Analyzing Text Coherence via Multiple Annotation in the Prague Dependency Treebank

Kateřina Rysová(✉) and Magdaléna Rysová

Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Charles University in Prague, Malostranské náměstí 25, 118 00 Praha, Czech Republic
{rysova,magdalena.rysova}@ufal.mff.cuni.cz
https://ufal.mff.cuni.cz

Abstract. Corpus-based research demonstrates an existence of a mutual interaction of bridging anaphoric relations in the text and sentence information structure. The research is carried out on large corpus data of the Prague Dependency Treebank 3.0 that contains almost 50 thousand sentences with manual annotation of both sentence information structure and bridging anaphora. We investigate in which way the bridging anaphora relations interconnect contextually bound and non-bound sentence items and how such types of connections contribute to the text coherence.

Keywords: Text coherence · Bridging anaphora · Sentence information structure · Topic-focus articulation · Prague dependency treebank

1 Introduction

The paper investigates the relation between two language phenomena: sentence information structure and bridging anaphoric text relations. Both of them have been studied as individuals in many research papers but in mutual interaction, they have been investigated only in the last recent years: so far, the theme of their interplay (on large corpus data) is elaborated especially by Hajičová [1], [2] or [3]. Hajičová principally deals with the relation between sentence information structure and coreference (and anaphora) and discourse relations. She analyzes e.g. under which circumstances, the anaphoric links lead from the sentence items that are contextually non-bound in terms of sentence information structure. In doing so, Hajičová emphasizes the need of complex text study, i.e. the need of exploration of the mentioned language phenomena in cooperation and mutual interaction because the text coherence results from the interplay of the individual intra- and inter-sentential phenomena.

For studying text coherence that covers several language areas or phenomena, it is necessary to use language data annotated on multiple language levels and planes. Nowadays, there are some corpora and computer programs enabling to see a mutual interaction of more individual language phenomena in a text at once, see [4], [5] or [6].

P. Král and V. Matoušek (Eds.): TSD 2015, LNAI 9302, pp. 71–79, 2015.
DOI: 10.1007/978-3-319-24033-6_8

One of the richest corpora (i.e. corpora with various types of annotation) is the Prague Dependency Treebank (PDT) [7]. It contains language annotation on morphological, analytical (surface syntactic) and tectogrammatical (deep syntactic) levels and includes, among others, manual annotation of sentence information structure, coreference and anaphoric relations, text genres and discourse relations. Therefore, the Prague Dependency Treebank is an ideal data source for our investigation of the interplay between bridging anaphoric relations and sentence information structure. In the paper, we examine the texts in Czech, but our methods may be used also for other languages in similarly annotated corpora.

2 Aim of Work

The main aim of the paper is 1) to find out whether and how the two language phenomena cooperate in a text; 2) to demonstrate linguistic and computational methods that may be used for further research of the language phenomena interplay; 3) to contribute to the general discussion on text coherence, i.e. how these two phenomena participate in text coherence.

When starting our analysis, we have concentrated on several crucial issues or questions. One of the most important is whether the bridging anaphoric links connect rather contextually bound sentence members (mutually) or rather contextually non-bound sentence members (mutually) or whether they rather interconnect both, i.e. contextually bound sentence members with the contextually non-bound sentence members.

Another aspect we focused on is whether the bridging anaphoric links operate within a Topic and Focus of the sentence in the same way.

Our assumption is that there will be more bridging anaphoric relations leading from the contextually bound sentence members (contrastive and non-contrastive, see the Section Annotation of Sentence Information Structure) looking for the connection with the previous (con)text than leading from the contextually non-bound sentence members. The reason is, in very simple terms, that the contextually bound sentence items often bring old and known information (deducible from the previous (con)text) while the contextually non-bound items often bring new and unknown (non-deducible) information.

3 Language Material – the Prague Dependency Treebank

To answer the above questions, we have analyzed the data from the Prague Dependency Treebank 3.0 (containing almost 50,000 sentences: 833,195 word tokens in 3,165 documents), a multilayer annotated corpus of Czech newspaper texts. As mentioned above, PDT contains various types of language annotations: among others, also manual annotation of sentence information structure and manual annotation of bridging anaphora.

3.1 Annotation of Sentence Information Structure

The annotation of sentence information structure in PDT is based on the theory of Functional Generative Description (FGD) [8]. During the annotation, the sentence items (nodes in a dependency tree) have been labeled as one of these three options:

a) non-contrastive contextually bound nodes (marked as t)[1];
b) contrastive contextually bound nodes (marked as c)[2];
c) contextually non-bound nodes (marked as f)[3].

For the examples of t , c and f nodes, see Figure 1. Contextually bound sentence items (contrastive or non-contrastive) are typical members of the sentence Topic. Contextually non-bound sentence items are typical members of the sentence Focus. For more details, see the annotation manual [9].

3.2 Annotation of Bridging Anaphora

Annotation of bridging anaphora (some authors use also other terms like indirect anaphora or associative anaphora) in PDT was carried out according to Nedoluzhko [10]. Bridging anaphoric relations in PDT annotation are considered the semantic or pragmatic relations between non-coreferential entities (nodes in a dependency tree) that participate in the text coherence. PDT contains the following types of such relations [11]: PART – WHOLE (e.g. *room – ceiling*), SUBSET – SET (*students – some students*), FUNCTION (*state – president*), CONTRAST (for coherence relevant discourse opposites; e.g. *this year – last year*), ANAF (for explicit anaphoric relations without coreference or one of the semantic relations mentioned above; e.g. *rainbow – that word*), REST (further underspecified group). The bridging anaphora contains both inter- and intra-sentential relations.

[1] Non-contrastive contextually bound expressions are expressions (both expressed and absent in the surface structure of the sentence) that introduce in the text some given information . Such expressions are repeated from the preceding text (not necessarily verbatim), they are deducible from it (e.g. using coreferential or inferential relations), or somehow related to a broader context. [9]

[2] A contrastive contextually bound expression is usually a choice from a set of alternatives. This set need not be explicitly specified in the text. A contrastive contextually bound expression can refer to a larger text segment and does not have to be deducible from the immediately preceding textual context. [...] The occurrence of a contrastive contextually bound expression is primarily determined by the thematic structure (progression) of the text. Contrastive contextually bound expressions usually occur in enumerations, at the beginning of paragraphs etc. In the spoken form of an utterance the contrastive contextually bound expression carries an optional contrastive stress. [9]

[3] Contextually non-bound expressions are expressions (both expressed and absent in the surface structure of the sentence) that represent in the text some unknown, new facts, or introduce known facts in new relations, i.e. they express information not deducible from context. [9]

Table 1. Numbers of occurrences of contextually bound (contrastive and non-contrastive: c , t) and non-bound (f) sentence items interlinked with bridging anaphoric relation (in the Prague Dependency Treebank).

	f (from)	**t (from)**	**c (from)**	**To (in total)**
f (to)	12,485	4,428	2,095	19,008
t (to)	7,091	4,348	1,720	13,159
c (to)	2,248	809	639	3,696
From (in total)	21,824	9,585	4,454	35,863

4 Methods

The aim of our work is to find out how the bridging relations in text correspond to the sentence information structure.

We may imagine the bridging relation as an arrow with two important aspects concerning sentence information structure: 1) where the bridging arrow leads FROM (i.e. whether rather from contextually bound or non-bound sentence items) and 2) where it leads TO (i.e. whether rather to contextually bound or non-bound sentence items).

To answer both questions, we have compiled a table (see Table 1) expressing all mutual possibilities of how many bridging anaphoric relations occur among contextually bound (non-contrastive; contrastive) and contextually non-bound sentence items.

5 Results

The main results of our analysis, i.e. the occurrences of bridging anaphoric relations connecting contextually bound and non-bound sentence items (nodes) in PDT in all possible combinations, are captured in Table 1.[4]

Table 1 demonstrates, for example, that a bridging anaphoric arrow leads from the sentence item (node) that is non-contrastive contextually bound (in PDT marked as t) to the sentence item (node) that is contextually non-bound (in PDT marked as f) in 4,428 cases.

Figure 1 shows an authentic PDT example of this combination – see Example (1) in the plain text:

(1) *Zpívají o nich v* ***písních****.*(f) *, např. i jedna z nejznámějších* [***písní****.*(t)] – *Čhajori romani – má sloku o utrpení Romů v koncentračních táborech.*
(In English: They sing about them[5] in **songs**.(f), e.g. one of the most famous [**songs**.(t)] – Čhajori romani – has a stanza about Gypsy suffering in concentration camps.)[6]

[4] For overall distributions of f , t and c nodes in PDT, see below.

[5] About the events of World War II.

[6] The preceding context is: [...] The Gypsies were affected by disaster during the World War II. [...] Events of that time are still a trauma for few gypsy survivors and their descendants.

A bridging anaphoric arrow leads from the node representing the lemma *song* (that is omitted in the surface structure but present in the deep (underlying) sentence structure) to the node representing the lemma *song* (present both in the surface and deep sentence structure).[7]

A bridging arrow starts in the node that is obviously contextually bound (known for the reader even to such extent that it is omitted from the surface word order) and points at the node that is contextually non-bound (in this case, the first occurrence of the word *song* is a part of the sentence Focus).

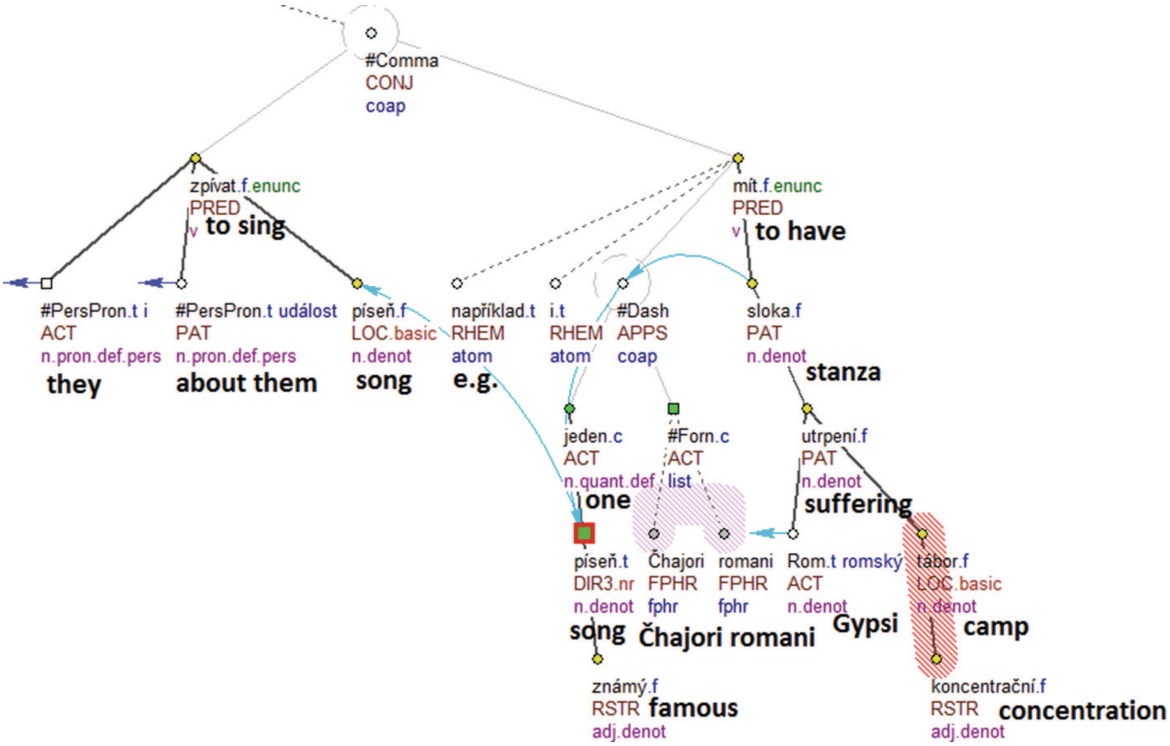

Fig. 1. PDT dependency tree representing the sentence from Example (1).

Such type of bridging relations in text is well expectable: contextually bound sentence items are connected to the preceding contextually non-bound items, i.e. the new information from one clause is repeated as the old information in the following clause where it is further elaborated. The text coherence often benefits right from this changing of the new information into the old . However, this type of text relations is not the main one, see the Table 1.

The most often type of bridging relation (in absolute numbers) is the relation between two contextually non-bound sentence items. The fact that this kind of bridging text connection is so common is quite surprising. On the other hand, although the sentence item brings new information, it can be also interlinked with other places of the text (with other sentence and text items). The general

[7] Another light blue bridging anaphoric arrow connects the nodes *stanza* and *song* in the picture.

Table 2. How many % of all f or t or c nodes are interlinked with a bridging anaphoric relation (in PDT).

	f (from)	**t (from)**	**c (from)**
f (to)	3.52	2.51	6.91
t (to)	2.00	2.47	5.67
c (to)	0.63	0.46	2.11

text coherence results exactly from this interlinking of various text items. In this connection, we can see that the text relations are complex relations created by more language phenomena in interplay.

However, the individual node types (c, t, f) do not occur in PDT with the same frequency. To find out the density of bridging anaphora relations within the individual node types (c, t, f), see the Table 2.
In PDT, there are (in total):

a) 354,841 contextually non-bound nodes (f);
b) 176,225 non-contrastive contextually bound nodes (t) and
c) 30,312 contrastive contextually bound nodes (c).

Among all of them, there are 35,863 bridging anaphoric arrows. The research results demonstrate that the distribution of bridging anaphoric relations among the individual node types (c, t, f) is not uniform.

The Table 2 demonstrates that in 3.52 % of all contextually non-bound nodes (i.e. 12,485 tokens within 354,841 f nodes), the bridging anaphoric arrow leads to another contextually non-bound node (i.e. from f to f). Another interesting result is that in 6.91 % of all contrastive contextually bound nodes (i.e. 2,095 tokens within 30,312 c nodes), the bridging anaphoric arrow leads to a contextually non-bound node (i.e. from c to f).

In general, the Table 2 shows that the contrastive contextually bound nodes (c nodes) have the highest probability and chance within all the nodes that the bridging anaphoric arrow will lead from them. In this respect, we may state the following main points gained from the results of our analysis:

1. The most typical bridging anaphoric connection leads from a contrastive contextually bound node to a contextually non-bound node (i.e. from c to f).
2. The second most typical bridging anaphoric connection leads from a contrastive contextually bound node to a non-contrastive contextually bound node (i.e. from c to t).
3. The third most typical bridging anaphoric connection leads from a contextually non-bound node to a contextually non-bound node (from f to f).
4. In general, the most favorite starting position for a bridging anaphoric arrow is a contrastive contextually bound sentence item (c).

Table 3 demonstrates which kinds of sentence items (from the perspective of sentence information structure) have the highest tendency to be the recipient and the sender of a bridging anaphoric relation. For example, 6.15 % within all contextually non-bound sentence items (i.e. 21,824 within 354,841) serve as a

Table 3. How many % of all f , t , c or t+c nodes are providing (to) or looking for (from) a bridging anaphoric relation (in PDT).

	f	**t**	**c**	**t+c**
from	6.15	5.44	14.69	6.80
to	5.36	7.47	12.19	8.16

bridging sender and 5.36 % of them (i.e. 19,008 within 354,841) as a bridging recipient . Thus, on the basis of our analysis, we came to the following points:
5. The bridging anaphoric arrow leads from every 7th and to every 8th contrastive contextually bound sentence item (c node);
6. the bridging anaphoric arrow leads from every 18th and to every 13th non-contrastive contextually bound sentence item (t node) and
7. the bridging anaphoric arrow leads from every 16th and to every 18th contextually non-bound sentence item (f node).
8. It is quite surprising that the contextually non-bound sentence items (f nodes; bringing typically the information that is non-deducible from the previous context) look for a bridging relation in the previous text even more often than non-contrastive contextually bound items (t nodes; bringing typically the information that is deducible from the previous context).

If we divide the items only into contextually non-bound and bound (without the contrastive and non-contrastive distinction), the proportion between contextually bound and non-bound nodes searching for the relation in the previous text is nearly balanced (6.15 % within all f nodes and 6.8 % within all t+c nodes are the starting position for the bridging anaphoric relation).

6 Conclusion

In PDT, we have found 35,863 bridging anaphoric relations interconnecting contextually bound or non-bound sentence items. The results of the research demonstrate that the bridging anaphoric relations are not uniformly distributed within them. Some types of sentence items (from the perspective of the sentence information structure) have a greater ability to attract bridging anaphoric relation than the other. This proves that the language phenomena of sentence information structure and bridging anaphora are closely interdependent – if the sentence item has a role of a contrastive contextually bound node in sentence information structure, there is a relatively high probability that it will be interconnected with other sentence items in the text (in sense of bridging anaphora relations).

The greatest ability to be a part of bridging anaphoric chains is proved at contrastive contextually bound sentence items. These items are the most favorite starting as well as landing positions for bridging anaphoric arrows. Among them (i.e. among the c nodes), there is the greatest density of bridging anaphora relations.

The contrastive contextually bound nodes serve as the most favorite sources of items to which the other (i.e. following) sentence items anaphorically refer

(in PDT, every 8th within all c nodes serves as a recipient of bridging relations, i.e. as a landing destination of bridging anaphora arrow). At the same time, they have also the highest tendency to look for a bridging relation in the previous (con)text (in PDT, every 7th within all c nodes serves as a sender of bridging relation, i.e. as a starting destination of a bridging anaphora arrow).

Since the contrastive contextually bound items appear in the initial sentence position or near to it very often, we may assume that the sentence beginnings are very important places of text coherence realized by bridging anaphora. Therefore, the contrastive contextually bound items may be seen as a significant pillar and backbone of the text coherence expressed by bridging anaphora.

In the paper, we tried to present how we may use the multilayer corpus data to demonstrate the crucial aspects of interplay of different language phenomena like sentence information structure and bridging anaphora, which could improve or deepen our general knowledge of text coherence.

Acknowledgments. The authors acknowledge support from the Czech Science Foundation (GACR; project n. P406/12/0658) and from the Ministry of Education, Youth and Sports of the Czech Republic (project n. LM2010013 LINDAT/CLARIN and n. LH14011).

References

1. Hajičová, E., Hladká, B., Kučová, L.: An annotated corpus as a test bed for discourse structure analysis. In: Proceedings of the Workshop on Constraints in Discourse, Maynooth, Ireland, pp. 82–89. National University of Ireland, National University of Ireland (2006)
2. Hajičová, E.: On interplay of information structure, anaphoric links and discourse relations. In: Societas Linguistica Europaea, SLE 2011, 44th Annual Meeting, Book of Abstracts, Logrono, Spain, pp. 139–140. Universidad de la Rioja, Center for Research in the Applications of Language, Universidad de la Rioja, Center for Research in the Applications of Language (2011)
3. Kučová, L., Veselá, K., Hajičová, E., Havelka, J.: Topic-focus articulation and anaphoric relations: a corpus based probe. In: Heusinger, K., Umbach, C. (eds.) Proceedings of Discourse Domains and Information Structure Workshop, pp. 37–46. Edinburgh, Scotland (2005)
4. Komen, E.R.: Coreferenced corpora for information structure research. Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources (Studies in Variation, Contacts and Change in English 10) (2012)
5. Stede, M., Neumann, A.: Potsdam commentary corpus 2.0: annotation for discourse research. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14), pp. 925–929 (2014)
6. Chiarcos, C.: Towards interoperable discourse annotation. Discourse features in the ontologies of linguistic annotation. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14), pp. 4569–4577 (2014)

7. Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., Zikánová, Š.: Prague dependency treebank **3** (2013)
8. Hajičová, E., Sgall, P., Partee, B.: Topic-focus articulation, tripartite structures, and semantic content. Kluwer, Dordrecht (1998). ISBN 0-7923-5289-0
9. Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z., Kučová, L.: Anotace na tektogramatické rovině pražského závislostního korpusu. anotátorská příručka. Technical Report TR-2005-28 (2005)
10. Nedoluzhko, A.: Rozšířená textová koreference a asociační anafora (Koncepce anotace českých dat v Pražském závislostním korpusu). Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha (2011)
11. Nedoluzhko, A.: Generic noun phrases and annotation of coreference and bridging relations in the prague dependency treebank. In: 51st Annual Meeting of the Association for Computational Linguistics Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, Sofija, Bulgaria, pp. 103–111. Bălgarska akademija na naukite, Omnipress, Inc (2013)