Chapter 3

Possibilities of text coherence analysis in the Prague Dependency Treebank

Kateřina Rysová
Charles University, Faculty of Mathematics and Physics

The aim of this paper is to examine the interplay of text coreference and sentence information structure and its role in text coherence. The study is based on the analysis of authentic Czech texts from the Prague Dependency Treebank 3.0 (PDT; i.e. on almost 50 thousand sentences). The corpus contains manual annotation of both text coreference and information structure – the paper tries to demonstrate how these two different corpus annotations may be used in examination of text coherence. In other words, the paper tries to describe where these two language phenomena meet and how important the interplay is in making text well comprehensible for the reader. Our results may be used not only in a theoretical way but also practically in automatic corpus annotations, as they may give us an answer to the general question whether it is possible to annotate the sentence information structure automatically in large corpora on the basis of text coreference.

1 Introduction and theoretical background

Studying text coherence is dependent on studying several individual language phenomena like coreference, anaphora, sentence information structure or discourse (mainly in terms of semantico-pragmatic discourse relations). In other words, a text may be imagined as a net of many different kinds of relations that are mutually interconnected and possibly influence each other.

So far, these phenomena have been studied primarily in isolation but recently, there is a growing need for more complex studies focusing on interaction (see, for example, Hajičová, Hladká & Kučová 2006; Hajičová 2011; Eckert & Strube 2000;

Kateřina Rysová. 2017. Possibilities of text coherence analysis in the Prague Dependency Treebank. In Katrin Menzel, Ekaterina Lapshinova-Koltunski & Kerstin Kunz (eds.), *New perspectives on cohesion and coherence*, 33–45. Berlin: Language Science Press. DOI:10.5281/zenodo.814462

Rysová & Rysová 2015). In other words, if we want to analyze text coherence deeply (i.e. to help to answer the question what are the general properties of a text), we have to pay closer attention to the interactions of several individual phenomena at once (operating both inter- and intra-sententially).[1]

The theme of interplay between coreference or anaphoric relations and sentence information structure has been studied recently especially in Nedoluzhko & Hajičová (2015) and Nedoluzhko (2015) who linguistically investigated contextually bound nominal expressions (explicitly present in the sentence) that do not have an anaphoric (bridging, coreference or segment) link to a previous (con)text. They draw the conclusion that three cases may be found when contextually bound expressions may not be linked by any coreference or anaphoric relation: (i) contextually bound nominal groups related to previous context (semantically or pragmatically) but not specified as bridging relations in the Prague Dependency Treebank (PDT); (ii) noun groups referring to secondary circumstances (like temporal, local, etc.) and (iii) nominal groups having low referential potential.

In this respect, this paper follows their work. It investigates a narrower data sample (only expressions interlinked by text coreference) with the aim to bring an overview of density of text coreference relations according to the sentence information structure values of the interlinked expressions.

The complex analysis of text coherence demands extensive language material of authentic texts, i.e. large language corpora with multilayer annotation. Such corpora are rather rare (cf., for example, Komen 2012; Stede & Neumann 2014; Chiarcos 2014). The corpus with one of the richest (i.e. multilayer) annotation is the Prague Dependency Treebank (PDT) for Czech (see Bejček et al. 2013). The PDT contains detailed annotation on morphological, analytical (surface syntactic) and tectogrammatical (deep syntactic) level as well as the annotation of sentence information structure, coreference and anaphoric relations, discourse relations and text genres. The PDT thus offers suitable language material for studies focusing on the annotated language phenomena in interaction.

The paper concentrates on the interplay of two of them – text coreference and sentence information structure (mainly in terms of contextual boundness) – as well as on the fact how and to what extent this interplay is projected into text coherence.

[1] For complex studying of coherence phenomena, see Accessibility Theory (Ariel 1988) or Centering Theory (Joshi & Weinstein 1981; Grosz & Sidner 1986).

2 Main objectives

Generally, as said above, the paper focuses on the relation between text coreference and sentence information structure. It describes where and in which aspects these two phenomena meet in the text and how they influence each other. It also presents methods that may be used for analyzing language interplays in general (demonstrated using the PDT data). Finally, the paper demonstrates whether and how the present (manual) annotation of text coreference in the PDT may be used for improving automatic annotation of sentence information structure.

To meet the goals, the paper focuses on the specific tasks concerning the relation of text coreference and sentence information structure (in sense of contextual boundness – see §3.1). The paper explores whether the text coreference relations (in the PDT texts) connect rather contextually bound or non-bound sentence members (mutually) or both of them in the same way, see Examples 1 and 2 and Figure 1 below.

Since the contextually bound sentence items usually carry information that is deducible from the previous (con)text (in contrast to the contextually non-bound items), we assume the higher number of text coreference links leading right from them. In other words, the assumption is that text coreference and sentence information structure meet especially in sentence items related somehow to the previous (con)text.

3 Methods and material

3.1 Sentence information structure in the PDT

The analysis uses the language data of the Prague Dependency Treebank. The PDT contains almost 50,000 sentences (833,195 word tokens in 3,165 documents) of Czech newspaper texts that are (mostly manually) annotated on several language levels at once.

The theoretical framework for sentence information structure in the PDT is based on Functional Generative Description (FGD) introduced by Sgall (1967) and further developed especially by Hajičová, Partee & Sgall (1998).

The annotation is carried out on tectogrammatical trees. Each relevant node of the tree is labeled with one of the three values of contextual boundness.[2] Contextual boundness has the following possible values: non-contrastive contextually

[2] In addition, the communicative dynamism is annotated – as deep order of the nodes in the tree.

bound nodes (marked as "t"), contrastive contextually bound nodes (marked as "c") and contextually non-bound nodes (marked as "f").

Non-contrastive contextually bound nodes represent units that are considered deducible from the broad (not necessarily verbatim) context and are known for the reader (or presented as known for him or her). Contrastive contextually bound nodes also are expressions related to the broad context and moreover, they usually represent a choice from a set of alternatives. They often occur at the beginning of paragraphs, in enumerations etc. In spoken language, such units carry an optional contrastive stress. Contextually non-bound expressions are not presented as known and are not deducible from the previous context – on the contrary, they represent new facts (or known facts in new relations). The particular occurrences of contextual boundness values can be found in (1).

(1) [*Jane is my friend.*] She.t is.f very.f fine.f. However, her.t brother.c is.t boring.f. I.t like.f rather.f her.f.

On the basis of contextual boundness, the division of the sentence into Topic and Focus is realized (Topic is formed especially by contextually bound items and Focus typically by non-bound items). In the first sentence, the Topic is *she* and the Focus *is very fine*. In the second sentence, the Topic part includes *however, her brother is* and the Focus part *boring*. The participant *I* is the Topic of the third sentence and the part *like rather her* is the Focus.[3]

For further examples of "t", "c" and "f" nodes, see (2) in §3.3. For more details about Topic-Focus Articulation, see Hajičová, Partee & Sgall (1998).

3.2 Text coreference in the PDT

Annotation principles of text coreference in the PDT were done according to Nedoluzhko (2011). In this concept, the text coreference is understood as the use of different language means for marking the same object of textual reference. The basic principle of text coreference is that the antecedent and the anaphor referents are identical (e.g. *a house – the house*; *Jane – she – her*; *Jane – 0*; *problem – this – that*).

The general aspect of text coreference is that the coreferential relation is symmetric (if A is coreferential with B, B is coreferential with A) and transitive (if A is coreferential with B and B is coreferential with C, then A is coreferential with C).

[3] For more details about annotation of sentence information structure in English texts, see Rysová, Rysová & Hajičová (2015).

Text coreference relations in the PDT are represented especially by personal or possessive pronouns (*Jane – she – her*), ellipsis (*Jane – 0*), demonstratives (*problem – this – that*) or by referential nominal phrases (concerning mainly nouns with specific, abstract or generic reference – for more details see Nedoluzhko (2011)) and they operate both inter- and intra-sententially.

3.3 Example of a dependency tree from the Prague Dependency Treebank

(2) illustrates the most common corpus occurrence – the text coreference connection leading from a non-contrastive contextually bound node to another non-contrastive contextually bound node (i.e. from "t" to "t").

(2) [Jestliže ve státě New Hampshire začne geometricky narůstat kriminalita mladistvých, veřejnost ocení svou přízní vládní akt zvýšení výdajů na boj se zločinností.]] Takové dobré **opatření** nakonec udělá každá druhá vláda, zvlášť půjde-li o **opatření** předvolební.

[If the juvenile delinquency will increase in the state of New Hampshire, the public will appreciate the government act to increase spending on the fight against crime.] Every other government eventually makes such good **measure**, regarding especially a pre-election **measure**.

Figure 1 represents the sentence from (2). The text coreference arrow leads from the second occurrence of the word *measure* (non-contrastive contextually bound ("t")) to the first occurrence of the word *measure* (that is also non-contrastive contextually bound ("t"), i.e. deducible from the previous context).

Another coreference relation is between the nodes *government* (Figure 1) and *government* (*act*) from the previous sentence, see (2). In Figure 1, only the starting position of this coreference relation can be seen. The final position of the coreference arrow is in the previous tree in the treebank and it is not displayed in Figure 1.

3.4 PML Tree Query

Our analysis of the interaction between information structure and text coreference was carried out with the client-server PML Tree Query (PML-TQ; the primary format of the PDT is called Prague Markup Language) (Štěpánek & Pajas 2010). The client part has been implemented as an extension to the tree editor TrEd (Pajas & Štěpánek 2008) that may be used also for editing data.

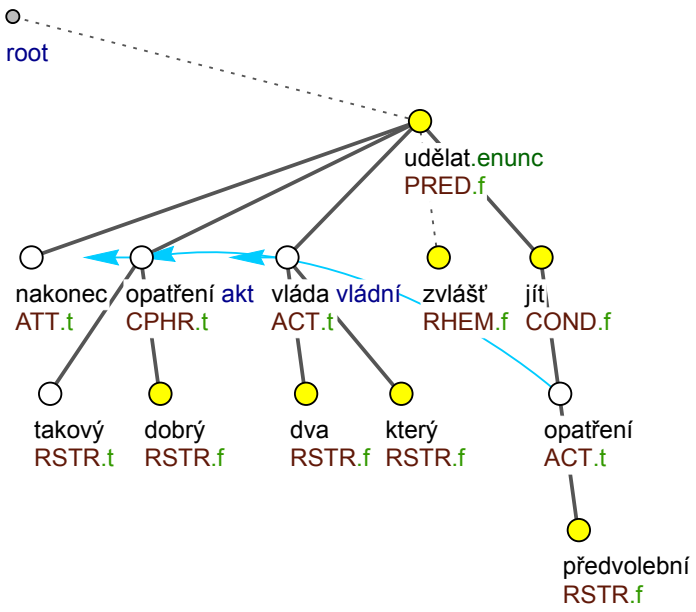

Figure 1: Dependency tree from the Prague Dependency Treebank depicting the sentence *Takové dobré opatření nakonec udělá každá druhá vláda, zvlášť půjde-li o opatření předvolební. – Every other government eventually makes such good measure, regarding especially a pre-election measure.*

Using PML-TQ engine, all the occurrences of text coreference relations in the PDT (annotated as arrows – see Figure 1) have been collected and we have examined the information structure of the sentence items (nodes in dependency trees) where the text coreference relations start and where they lead to. In other words, identifying whether the items participating in text coreference are rather contextually bound or non-bound.

4 Results and evaluation

Table 1 shows text coreference relations connecting contextually bound and non-bound sentence items (nodes) in the PDT.[4]

From the comparison of Figure 2 and 3, we may observe that among all the 86,590 text coreference relations marked in the PDT, mainly the non-contrastive

[4] The distributions of "f", "t" and "c" nodes in the PDT are presented below.

Table 1: Contextually bound and non-bound sentence items interconnected with text coreference relation in the Prague Dependency Treebank

	f (from)	t (from)	c (from)	To (in total)
f (to)	19,571	20,354	2,754	42,679
t (to)	7,980	27,109	1,762	36,851
c (to)	2,322	3,671	1,067	7,060
>From (in total)	29,873	51,134	5,583	**86,590**

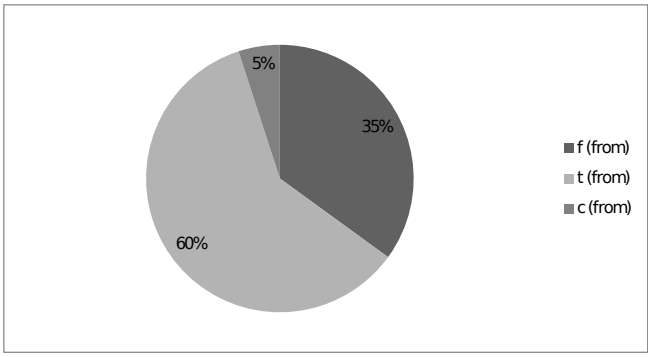

Figure 2: Percentage of individual node types participating in text coreference as the sender of the coreference arrow (its starting point)

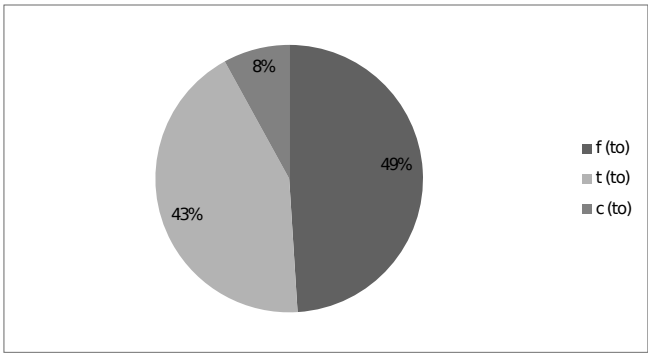

Figure 3: Percentage of individual node types participating in text coreference as the recipient of the coreference arrow (its ending point)

contextually bound sentence items ("t" nodes) (60%) are referring to the previous text (51,134 within 86,590). On the contrary, mainly the contextually non-bound sentence items ("f" nodes) (49%) serve as recipients of text coreference relations (42,679 within 86,590), see Figure 2. More specifically, if there is the coreference text relation between the words *Jane* and *she* (i.e. from *she* to *Jane*), *she* is mostly (in 60%) "t" node (i.e. non-contrastive contextually bound sentence item) and *Jane*, on the other hand, "f" node (i.e. contextually non-bound sentence item) in 49%, see Figure 3.

The particular "c", "t" and "f" node types are not distributed with the same frequency in the PDT, see Table 2 reflecting the ratio of occurrences of particular node types in the data (the PDT contains 354,841 contextually non-bound nodes ("f"), 176,225 non-contrastive contextually bound nodes ("t") and 30,312 contrastive contextually bound nodes ("c")).

Table 2: The PDT distribution of "f", "t" and "c" interconnected with a text coreference relation

%	f (from)	t (from)	c (from)
f (to)	5.52	11.55	9.09
t (to)	2.25	15.38	5.81
c (to)	0.65	2.08	3.52

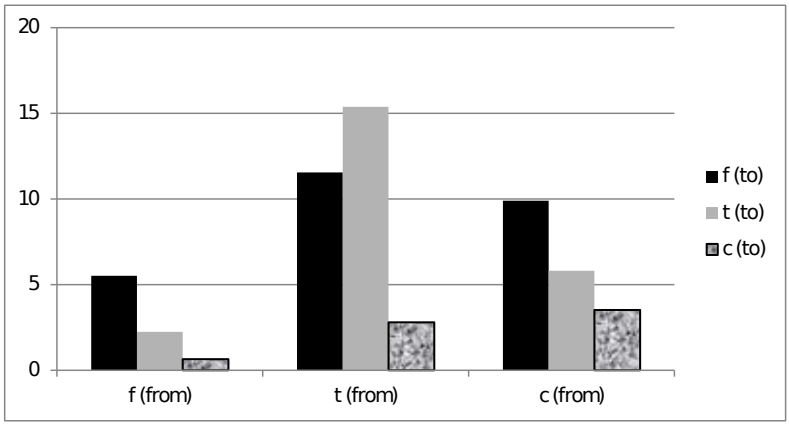

Figure 4: The PDT distribution of "f", "t" and "c" interconnected with a text coreference relation

The contextually bound nodes ("t" and "c" nodes) generally have higher probability that the text coreference arrow will lead from them and also to them than contextually non-bound nodes ("f" nodes). Based on this, the most typical text coreference connection leads from a non-contrastive contextually bound node to another non-contrastive contextually bound node (i.e. from "t" to "t"), see (2) in §3.3. The second most typical text coreference connection leads from a non-contrastive contextually bound node to a contextually non-bound node (i.e. from "t" to "f"). The third most typical text coreference connection leads from a contrastive contextually bound node to a contextually non-bound node (from "c" to "f"). Generally, the most favored "starting" position for a text coreference arrow is a non-contrastive contextually bound sentence item ("t").

Table 3: Percentage of all "f" or "t+c" nodes interlinked with a text coreference relation in the PDT

%	(from)	t+c (from)
f (to)	5.52	11.19
t+c (to)	2.90	16.27

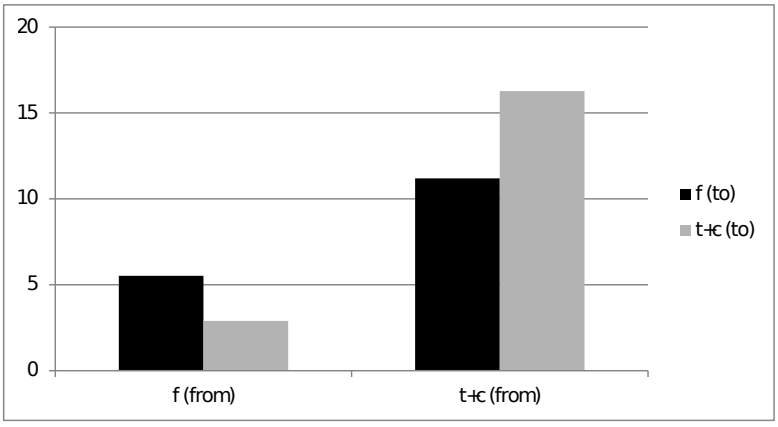

Figure 5: Percentage of all "f" or "t+c" nodes interlinked with a text coreference relation in the PDT

Contextually bound sentence items (both contrastive and non-contrastive that are mostly part of sentence Topic) are interlinked with text coreference relations more often than contextually non-bound (i.e. from the context non-deducible) items that are mostly part of sentence Focus, see Table 3 and Figure 5. Thus, the

two described language phenomena, text coreference and sentence information structure, mutually cooperate in building the text coherence.

The individual node types differ in the fact where they find their parts of coreference chains. While the non-contrastive contextually bound nodes ("t") most likely are interconnected with contextually bound nodes, the contextually non-bound nodes ("f") mostly interconnected with contextually non-bound nodes (in terms of text coreference). The contrastive contextually bound nodes stand between these two tendencies – they are connected both with contextually bound and non-bound nodes (in relatively equal way). Such inclinations also demonstrate that it is worth distinguishing two different kinds of contextually bound nodes (contrastive and non-contrastive) because they contribute to the text coherence in different ways.

The individual node types ("t", "c" and "f") have in common that they all refer to the contrastive contextually bound nodes ("c") in the slightest degree (among them, the "c" nodes have the highest tendency to be interlinked with other "c" nodes).

Table 4: Percentage of "f", "t", "c" or "t+c" nodes interlinked with a text coreference relation in the PDT

%	f	t	c	t+c
from	8.42	29.02	18.42	27.46
to	12.03	20.91	23.29	21.26

Table 4 and Figure 5 shows a percentage of bound vs. non-bound nodes participating in text coherence relations (either as "recipients" or "senders"). The biggest text coreference "recipient" and "sender" are contextually bound nodes (without further distinguishing between contrast and non-contrast) – 27.46 % within all of them (i.e. 56,717 within 206,537) serve as a "text coreference sender" and 21.26 % of them (i.e. 43,911 within 206,537) as a "text coreference recipient".

Based on the presented analysis, the following conclusions can be drawn:

- Generally, a text coreference arrow (i) starts in every 5th–6th and leads to every 4th contrastive contextually bound sentence item ("c" node);
  (ii) starts in every 3rd–4th and to every 5th non-contrastive contextually bound sentence item ("t" node) and (iii) starts in every 12th and to every 8th contextually non-bound sentence item ("f" node).

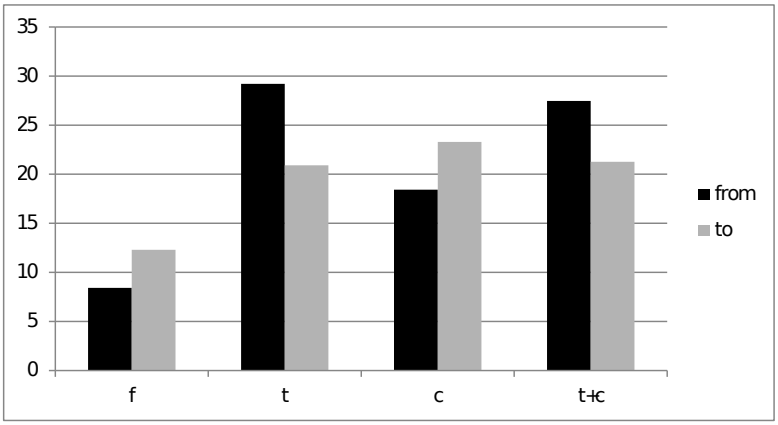

Figure 6: Percentage of "f", "t", "c" or "t+c" nodes interlinked with a text coreference relation in the PDT

- The contextually non-bound nodes ("f") as well as contrastive contextually bound ("c") nodes serve more often as text coreference "recipient" than "sender".
- Conversely, the non-contrastive contextually bound nodes ("t") serve more often as a text coreference "sender" than "recipient".

5 Conclusions

The paper has examined the correlation between sentence information structure and text coreference on the data of the Prague Dependency Treebank. Altogether, the PDT contains 86,590 text coreference relations interconnecting contextually bound or non-bound sentence items. The analysis shows that the text coreference relations operate rather within contextually bound nodes, i.e. if a sentence item is contextually bound (in terms of sentence information structure), it has a relatively high probability to be interconnected with another sentence item in a text coreference relation.

On the other hand, there is also a relatively significant part of contextually non-bound sentence items interconnected with another part of text through text coreference. The text coreference arrow leads from every 12th contextually non-bound sentence item ("f" node). It means that every 12th contextually non-bound sentence item clearly refers to the previous language context (in terms of text coreference). However, these two facts are not in contradiction. It is well known

that entities mentioned in the previous text can be used in a new perspective (i.e. as contextually non-bound items) and they can bring new and unknown information to the text addressee (cf. *Do you want tea or coffee? – Tea, please.*).

In this context, contextually bound sentence items cannot be defined simply as coreferentially referring to the previous language context. They refer to the previous text (through text coreference) clearly more often than the contextually non-bound items. However, such kind of text referring is also not rare – according to the PDT, the contextually non-bound items participate in the text coreference in about 35%, non-contrastive contextually bound items in 60% and contrastive contextually bound items in 5%.

In this respect, the corpus-based research also demonstrates that the annotation of text coreference cannot be (without further specification) a reliable basis for the automatic annotation of sentence information structure in large corpora. If every sentence item annotated as referring to the previous context (in terms of text coreference) were automatically annotated also as contextually bound, it would constitute a large degree of error (based on the data from the PDT, the error rate would be about 35%).

Acknowledgments

The author acknowledges support from the Ministry of Culture of the Czech Republic (project n. DG16P02B016 *Automatic Evaluation of Text Coherence in Czech*). This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

Ariel, Mira. 1988. Referring and accessibility. *Journal of Linguistics* 24(01). 65–87.

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek & Šárka Zikánová. 2013. *Prague Dependency Treebank 3.0*. Prague, Czech Republic: Univerzita Karlova v Praze, MFF, ÚFAL. http://ufal.mff.cuni.cz/pdt3.0/.

Chiarcos, Christian. 2014. Towards interoperable discourse annotation. Discourse features in the ontologies of linguistic annotation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 4569–4577. Reykjavik, Iceland: European Language Resources Association (ELRA).

Eckert, Miriam & Michael Strube. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics* 17(1). 51–89.

Grosz, Barbara J. & Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3). 175–204.

Hajičová, Eva, Barbara Partee & Petr Sgall. 1998. *Topic-focus articulation, tripartite structures, and semantic content*. Vol. 71. Dordrecht: Kluwer.

Hajičová, Eva. 2011. On interplay of information structure, anaphoric links and discourse relations. In. *Societas Linguistica Europaea, SLE 2011 – 44th Annual Meeting*. Javier Martin Arista (ed.). Universidad de La Rioja. 139–140.

Hajičová, Eva, Barbora Hladká & Lucie Kučová. 2006. An annotated corpus as a test bed for discourse structure analysis. In *Proceedings of the Workshop on Constraints in Discourse*, 82–89. Maynooth, Ireland: National University of Ireland.

Joshi, Aravind K. & Scott Weinstein. 1981. Control of inference: Role of Some aspects of discourse Structure-Centering. In *IJCAI*, 385–387.

Komen, Erwin R. 2012. Coreferenced corpora for information structure research. Studies in Variation, Contacts and Change in English (10).

Nedoluzhko, Anna. 2011. *Rozšířená textová koreference a asociační anafora (Koncepce anotace českých dat v Pražském závislostním korpusu)* (Studies in Computational and Theoretical Linguistics). Praha, Česká Republika: Ústav formální a aplikované lingvistiky.

Nedoluzhko, Anna. 2015. Contextually Bound Expressions without a Coreference Link. In Zikánová, Šárka, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová & Jan Václ. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. (Studies in Computational and Theoretical Linguistics). Prague, Czechia: UFAL. 199–215.

Nedoluzhko, Anna & Eva Hajičová. 2015. Information structure and anaphoric links – a case study and probe. In *Corpus Linguistics 2015. Abstract book*, 252–254. Lancaster University, UK. Lancaster, UK: UCREL.

Pajas, Petr & Jan Štěpánek. 2008. Recent advances in a Feature-Rich framework for treebank annotation. In Donia Scott & Hans Uszkoreit (eds.), *The 22nd International Conference on Computational Linguistics – Proceedings of the Conference*, vol. 2, 673–680. Manchester, UK: The Coling 2008 Organizing Committee.

Rysová, Kateřina & Magdaléna Rysová. 2015. Analyzing text coherence via multiple annotation in the Prague Dependency Treebank. In Pavel Král & Václav Matoušek (eds.), *Text, Speech, and Dialogue: 18th International Conference, TSD 2015* (Lecture Notes in Artificial Intelligence 9302), 71–79. University of West Bohemia. New York: Springer International Publishing.

Rysová, Kateřina, Magdaléna Rysová & Eva Hajičová. 2015. *Topic–focus articulation in English texts on the basis of Functional Generative Description*. Tech. rep. TR 2015-59. Prague, Czechia.

Sgall, Petr. 1967. *Generativní popis jazyka a česká deklinace*. Prague, Czech Republic: Academia.

Stede, Manfred & Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 925–929. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/579_Paper.pdf.

Štěpánek, Jan & Petr Pajas. 2010. Querying diverse treebanks in a uniform way. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, 1828–1835. Valletta, Malta: European Language Resources Association.