Topics in Linguistics (2017), 18(2), pp. 36-47

DE GRUYTER
OPEN

10.1515/topling-2017-0009

Studying text coherence in Czech – a corpus-based analysis

Magdaléna Rysová
University of Economics, Czech Republic

Abstract
The paper deals with the field of Czech corpus linguistics and represents one of various current studies analysing text coherence through language interactions. It presents a corpus-based analysis of grammatical coreference and sentence information structure (in terms of contextual boundness) in Czech. It focuses on examining the interaction of these two language phenomena and observes where they meet to participate in text structuring. Specifically, the paper analyses contextually bound and non-bound sentence items and examines whether (and how often) they are involved in relations of grammatical coreference in Czech newspaper articles. The analysis is carried out on the language data of the Prague Dependency Treebank (PDT) containing 3,165 Czech texts. The results of the analysis are helpful in automatic text annotation – the paper presents how (or to what extent) the annotation of grammatical coreference may be used in automatic (pre-)annotation of sentence information structure in Czech. It demonstrates how accurately we may (automatically) assume the value of contextual boundness for the antecedent and anaphor (as the two participants of a grammatical coreference relation). The results of the paper demonstrate that the anaphor of grammatical coreference is automatically predictable – it is a non-contrastive contextually bound sentence item in 99.18% of cases. On the other hand, the value of contextual boundness of the antecedent is not so easy to estimate (according to the PDT, the antecedent is contextually non-bound in 37% of cases, non-contrastive contextually bound in 50% and contrastive contextually bound in 13% of cases).

Key words
sentence information structure, coreference, corpus analysis, Czech

Introduction

Studying language interactions is one of the essential themes of current corpus linguistics. The analysing of relationships between various language aspects is of great importance, in particular for such complex phenomena as text coherence (see, e.g., Burke, 2016 or Povolná, 2016). Text coherence may be viewed as a network of relations of many different kinds, including coreference and anaphoric relations, discourse relations and relations in terms of sentence information structure. Any text must contain this (unbroken) network of relations so that it can be considered coherent and understandable to the reader.

In the international context, language interactions (concerning coherence and discourse analysis) have been studied by Grosz and Sidner (1986; introducing a new theory of discourse structure and focusing on examining the mutual relationship between the linguistic structure, the attentional state and intentional structure), by Long and Chong (2001; focusing on the relationship between comprehension skill and global coherence), or more recently by Camblin et al. (2007; studying the interplay of word-level and discourse-level information during sentence processing) and Ledoux et al. (2007; focusing on

coreference and lexical repetition and examining the relationship between basic processes of word recognition and higher processes involving the integration of information into a discourse model).

In the Czech context, the need to study text coherence through interactions of various language phenomena such as coreference, sentence information structure and semantic discourse relations has been formulated in many recent papers – see, e.g., Hajičová (2011), Nedoluzhko and Hajičová (2015), Rysová and Rysová (2015) etc. The present paper represents one of various studies of corpus linguistics in Czech and continues the Czech linguistic tradition of analysing coreference and sentence information structure and their role in text coherence (see especially Daneš, 1974; Hajičová, Partee and Sgall, 1998 or Hajičová, Havelka and Sgall, 2014). The need to study these phenomena at present is primarily the result of the rise of richly annotated corpora that enable linguists to verify earlier hypotheses using larger authentic data. At the same time, the new linguistic results are highly useful in computational and corpus linguistics, as they help to improve automatic data annotations.

1. Methodology and key concepts

This paper investigates the interplay of grammatical coreference and sentence information structure in Czech based on the data of the Prague Dependency Treebank (PDT, a corpus incorporating dependency grammar and the only corpus of Czech containing multiple manual annotations of coreference and sentence information structure). The paper examines the mutual relationship between these two phenomena (based on their manual annotation in the PDT) and it presents how the existing annotation of grammatical coreference may be used for improving automatic (pre-)annotation of sentence information structure in other corpora.

The PDT is a large corpus of Czech newspaper texts with almost 50,000 annotated sentences (in 3,165 documents). The PDT contains annotations (both automatic and manual) of more language levels at once: the morphological layer, analytical layer (i.e. surface syntax) and the so-called tectogrammatical layer (i.e. deep syntactico-semantic layer). At the same time, the PDT is annotated for phenomena that often go beyond the sentence boundary, such as discourse relations (i.e. annotation of semantico-pragmatic relations expressed by connectives), coreference and anaphoric relations and sentence information structure. The newest version of the PDT is PDT 3.0 (see Bejček et al., 2013).

The following subsections briefly present the key concepts connected with the PDT: the theory of Functional Generative Description (FGD) establishing the annotation scheme of the PDT, sentence information structure and grammatical coreference and their annotation in the PDT and, finally, the client-server PML Tree Query, which is used for searching the PDT corpus.

1.1 Theory of Functional Generative Description

The annotation framework of the PDT corpus is based on the well-developed theory of language description, the Functional Generative Description – FGD (see Sgall, 1967; Sgall et al., 1969; Sgall, Hajičová and Panevová, 1986 etc.). The FGD principles follow the functional approach of the Prague School and linguistic methodological requirements presented by Chomsky (1964).

The FGD is conceived of as a multi-level system going from linguistic function to linguistic form, i.e. proceeding from the deep syntactico-semantic representation of a sentence to its surface syntax, morphemic and phonemic levels down to the phonetic form of the sentence.

The FGD concentrates especially on the deep syntactico-semantic level of sentences (called tectogrammatical – a term adopted from Putnam, 1961). The tectogrammatical level captures sentences such as dependency trees whose roots are the predicates of main clauses, and the tree edges establish the dependency relations between the dependency nodes.

1.2 Sentence information structure and contextual boundness

In the international linguistic context, the phenomenon of sentence information structure is based traditionally on a dichotomy – on a distinction between two notions called variously as psychological subject and psychological predicate, theme and rheme, topic and comment, topic and focus, given and new

information etc. The crucial ideas on this issue were discussed by Weil as early as 1844, and by linguists around the Zeitschrift für Völkerpsychologie (see von der Gabelentz, 1868; Paul, 1886 or Wegener 1885; using the terms of psychological subject and psychological predicate, later criticized by Mathesius, 1907).

The research topics of sentence information structure (the term used by Steedman, 1991 or Lambrecht, 1996) in an international context include e.g. intonative structure of focalization (Le Gac and Yoo, 2002), studying interactions between information and syntactic structure (Birner and Ward, 2009), the relation between information structure and word-order variation (Petrova, 2009) or informativity of sentence information structure (Peti-Stantić, 2013).

Currently, the phenomenon of sentence information structure is also a part of many formal and empirical language descriptions. As indicated above, the annotation principles for sentence information structure for Czech in the PDT were formulated in accordance with the theory of Functional Generative Description (FGD) established by Sgall and his students and colleagues (1964) and further elaborated especially by Hajičová (e.g. 1998).[1]

The annotation in the PDT includes two types of information: all sentence items are evaluated firstly in terms of contextual boundness (they are labelled as contextually bound or non-bound, see below), and secondly in terms of communicative dynamism, i.e. the annotators mark the relative degree of importance of the individual sentence items. In general, the contextually non-bound sentence items are considered more dynamic than the contextually bound ones. In the PDT, communicative dynamism is expressed by the so-called underlying word order (order of nodes in dependency trees). More details are available in Hajičová et al. (1998), Mikulová et al. (2005) or Rysová (2014).

Contextual boundness in the PDT is considered a property of a sentence item (that may be present or absent in the surface sentence structure) which indicates whether the author uses the item as given for the recipient, i.e. deducible from the broader context. All the relevant sentence items (represented by nodes in dependency trees) are thus evaluated as one of the following three values: "t", "c" or "f", see Example (1).

(1) [*Petr je můj kolega*.] (*On*.t) *je*.f *velmi*.f *pracovitý*.f. *Ale jeho*.t *sestra*.c *je*.t *líná*.f. (*Já*.t) *mám rád*.f *spíš*.f *jeho*.f.

English translation:
'[*Peter is my colleague*.] *He*.t *is*.f *very*.f *hardworking*.f. *However, his*.t *sister*.c *is*.t *lazy*.f. *I*.t *like*.f *rather*.f *him*.f.'

The "t" value is a label for non-contrastive contextually bound nodes introducing deducible information. They can be mentioned already in the preceding (con)text (that is not necessarily verbatim) or we may easily deduce them. These items may also be related to a broader social (con)text.

The "c" value is given to contrastive contextually bound nodes that usually represent a choice from other possible alternatives (that do not have to be explicitly mentioned in the text). Their typical position is at the beginning of a sentence. In spoken form, they carry an optional contrastive stress.

The "f" value is a label for contextually non-bound nodes representing some "new" or "unknown" information (or "known" information presented from a new perspective) and they are thus non-deducible from the previous (con)text. More details are given in Hajičová et al. (1998).

Typically, contextually bound sentence items (both contrastive and non-contrastive) represent the sentence Topic whereas contextually non-bound sentence items represent the sentence Focus. Topic is defined as all contextually bound nodes directly dependent on the governing verb with all their dependents (modifiers). At the same time, Focus consists of all contextually non-bound sentence items directly dependent on the governing verb with all their dependents. The governing verb itself is a part of the sentence Topic

[1] A comparison between the FGD approach and other approaches to topic–focus articulation may be found in Hajičová (1972); Sgall, Hajičová and Benešová (1973); Sgall (1975); Hajičová, Partee and Sgall (1998) or Hajičová (2012).

or Focus according to whether it is contextually bound or non-bound.

1.3 Grammatical coreference

The paper examines the relationship between sentence information structure and grammatical coreference in the PDT data. Therefore, we adopt also the definition and principles of annotation of grammatical coreference stated for this corpus, see Nedoluzhko (2011).

The concept of (grammatical) coreference in Czech presented in Nedoluzhko (2011) is based especially on the works by Paducheva (1985), van Hoek (1995) and Langacker (2008), understanding coreference as the relation between entities realized within the world of discourse (i.e. not between word meanings as traditionally stated in classical logical semantics from which the general theory of reference originates, see Frege, 1892; Russel, 1905 or Carnap, 1947). In theoretical linguistics (especially concerning theory of communication and studies on coherence and cohesion), coreference (along with anaphora) is traditionally defined as a basic means contributing to text coherence (cohesion) – see, e.g., Halliday and Hasan (1976), Hobbs (1979), Kehler et al. (2008) etc. Many studies then focus on topics such as theory of reference, types of reference or anaphora resolution (see Hlavsa, 1975; Miodunka, 1974; Topolińska, 1984; Palek, 1988; Sorace and Filiaci, 2006 and Mitkov, 2014). Grammatical coreference (as one of the coreference subtypes) is elaborated in detail by Gross (1973), and Gordon and Hendrick (1998).

In the Czech linguistic tradition (see Hajičová, Panevová and Sgall, 1985; 1986; 1987), grammatical coreference is understood as a type of coreference where the antecedent is determinable on the basis of grammatical rules of the given language and is thus associated with the syntactic structure of sentences. At the same time, both the antecedent and anaphor of the grammatical coreference relation are placed in the same sentence.[2]

Three cases of grammatical coreference relation can be found in the following examples:

(2) ***Jana*** ***se*** *dala na melounovou dietu.*

English translation:
*'**Jane** put **herself** on a watermelon diet.'*

(3) ***Fotografie***, ***která*** *není taková, jak se na první pohled zdá*

English translation:
*'A **Photo** **That**'s Not What It First Seems to Be'*

(4) *Stále rostoucí ceny centrálně vyráběného tepla nutí* ***spotřebitele*** *hledat* [= *aby* ***oni/spotřebitelé***$_{\text{zero Actor}}$ *hledali*] *alternativní řešení.*

English translation:
*'The constantly rising prices of centrally produced heat force **consumers** to look for* [= ***they/the consumers***$_{\text{zero Actor}}$ *look for*] *alternative solutions.'*

As can be seen, the grammatical coreference occurs e.g. between nouns and reflexive pronouns (Example 2) between nouns and relative expressions (Example 3) or between nouns and zero/omitted items (often in the semantic function of Actors; Example 4). The basis of coreference lies in the fact that the antecedent (e.g. *Jane*) and anaphor (e.g. *herself*) (co)refer to the same object (e.g. to the same person).

The annotation of grammatical coreference in the PDT was carried out manually on the whole corpus data and it appears (similarly as the annotation of sentence information structure) at the tectogrammatical level of the PDT.

1.4. Example of a dependency tree from the PDT

Example 5 represents one of the occurrences of grammatical coreference in the PDT – the coreference relation leading from a non-contrastive contextually bound node to a contextually non-bound node (i.e. from "t" to "f").

(5) [*Majitelé podniku chtějí kvalitou a tradicí dosáhnout výroby piva Bernard jako značkového nápoje… Chmelové extrakty dovážejí z Německa pro vysoký obsah hořkých látek*.]

[2] With some exceptions, see Hajičová, Oliva and Sgall (1987).

*V pivovaru se snaží o profilaci piva Bernard jako **značky**, **která** není pro spotřebitele levná.*

English translation:
'[*The owners of the company want to achieve brewing Bernard beers as branded drinks with the help of quality and tradition…*
Extracts of hops are imported from Germany for their high content of bitter substances.]
*In the brewery, they try to profile Bernard Beer as a **brand** **that** is not cheap for the consumer*.'

Figure 1 (placed at the end of the paper) represents the sentence from Example 5 (in the square brackets, there is a part of the previous context). The grammatical coreference arrow leads from the node *that* that is non-contrastive contextually bound ("t") to the node *brand* that is contextually non-bound ("f").[3]

The annotations of sentence information structure and grammatical coreference in the PDT were carried out independently by two different research teams; the annotation principles and theories on both phenomena thus did not influence each other in the PDT.

1.5. PML Tree Query

To explore the relation between grammatical coreference and sentence information structure on the PDT data, we used the client-server PML Tree Query (the primary format of the PDT is called Prague Markup Language; PML-TQ, see Štěpánek and Pajas, 2010). The client part has been implemented as an extension to the tree editor TrEd (Pajas and Štěpánek, 2008).

Using the query engine, we searched for places where the grammatical coreference relations (in the PDT captured as arrows, see Figure 1) lead *from* and where they lead *to* in terms of sentence information structure. In other words, we examined whether the grammatical coreference arrows lead rather from contextually bound or non-bound items (nodes) and rather to contextually bound or non-bound items (nodes).

2. Results and evaluation

2.1 Interaction of sentence information structure and grammatical coreference in the PDT

Grammatical coreference relations operating among contextually bound (both non-contrastive and contrastive) and non-bound sentence items with their occurrences in the PDT are presented in Table 1.

For example, the abbreviation "f (to)" in the table represents a sentence item (node) that is marked as contextually non-bound and, at the same time, it is a place where a grammatical coreference relation (arrow) leads to.

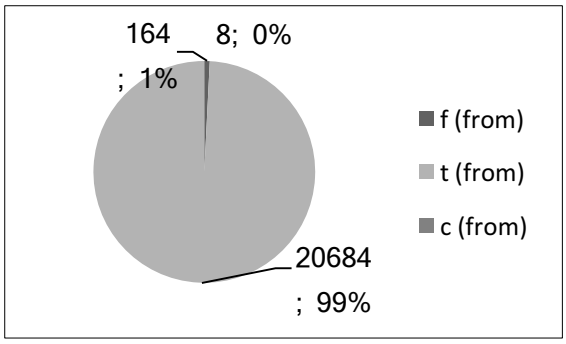
	f (from)	t (from)	c (from)	To (in total)
f (to)	22	7,687	5	7,714
t (to)	122	10,267	3	10,392
c (to)	20	2,730	0	2,750
From (in total)	164	20,684	8	**20,856**

Table 1: Interaction of contextual boundness and grammatical coreference in numbers

We can observe that the Prague Dependency Treebank contains 20,856 annotated relations expressing grammatical coreference (between sentence items that are relevant also for information structure). Table 1 shows that the relations of grammatical coreference are not distributed uniformly in the PDT. Most of them (in absolute numbers) occur between two non-contrastive contextually bound sentence items ("t" nodes in a dependency tree, see Graphs 1 and 2). Graph 1 illustrates that non-contrastive contextually bound ("t") nodes form 99% of all the items from which the grammatical coreference leads.

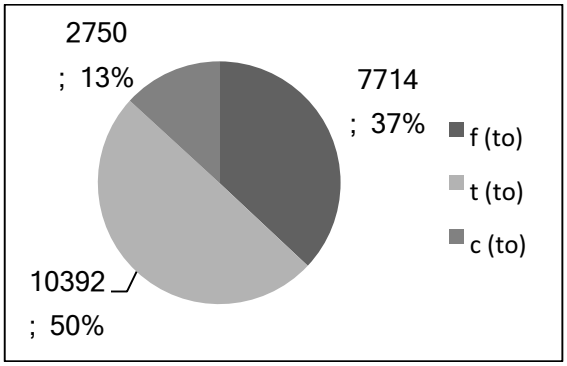
Graph 2 demonstrates that most of the relations (arrows) of grammatical coreference also lead to "t" nodes but not with such predominance as in Graph 1. The ending points of grammatical coreference relations are "t" nodes in 50% of cases; the other half is formed by "f" nodes (37%) and "c" nodes (13%).

[3] Another relation is annotated between the nodes *brand* and *beer* but the type of relation is not grammatical coreference (such cases are not taken into account in this paper).



Graph 1: Percentage of individual items participating in grammatical coreference as senders of the coreference relation (its starting point)

The most typical relations of grammatical coreference are thus a) between two "t" nodes ("t"←"t"), and b) leading from the "t" to the "f" node ("f"←"t"). This result empirically supports the idea presented by Daneš (1974:118) who describes the "t"←"t" relation as the "thematic progression with a continuous (constant) theme" and the "f"←"t" relation as the "simple linear thematic progression (or TP with linear thematization of rhemes)".



Graph 2: Percentage of individual items participating in grammatical coreference as recipient of the coreference relation (its ending point)

More specifically, if we have an example of grammatical coreference relation leading backwards from the word *herself* to the word *Jane*, we may assume (without any knowledge of context and based only on the presented results) that the item *herself* is (non-contrastive) contextually bound (with a probability of 99%, see Graph 1).

When estimating the value of contextual boundness of the item *Jane*, the probability is lower. It may be assumed that it is non-contrastive contextually bound (with a probability of 50%), less probably contextually non-bound (37%) or contrastive contextually bound (13%). If contextual boundness were not divided into contrastive and non-contrastive, i.e. if we distinguished only boundness and non-boundness, there would be a 63% probability that the item *Jane* is contextually bound (see Graph 2).

The most typical cases of grammatical coreference in the PDT are illustrated in the following examples. Example 6 demonstrates the grammatical coreference between two "t" nodes. The previous context is given in the square brackets.

(6) [*OSKAR. Za mimořádný výkon. Firmě Ilja Běhal a spol., zajišťující umělecko-kovářské a restaurátorské práce hlavně na střední Moravě.*] *Zejména v Olomouci* ***firma*****.t** ***svými*****.t** *výrobky přispívá ke zvýraznění koloritu historického jádra města.*

English translation:
'[*OSCAR. For outstanding performance. To Ilja Běhal et al. company providing artistic blacksmith and restoration work mainly in Central Moravia.*] *Especially in Olomouc, the* ***company*****.t** *contributes* ***its*****.t** *products to accentuate the atmosphere of the historic city center.'*

Example 7 illustrates the grammatical coreference relation leading from the "t" node to the "f" node, which is the second most frequent type of grammatical coreference in the PDT.

(7) [*První informaci o novém výrobku, ceně a podobně získá volající z řečové paměti.*] *Potřebuje-li další informace, je přepojen na* ***oddělení*****.f**, ***které*****.t** *jeho dotazy v co nejkratším čase zodpoví.*

English translation:
'[*The first information about a new product, price etc. is given to the caller from a speech memory.*] *If he needs additional information, he is redirected to the* ***department*****.f** ***that*****.t** *answers his questions in the shortest time.*'

Example 8 is an illustration of the minor cases of grammatical coreference, specifically of a relation leading from the "f" node to the "f" node.

(8) *A: Jak se cítíte ve své nové divadelní* ***roli*****.f***? B: A ve* ***které*****.f***? [Já teď mám za sebou dvě premiéry.]*

English translation:
'*A: How do you feel in your new* ***role*****.f** *in the theatre? B: And in* ***which*****.f***? [I have two premieres now.]'*

2.2 Statistical measurements: Chi-Square Test

This subsection presents statistical measurements carried out on the presented results. The chi-square test is used to verify the hypotheses that the above-mentioned frequencies of occurrences of the coreference relations captured in Table 1 are distributed equally among the individual "f", "t" and "c" nodes. More specifically, we verify the zero hypothesis (H_0) stating that the frequencies (occurrences found in the PDT) do not differ. Calculation of the chi-square test was performed using R-Studio. Tables 2–7 capture the following information: X-squared, the degree of freedom (df), p-value and information whether the hypothesis H_0 has been rejected or not. The calculation is made for all the options offered by Table 1, i.e. for the values f_{from}, t_{from}, c_{from}, f_{to}, t_{to} and c_{to}.

X-squared	df	p-value	H_0 rejected
124.44	2	< 2.2e-16	Yes

Table 2: f_{from} (f: 22, t: 122, c: 20)

X-squared	df	p-value	H_0 rejected
4,256.2	2	< 2.2e-16	Yes

Table 3: t_{from} (f: 7,687, t: 10,267, c: 2,730)

X-squared	df	p-value	H_0 rejected
4.75	2	0.09301	Yes

Table 4: c_{from} (f: 5, t: 3, c: 0)

X-squared	df	p-value	H_0 rejected
15,266	2	< 2.2e-16	Yes

Table 5: f_{to} (f: 22, t: 7,687, c: 5)

X-squared	df	p-value	H_0 rejected
20,043	2	< 2.2e-16	Yes

Table 6: t_{to} (f: 122, t: 10,267, c: 3)

X-squared	df	p-value	H_0 rejected
5,380.9	2	< 2.2e-16	Yes

Table 7: c_{to} (f: 20, t: 2,730, c: 0)

	f	t	c
Total number of nodes in the PDT	354,841	176,225	30,312
Involved in grammatical coreference (from)	164 (0%)	20,684 (12%)	8 (0%)
Involved in grammatical coreference (to)	7,714 (2%)	10,392 (6%)	2,750 (9%)

Table 8: Total numbers of "f", "t" and "c" nodes in the PDT and tokens and percentage of "senders" and "recipients" of grammatical coreference in the PDT

The results demonstrate that the zero hypothesis (H_0) was rejected with very high reliability in almost all the tested cases (except for c_{from} in Table 4 where the p-value is very high compared to other cases and the result cannot be considered reliable).

2.3 Ratio of occurrences of contextually bound and non-bound nodes in the PDT

Subsections 2.1 and 2.2 used the absolute numbers of tokens of grammatical coreference relations concerning "t", "c" and "f" nodes, see Table 1 and Graphs 1 and 2. However, the PDT does not contain the same (or at least similar) number of tokens of the individual types of nodes. Some of them (mainly "f") appear with significantly higher frequency than the

others (typically, "c" nodes are the rarest). For example, the PDT contains almost a 12 times lower number of "c" than "f" nodes (see Table 8).

Table 8 shows that the non-contrastive contextually bound sentence items ("t" nodes) are the "senders" of grammatical coreference relations most often (i.e. 20,684 tokens within 176,225 "t" nodes; 12% of all "t" nodes serve as grammatical coreference "senders"). An interesting result is to which type of node the relation leads most often (reflecting the fact that the "t", "c" and "f" nodes occur in the corpus with different frequencies). The most numerous "recipients" of grammatical coreference are the contrastive contextually bound nodes ("c" nodes, i.e. 2,750 tokens within 30,312 "c" nodes forming 9% of all "c" nodes serving as grammatical coreference "recipients"), see Example (9).

(9) [*Letos nebyla podle zemědělců dobrá úroda ovoce.*] *Příští* ***rok*****.c**, *ve* ***kterém*****.t** *má více pršet, by měl být lepší.*

English translation:
'[*This year, farmers did not have a good harvest of fruit.*] *The next* ***year*****.c**, ***which*****.t** *is expected to be rainy, should be better.*'

However, in absolute numbers, the most numerous "recipients" as well as "senders" of grammatical coreference are still the non-contrastive contextually bound nodes (see Table 1).

3. Implications: Usability of the results in automatic annotation of sentence information structure

Based on the results of the corpus analysis, it is possible to answer the above-stated question as to how the existing manual annotation of grammatical coreference could be used in improving automatic (pre-)annotation of sentence information structure.

The results demonstrated that the grammatical coreference almost always leads from the non-contrastive contextually bound sentence items ("t" nodes) but leads to various types of nodes ("f", "t" and "c"). 20,684 grammatical coreference arrows out of a total 20,856 grammatical coreference arrows (i.e. 99.18% of all grammatical coreference relations) start in the "t" nodes (i.e. in the non-contrastive sentence items that are deducible from the context), see Table 1. This uniformity of starting points of grammatical coreference relations can help markedly in the automatic annotation of information structure (specifically in the annotation of contextual boundness). If every node of a dependency tree from which a grammatical coreference arrow leads were automatically annotated as "t" node (i.e. as non-contrastive contextually bound) and then the correct value of contextual boundness of these nodes were checked by human annotators, the estimated error rate of this procedure would be less than 1%. Thus, the manual annotation of grammatical coreference can assist in the automatic annotation of sentence information structure.

For example, some other Prague corpora like the Prague Czech English Dependency Treebank (PCEDT) contain a manual annotation of coreference but not of sentence information structure. The existing annotation of coreference could thus serve well in automatic (pre)annotation of sentence information structure in this treebank.

4. Conclusions

The paper examined the mutual interaction of grammatical coreference and sentence information structure in Czech. The Prague Dependency Treebank contains 20,856 grammatical coreference relations (between nodes relevant in information structure). As demonstrated in the analysis (based on statistical measurements), the grammatical coreference and sentence information structure meet especially at one point – if the sentence item refers to some of the previously mentioned sentence items in the sense of grammatical coreference (like e.g. *brand*←*that*), there is a very high probability (more than 99%) that the anaphor (e.g. *that*) will be (non-contrastive) contextually bound.

This result specifies the concept of anaphor in grammatical coreference (not only) in the PDT. Apart from the fact that an anaphor is a part of coreferential relation that provides grammatical information about the previous antecedent, it also has a very strong tendency to be (non-contrastive) contextually bound in terms of the sentence information structure.

This information may be used in improving automatic (pre-)annotation of sentence information structure. On the other hand, the value of contextual boundness of the antecedent (e.g. *brand*) is not so easy to estimate (according to the PDT, the antecedent is contextually non-bound in 37% of all cases, non-contrastive contextually bound in 50% and contrastive contextually bound in 13%).

The paper tried to demonstrate the possibilities of analysing language interactions in the Prague Dependency Treebank, which represents an indispensable step towards studying such complex phenomena as text coherence and text understanding.

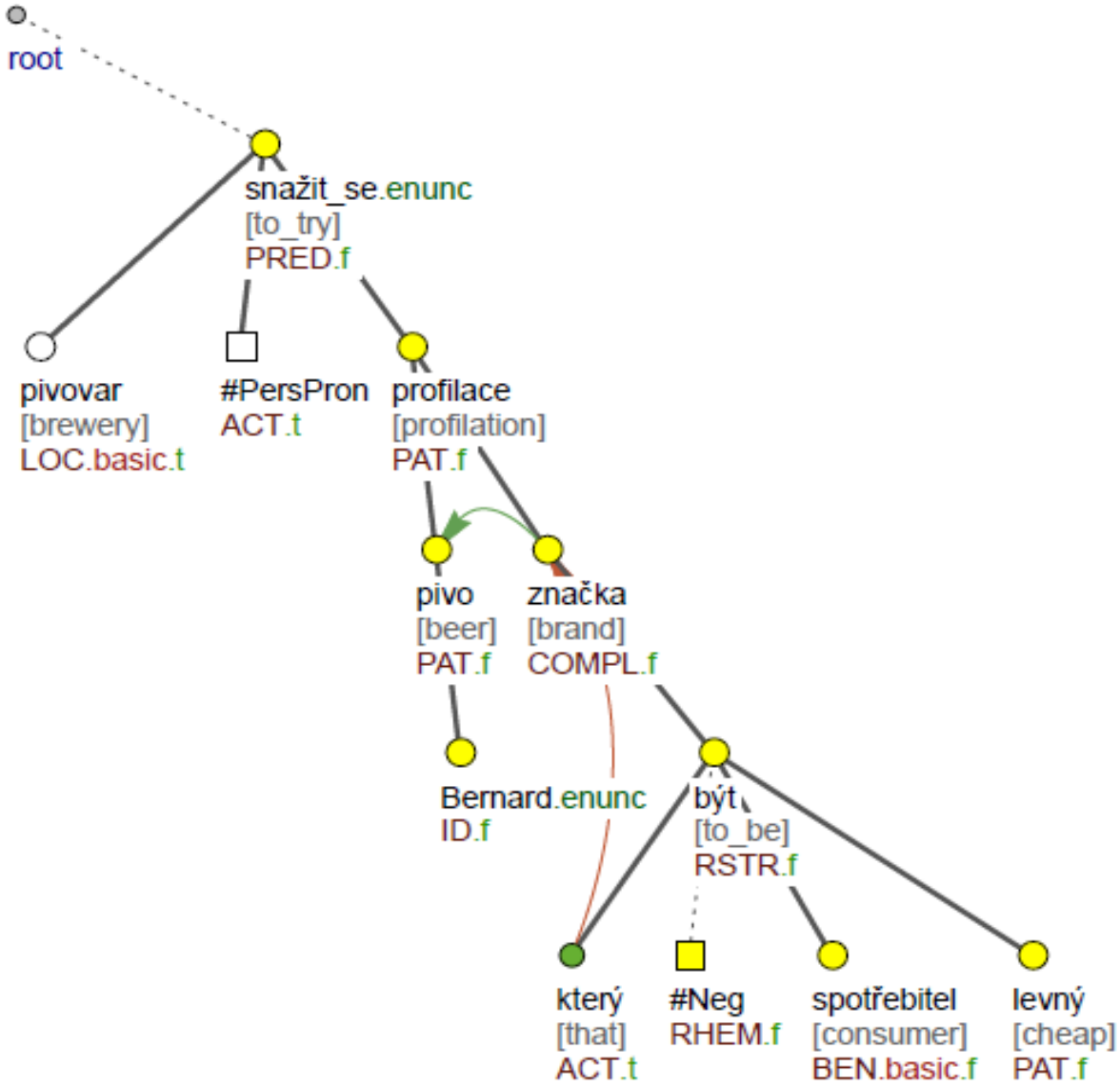


Figure 1: Dependency tree from the Prague Dependency Treebank representing the sentence *In the brewery, they try to profile Bernard Beer as a brand that is not cheap for the consumer*

References

BEJČEK, E. et al., 2013. *Prague Dependency Treebank 3.0*. Data/software. Charles University in Prague, MFF, ÚFAL, Prague, Czech Republic, http://ufal.mff.cuni.cz/pdt3.0/.

BIRNER, B. J. and WARD, G., 2009. Information structure and syntactic structure. *Language and Linguistics Compass,* vol. 3, no.4, pp. 1167–1187.

BURKE, M., 2016. Discourse implicature, Quintilian and the Lucidity Principle: Rhetorical phenomena in pragmatics. *Topics in Linguistics,* vol.17, no. 1, pp. 1–16.

CAMBLIN, C. Ch., GORDON, P.C. and SWAAB, T.Y., 2007. The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, vol. 56, no.1, pp. 103–128.
CARNAP, R., 1947. *Meaning and necessity*. Chicago: University of Chicago Press.
CHOMSKY, N., 1964. *Aspects of the theory of syntax*. Massachusetts Inst. of Tech. Cambridge Research Lab of Electronics.
DANEŠ, F., 1974. FSP and the organization of the text. In: *Papers on Functional Sentence Perspective*, pp. 106–128. Prague: Academia.
FREGE, G., 1892. Über Sinn und Bedeutung. *Zeitschift für Philosophie und philologische Kritik*, 100, pp. 25–50.
GORDON, P. C. and HENDRICK, R., 1998. Dimensions of grammatical coreference. In: *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pp. 424–429.
GROSS, M., 1973. On grammatical reference. *Generative grammar in Europe*. Springer Netherlands, pp. 203–217.
GROSZ, B. J. and SIDNER, C.L., 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, vol.12, no. 3, pp. 175–204.
HAJIČOVÁ, E., 1972. Some remarks on presuppositions. *The Prague Bulletin of Mathematical Linguistics*, vol.17, pp. 11–23.
HAJIČOVÁ, E., 2011. On interplay of information structure, anaphoric links and discourse relations. In: *Societas linguistica europaea SLE 2011, 44th Annual Meeting, Book of Abstracts*. Universidad de la Rioja, Center for Research in the Applications of Language, Logrono, pp. 139–140.
HAJIČOVÁ, E., 2012. Topic–focus revisited (Through the eyes of the Prague Dependency Treebank). In: J. D. Apresjan, ed. *Smysly, teksty i drugie zachvatyvajuščie sjužety. Sbornik statej v čest 80-letija Igorja Aleksandroviča Melčuka. Moscow: Jazyky slavjanskoj kultury*, pp. 218–232.
HAJIČOVÁ, E., PARTEE, B.H. and SGALL, P., 1998. *Topic-focus articulation, tripartite structures and semantic content*. Kluwer, Dordrecht.
HAJIČOVÁ, E., HAVELKA, J. and SGALL, P., 2014. Topic and focus, anaphoric relations and degrees of salience. Accepted for publication *Prague Linguistic Circle Papers*, vol.2, no.4, John Benjamins Publishing Company, Amsterdam.
HAJIČOVÁ, E., PANEVOVÁ, J. and SGALL, P. 1985. Coreference in the grammar and in the text. Part I. *The Prague Bulletin of Mathematical Linguistics,* vol.44, pp. 2–22.
HAJIČOVÁ, E., PANEVOVÁ, J. and SGALL, P. 1986. Coreference in the grammar and in the text. Part II. *The Prague Bulletin of Mathematical Linguistics,* vol. 46, pp. 1–11.
HAJIČOVÁ, E., PANEVOVÁ, J. and P. SGALL, P. 1987. Coreference in the grammar and in the text. Part III. *The Prague Bulletin of Mathematical Linguistics,* vol. 48, pp. 3–12.
HAJIČOVÁ, E., OLIVA, K. and SGALL, P. 1987. Odkazování v gramatice a v textu [Coreference in the grammar and in the text]. *Slovo a slovesnost*, vol.48, no.3, pp. 199–212.
HALLIDAY, M. A. K. and HASAN, R., 1976. *Cohesion in English*. London: Longman.
HLAVSA, Z., 1975. *Denotace objektu a její prostředky v současné češtině*. Vol. 10. Acad. Naklad.
HOBBS, J. R., 1979. Coherence and coreference. *Cognitive Science*, vol.3, no.1, pp. 67–90.
KEHLER, A., 2002. *Coherence, reference, and the theory of grammar*. Stanford: CSLI Publications.
LAMBRECHT, K., 1996. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge, UK: Cambridge University Press.
LANGACKER, R., 2008. *Cognitive grammar: A basic introduction*. New York: Oxford University Press.
LE GAC, D. and YOO, H.Y.,2002. Intonative structure of focalization in French and Greek. *Amsterdam Studies in the Theory and History of Linguistic Science,* 4, pp. 213–232.
LEDOUX, K., GORDON, P. C., CAMBLIN, C.C. and SWAAB, T.Y., 2007. Coreference and lexical repetition: Mechanisms of discourse integration. *Memory & Cognition*, vol.35, no.4, pp. 801–815.
LONG, D. L and CHONG, J. L., 2001. Comprehension skill and global coherence: A paradoxical picture of poor comprehenders abilities. *Journal of Experimental Psychology Learning, Memory and Cognition*, vol. 27, pp. 1424–1429.

MATHESIUS, V., 1907. Studie k dějinám anglického slovosledu [A study on history of English word order]. *Věstník České akademie*, vol.16, no.1, pp. 261–265.
MIKULOVÁ, M. et al., 2005. *Annotation on the tectogrammatical layer in the Prague Dependency Treebank.* Praha: Universitas Carolina Pragensia. <http://ufal.mff.cuni.cz/pdt2.0/browse/doc/manuals/en/t-layer/html/>.
MIODUNKA, W., 1974. *Funkcje* zaimków w grupach nominalnych współczesnej polszczyzny mówionej. *Zesz. Nauk. UJ. Prace językoznawcze, Zesz.* 43. Krakow: PWN.
MITKOV, R., 2014. *Anaphora resolution*. Routledge.
NEDOLUZHKO A. and HAJIČOVÁ, E., 2015. Information structure and anaphoric links – a case study and probe. In: *Corpus Linguistics 2015. Abstract book*. Lancaster: Lancaster University, pp. 252–254.
NEDOLUZHKO, A., 2011. *Extended nominal coreference and bridging anaphora (An approach to annotation of Czech data in the Prague Dependency Treebank*). Prague: ÚFAL.
PADUCHEVA, E., 1985. *Vyskazyvanie i ego sootnesennost' s dejstvitel'nost'ju* [The utterance and its realization in the text]. Moscow: Nauka.
PAJAS, P. and ŠTĚPÁNEK, J., 2008. Recent advances in a feature-rich framework for treebank annotation. In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008).* Manchester, pp. 673–680.
PALEK, B., 1988. *Referenční výstavba textu*. Univerzita Karlova, Praha.
PAUL, H., 1886. *Prinzipien der Sprachgeschichte*. Halle: Max Niemeyer.
PETI-STANTIĆ, A., 2013. Informativity of sentence information structure: The role of word order. *Language as Information*, pp. 155–178.
PETROVA, S., 2009. Information structure and word order variation in the Old High German Tatian. *Information structure and language change: New approaches to word order variation in Germanic*, pp. 251–280.
POVOLNÁ, R., 2016. A cross cultural analysis of conjuncts as indicators of the interaction and negotiation of meaning in research articles. *Topics in Linguistics,* vol.17, no.1, pp. 45–63.
PUTNAM, H., 1961. Some issues in the theory of grammar. In: R. Jakobson, ed. *The structure of language and its mathematical aspects. Proceedings of Symposia in Applied Mathematics*. Providence: American Mathematical Society, pp. 25–42.
RUSSEL, B., 1905. On denoting. *Mind*, vol.14, no.56, pp. 479–493.
RYSOVÁ, K., 2014. *O slovosledu z komunikačního pohledu* [On word order from the communicative point of view]. Prague: ÚFAL.
RYSOVÁ, K. and RYSOVÁ, M. 2015. Analyzing text coherence via multiple annotation in the Prague Dependency Treebank. In: *Lecture Notes in Computer Science, No. 9302, Text, Speech, and Dialogue: 18th International Conference, TSD 2015*. Cham / Heidelberg / New York / Dordrecht / London: Springer International Publishing, pp. 71–79.
SGALL, P., 1964. Generativní systémy v lingvistice [Generative systems in linguistics]. *Slovo a slovesnost*, vol.25, no.4, pp. 274–282.
SGALL, P., 1967. Functional sentence perspective in a generative description of language. *Prague Studies in Mathematical Linguistics*, 2, pp. 203–225.
SGALL, P., 1975. On the nature of topic and focus. In: H. Ringbom, ed. *Style and text (Studies Presented to Nils Erik Enkvist)*. Stockholm: Scriptor, pp. 409–15.
SGALL, P., HAJIČOVÁ, E. and BENEŠOVÁ, E. 1973. *Topic, focus and generative semantics*. Kronberg/Taunus: Scriptor.
SGALL, P., HAJIČOVÁ, E. and PANEVOVÁ, J. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Dordrecht: Reidel Publishing Company.
SGALL, P., NEBESKÝ, L., GORALČÍKOVÁ, A. and HAJIČOVÁ, E., 1969. *A functional approach to syntax in generative description of language*. New York: American Elsevier Publishing Company.
SORACE, A. and FILIACI, F., 2006. Anaphora resolution in near-native speakers of Italian. *Second Language Research,* vol.22, no. 3, pp. 339–368.
STEEDMAN, M., 1991. Structure and intonation. *Language*, vol.67, no.2, pp. 260–296.
ŠTĚPÁNEK, J. and PAJAS, P., 2010. Querying diverse treebanks in a uniform way. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010).* European Language Resources Association, Valletta, Malta, pp. 1828–1835.

TOPOLIŃSKA, Z., 1984. Składnia grupy imiennej. *Gramatyka współczesnego języka polskiego*, pp. 301–384.
VAN HOEK, K., 1995. Conceptual reference points: A cognitive grammar account of pronominal anaphora constraints. *Language*, pp. 310–340.
VON DER GABELENTZ, G., 1868. Ideen zu einer vergleichenden Syntax – Wort- und Satzstellung. *Zeitschrift für Völkerpsychologie und Sprachwissenschaft*, vol.6, no.1, pp. 376–384.
WEGENER, P., 1885. *Untersuchungen über die Grundfragen des Sprachlebens*. Amsterdam: Benjamins.
WEIL, H., 1844. *Question de grammaire générale: de l'ordre des mots dans les langues anciennes comparées aux langues modernes (thèse française)*. Paris: Joubert. Translated by Charles W. Super as Weil, H. 1887. *The order of words in the ancient languages compared with that of the modern languages*. Boston: Ginn.

Author's address and contact details
PhDr. Magdaléna Rysová, PhD.
Department of English
Faculty of International Relations
University of Economics
W. Churchill Sq. 4
Prague
Czech Republic
E-mail: magdalena.rysova@post.cz