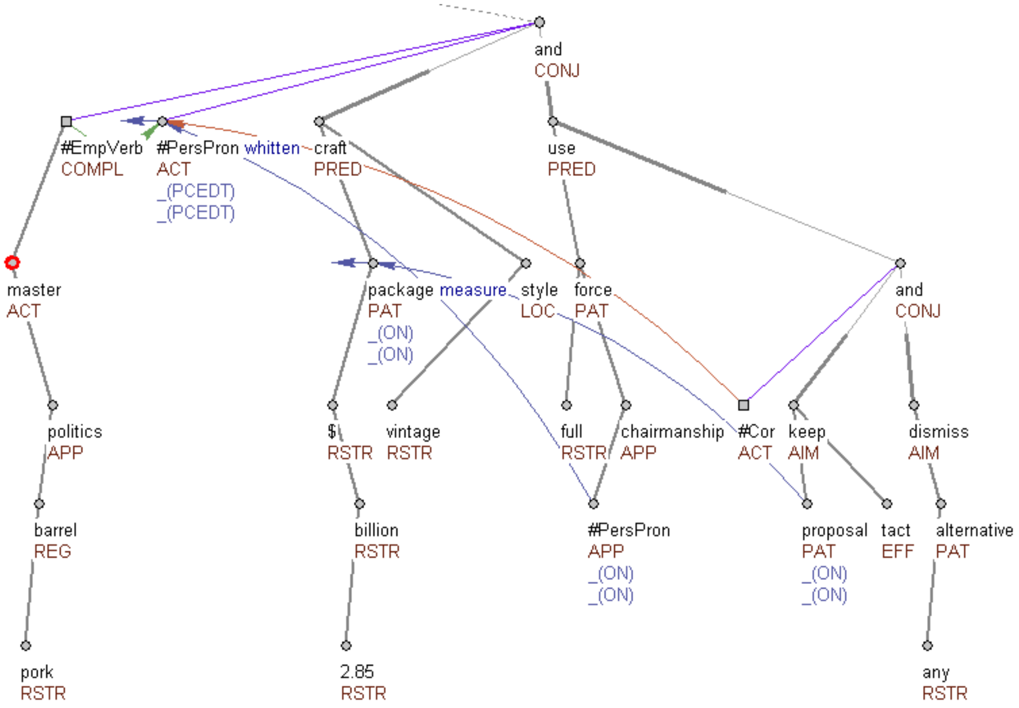

Figure 3: *A master of pork-barrel politics, he had crafted the $2.85 billion package in vintage style and used the full force of his chairmanship to keep the proposal intact and dismiss any alternative*.

relationship indicated in the tectogrammatical tree, going from the node with the #Rcp t-lemma to the node with which it is in the reciprocal relation: *The lovers kissed* {#Rcp.PAT}.

- If it is clear (and possible to find in the text) which noun has been omitted in the surface structure of the sentence (the case of textual ellipsis), a copy of the node representing the same lexical unit as the omitted element is inserted into the appropriate position. Cf. Fig. 2.

Other newly established nodes are not supposed to be linked by coreference. These are e.g. **#Gen** for a general participant (*Houses are built* {#Gen.ACT} *from bricks.*), **#Unsp** for valency modifications with vague (non-specific) semantic content (*U Nováků* {#Unsp.ACT} *dobře vaří. (=They cook well at Nováks'.)*), **#EmpNoun** for non-expressed nouns governing syntactic adjectives, which are not the case of textual ellipsis (*Přišli jen* {#EmpNoun.ACT} *mladší.* (=*lit. Came only {#EmpNoun} younger.*)), **#Oblfm** for obligatory adjuncts that are absent at the surface level (*Ta vypadá.* {#Oblfm.MANN} *(=lit. That.fem looks; meaning: She looks awful/so strange...*)) and some other newly established nodes used in comparative constructions.

Such a detailed linguistically elaborated method provides very consistent information of the analyzed language and thus a reliable base for a theoretical linguistic research, but corpora annotated in this way are problematic for many automatic resolving systems, as the state of the art at extracting full syntactic structure / argument structure from text is still not good enough. However, the results for #PersPron resolution in PDT are not so bad. A rule-based system employed in Nguy and Žabokrtský (2007) to resolution of pronominal textual coreference got the success rate of 74 % (F1-measure). Applying machine learning methods, particularly perceptron ranking in Nguy et al. (2009) on the same task outperformed the rule-based method with F1-measure over 79 %.

4.2 Processing non-referring expressions

Non-referring expressions such as appositions, verbal complements and noun phrases in predicative positions are a special problematic issue in coreference annotation projects that mark coreference on raw texts. Coding coreference on dependency trees may solve the problem. In PDT, appositions, verbal complements and noun phrases in predicative

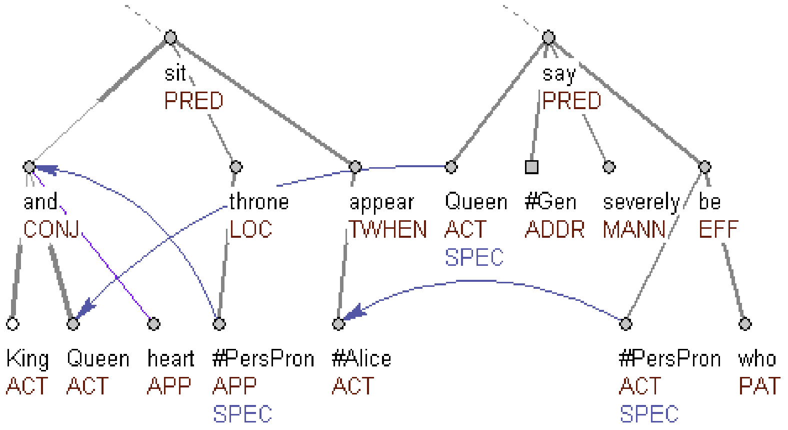

Figure 4: *The King and Queen of Hearts were sitting on their throne when Alice appeared. The Queen said severely "Who is she?"*

positions are resolved on the syntactic level and they do not need to be additionally annotated for coreference. This information can be easily extracted from the tectogrammatical layer. Thus, for appositions, the whole construction serves as a markable for coreference annotation, its parts beeing connected with a node with a special tectogrammatical functor APPS. The predicative relation is obvious. For verbal complements, the functor COMPL is used, the dependency on a noun being additionally represented by means of a special attribute compl.rf, in TrEd visualized by a (non-coreference) arrow (see Fig. 3).

4.3 Coordinative constructions

Coordination structures and their connection with plural reference are another difficult issue for processing coreference relations. E.g. the semantics of plural reference to a coordination like *John and Mary met. They had not seen each other for a long time* is fairly uncontroversial from a semantic point of view and can be solved satisfactorily by any annotation system (the coordination construction as a whole and its parts separately may be lined by coreference relations). On the contrary, the problem of multiple antecedents for *they* in *John visited Ellen, and they went to the seaside* will present a problem for all, no matter if dependency-based or raw-text annotations. Still, there are coordinative constructions that are complicated for annotation on texts (a special split-antecedent mechanism is needed) and have an elegant solution on dependency trees.

Annotating coreference link for *the Queen* in Fig. 4[3] on the raw text is problematic because of its modifier *of Hearts* is common for both NPs, *The King* and *the Queen*. In PDT tectogrammatics, this is resolved by a dependency structure, as shown in Fig. 4.

The reconstruction of a complex syntactic construction makes it possible to refer to a coordination *Latitude or Longitude* in Fig. 5.

4.4 Contribution of tectogrammatics

There are some additional helpful features that do not necessarily depend on the dependency structure of the text representation but are present in the tectogrammatical level of the Prague Dependency Treebank and are very useful for the consistent annotation of coreference and bridging relations and its further analysis. According to its semantic part of speech, each node contains grammatical information about gender, number, resp. person, tense, mood, etc. Direct speech and parenthesis are marked in a special attribute. Discourse annotation supplies the information about which expressions are parts of titles or subtitles. Very important for the analysis of text cohesion is the topic-focus articulation that is annotated manually for the whole PDT.

Furthermore, PDT uses a special approach to the syntactic annotation of quantifiers, measure NPs and constructions with similar semantic meaning. In PDT annotation guidelines, they are called nouns with a 'container' meaning. Their

[3] For examples Fig. 4 and Fig. 5, the sentences from the discussion on the workshop RAIS are used (http://wiki.ims.uni-stuttgart.de/RAIS/Stuttgart-Workshop).

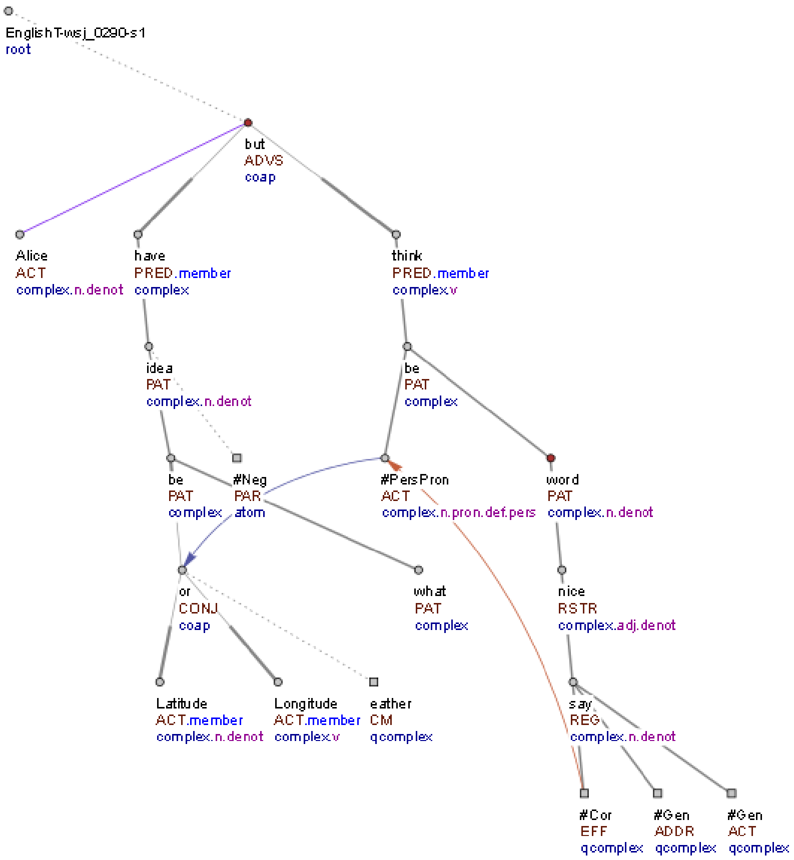

Figure 5. *Alice had no idea what Latitude was, or Longitude either, but thought they were nice grand words to say*.

arguments have typical semantic label MAT and can be easily recognized in the dependency tree. This fact was widely exploited for coreference annotation. The 'container' words were basically considered to be markables for coreference relations, their dependent elements were annotated for coreference only in some special (rare) cases where annotation to 'containers' was not possible for technical or semantical reasons (Nedoluzhko 2011).

Also information about the tectogrammatical functors of potentially coreferring nodes is widely explored for annotation of coreference and bridging relations. In PDT, functors mainly represent the semantic values of syntactic dependency relations, they express the functions of individual modifications in the sentence. As mentioned above, functors for appositive, predicative and coordinative constructions, as well as the special functor for verbal complements, are of great use for consequent coreference annotation. Moreover, when annotating coreference and bridging relations, there should be considered such functors as ID, ACMP, AUTH etc.

4.4.1 The ID functor

The functor ID (identity) is used as a functor for an identifying expression, which is represented as an identification structure. The ID functor is assigned to adnominal adjuncts representing meta-language expressions, proper nouns and names of animals, objects and events, e.g. *v případu Kott - Kutílek* (= *in the case of Kott - Kutílek*); *agentura Reuters* (=*Reuters agency*); *pojem čas* (=*notion of time*). In such cases, noun phrases do not refer to objects but to themselves. For this reason, all constructions containing expressions with the ID functor are annotated for coreference as one unit. The coreference arrow links the governing node to the node with the ID functor (i.e. *agency* in case of *Reuters agency*, *notion* in case of *notion of time* and so on).

4.4.2 The ACMP functor

The ACMP functor (accompaniment) is a functor for such an adjunct which expresses manner by specifying a circumstance (an object, person, event) that accompanies (or fails to accompany) the event or entity modified by the adjunct. The meaning of the ACMP functor may appear in conflict with some bridging relations (mainly SUBSET). In this case, the bridging relations are not annotated. Cf. *válečná plavidla včetně bojových letadel*.ACMP *a bitevních vrtulníků*.ACMP (=*warships including air force* ...).

5 Problematic Issues

Of course, dependency trees and tectogrammatics do not solve all coreference annotation problems. One of the problematic issues that remains daunting for coreference annotation is the identity of prepositional phrases and included noun phrases. In PDT, prepositions are hidden in sub-functors and can be taken into account if annotating on tectogrammatical trees only by looking at these subfunctors. Although the semantic distinction between prepositional phrases with the same head and different preposition is very important, we ignore it in the annotation. So, if two noun phrases are coreferential, we mark coreferential relation between them also in case when they are parts of prepositional phrases which are not coreferential. Although contraintuitive, the following expressions will be marked as coreferential: *Prague – near Prague, before the war – during the war – after the war*. The distinction between PPs and NPs beeing in Prague tectogrammatics complicated (though technically possible), the question about the ability of PPs to corefer still remains open, so our decision to mark coreference for NPs ignoring PPs still remains quite consequent.

6 Conclusion

In this paper, we demonstrated how manual annotation of coreference relations may benefit from the use of dependency trees and the tectogrammatical structure of the Prague Dependency Treebank. We considered separately the contribution of dependency trees and the tectogrammatics. Although dependency syntactic annotation is quite costly and time-consuming, it gives good structural solutions for processing coreference in predicative, appositive and coordinative structures, constructions with ellipses of different kinds and so on. The connection to the syntactico-semantic analysis of the tectogrammatical layer in the Prague Dependency Treebank appears as a rather convenient tool. In addition to issues already mentioned, it makes it possible to work with already established and coherent solutions of typical syntactic constructions and tectogrammatic functors. Along with other similar tasks being performed on the same PDT level (topic-focus articulation, discourse annotation), it creates a reliable basis for a deeper linguistic research in the field of language phenomena that cross the sentence boundary.

Acknowledgements

We gratefully acknowledge support from the Grant Agency of the Czech Republic (grants P406/12/0658 and P406/2010/0875).

Bibliography

J. Hajič et al. 2006. Prague Dependency Treebank 2.0.CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.

J. Hajič, J. Panevová, Z. Urešová, A. Bémová and V. Kolářová. 2003. PDT-Vallex: Creating a Large-coverage Valency Lexicon for Treebank Annotation**.** In Proceedings of The Second Workshop on Treebanks and Linguistic Theories, 57-68. Vaxjo University Press.

I. Meľčuk. 1974. O sintaksicheskom nule. In Cholodovich A.A. (ed.) *Tipologija passivnych konstrukcij. Diatezy i zalogi*. Leningrad.

M. Mikulová. 2011. *Významová reprezentace elipsy.* ÚFAL, Prague, 230 pp.

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

J. Mírovský, P. Jínová, L. Poláková. 2012. Does Tectogrammatics help the Annotation of Discourse? In *Procedings of the 24th International Conference on Computational Linguistics* (COLING 2012), Mumbai, India.

L. Mladová, Š. Zikánová, E. Hajičová. 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)*, Marakéš, Maroko.

Ch. Müller, M. Strube. 2001. Annotating anaphoric and bridging relations with MMAX. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*. Aalborg, Denmark, pp. 90–95.

A. Nedoluzhko. 2011. *Rozšířená textová koreference a asociační anafora. Koncepce anotace českých dat v Pražském závislostním korpusu.* ÚFAL, Prague.

G. L. Nguy, Z. Žabokrtský. 2007. Rule-based Approach to Pronominal Anaphora Resolution Applied on the Prague Dependency Treebank 2.0 Data. In *Proceedings of the 6th Discourse Anaphroa and Anaphor Resolution Colloquium*, Lagos, 2007.

G. L. Nguy, V. Novák, Z. Žabokrtský. 2009. Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech. In *Proceedings of the SIGDIAL 2009 Conference*, London.

P. Pajas, J. Štěpánek. 2008. Recent advances in a feature-rich framework for treebank annotation. In *The 22nd Interntional Conference on Computational Linguistics – Proceedings of the Conference*. Manchester, pp. 673-680.

J. Panevová. 1986. The Czech infinitive in the function of objective and the rules of reference. In Mey J. (ed) *Language and discourse: Text and protest*. Amsterdam; Philadelphia: John Benjamins.

M. Poesio, R. Delmonte, A. Bristot, L. Chiran, S.Tonelli. 2004. *The VENEX corpus of anaphoric information in spoken and written Italian*. Manuscript.

M. Poesio. 2004. The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. In *Proc.of SIGDIAL*, Boston.

M. Poesio, R. Artstein. 2008. Anaphoric annotation in the ARRAU corpus In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC).* Marrakech, Morocco.

S. Pradhan, L. A. Ramshaw, R. M. Weischedel, J. MacBride, L. Micciulla. 2007. Unrestricted Coreference: Identifying Entities and Events in OntoNotes. In *Proceedings of the International Conference on Semantic Computing (ICSC '07)*. 446-453.

M. Recasens, A. Martí. 2010. *AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan*. In Language Resources and Evaluation.

I. Roberts. 1997. *Comparative syntax.* London: Arnold.

K. J. Rodriguez, F. Delogu, Y. Versley, E. Stemle and M. Poesio. 2010. Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May 2010.

R. Růžička. 1999. *Control in grammar and pragmatics: a cross-linguistic study*. John Benjamins Publishing.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and M. Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.