Anaphoric Connectives and Long-Distance Discourse Relations in Czech

Lucie Poláková and Jiří Mírovský

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
[polakova|mirovsky]@ufal.mff.cuni.cz

Abstract. This paper is a linguistic as well as technical survey for the development of a shallow discourse parser for Czech. It focuses on long-distance discourse relations signalled by (mostly) anaphoric discourse connectives. Proceeding from the division of connectives on "structural" and "anaphoric" according to their (in)ability to accept distant (non-adjacent) text segments as their left-sided arguments, and taking into account results of related analyses on English data in the framework of the Penn Discourse Treebank [3, 11], we analyze a large amount of language data in Czech. We benefit from the multilayer manual annotation of various language aspects from morphology to discourse, coreference and bridging relations in the Prague Dependency Treebank 3.0. We describe the linguistic parameters of long-distance discourse relations in Czech in connection with their anchoring connective, and suggest possible ways of their detection. Our empirical research also outlines some theoretical consequences for the underlying assumptions in discourse analysis and parsing, e.g. the risk of relying too much on different (language-specific?) part-of-speech categorizations of connectives or the different perspectives in shallow and global discourse analyses (the minimality principle vs. higher text structure).

1 Introduction

In the area of discourse coherence research, the so-called anaphoric connectives (ACs) represent a unique phenomenon, as they combine two pillars of coherence: as discourse connectives, they connect two text units – arguments expressing abstract objects [1] – and express a type of meaning between them (e.g. causality, conjunction, contrast, generalization), compare Example 1 with the connective *přesto* (*nevertheless*) and the meaning of concession.

(1) *Kapacita sálu musela být rozšířena o 150 míst, tj. na 700 sedadel.* **Přesto je zájem třikrát vyšší.**

[*The capacity of the hall had to be expanded by 150 seats, i. e. to 700 seats.* **Nevertheless, the demand is three times higher.**][1]

[1] As a typographical convention for examples of discourse relations, the left-sided argument is highlighted in italics, the right-sided argument in bold and the connective is underlined.

At the same time, the connectives act as event anaphors, taking their left-sided argument anaphorically, which also means the possibility of long-distance discourse relations. We follow the distinction in [17] of "structural" and "anaphoric" (non-structural) discourse connectives according to their syntactic relations to either both of their arguments (subordinating and coordinating conjunctions like *because, although, and, but*), or to only one of them (mainly sentence adverbs, according to the prevalent classification in English, e.g. *however, therefore, instead*).

Discourse connectives[2] are typically located within one of the two discourse arguments they connect (the internal argument), the other argument is called external[3]. Arguments of structural connectives in inter-sentential relations are determined by syntactic rules and thus they are both relatively easily retrievable. Non-structural connectives provide an anaphoric link to their antecedent, i.e. the first discourse argument in the linear order, the external argument. Most often the external argument directly precedes the sentence including the AC, but non-adjacency (a long-distance discourse relation) is also possible, compare Example 2 from the Czech corpus data.

(2) *Vedení Pojišťovny Investiční a Poštovní banky nás upozornilo, že jejich pojišťovna nebyla zařazena mezi ty, které umožňují úrazové připojištění, ač tuto službu poskytují.* Omlouváme se za toto nedopatření, dotyčná redaktorka byla pokutována. **Informaci o úrazovém připojištění v Pojišťovně IPB tedy doplňujeme.**

[*The management of the insurance company notified us that their insurance company was not listed among those that allow accident insurance, although they provide this service.* We apologize for this mistake, the editor in question was fined. **We therefore complete the information on accident insurance in the insurance company.**]

Lit.: Information on accident insurance in insurance_company IPB therefore we_complete.

The possible non-adjacency of the external argument has been a known issue in discourse analysis and parsing (e.g. [6, 11, 5]). If a discourse parser applies the default strategy (choosing the immediately preceding sentence as the external argument) with anaphoric connectives, it may lead to incorrect results.

The aim of this paper is to study properties of ACs and long-distance relations in Czech empirically in large extent on discourse-annotated data and draw possible conclusions for automatic identification of the text units (arguments) entering discourse relations. This is a crucial task, since the correct understanding of text meaning presumes the knowledge of which parts of the text actually enter the relations.

[2] In this paper, we only focus on primary discourse connectives [15].

[3] or Arg1 according to Penn Discourse Treebank 2.0 annotation of inter-sentential relations

connective	PoS	distant	all inter	distant in inter	all	distant in all
však [however]	Coord	113	1,120	10%	1,356	8%
také [also]	Adv	54	201	27%	208	26%
ale [but]	Coord	37	376	10%	1,134	3%
dále [next]	Adv	37	104	36%	110	34%
pak [then]	Adv	31	191	16%	257	12%
tedy [so]	Coord	30	239	13%	269	11%
a [and]	Coord	27	313	9%	5,128	1%
naopak [on the contrary]	Adv	27	108	25%	134	20%
rovněž [also]	Adv	26	91	29%	97	27%
proto [therefore]	Coord	22	307	7%	339	6%
ovšem [however]	Coord	21	200	11%	257	8%
i [also]	Coord/Part	17	56	30%	73	23%
navíc [moreover]	Adv	15	145	10%	169	9%
totiž [actually]	Coord/Part	13	385	3%	405	3%
zároveň [at the same time]	Adv	12	71	17%	81	15%
přitom [and/yet]	Adv	10	156	6%	162	6%
například [for example]	Adv	8	78	10%	87	9%
zase [again]	Adv	8	32	25%	38	21%
ani [neither]	Coord	8	17	47%	35	23%
přesto [yet]	Adv/Coord?	7	79	9%	89	8%

Table 1. 20 connectives with most occurrences in long-distance relations in the PDT 3.0, their *prevalent* translation, PoS, occurences in long-distance relations and their proportion in inter-sentential and in all discourse relations.

2 Language Data and Tools

The dataset used in this study, the Prague Dependency Treebank 3.0 (PDT 3.0; [2]), contains approx. 50 thousand sentences of Czech journalistic texts annotated manually on several layers of language description [4]. Annotations "beyond the sentence boundary" include discourse relations (with connectives, arguments and semantic types), pronominal and nominal coreference, bridging relations and genres of corpus documents [18]. The annotation of discourse relations was to a great extent inspired by the Penn Discourse Treebank 2.0 lexical approach (PDTB 2.0, [12]). The Prague approach [10] follows the PDTB style in marking discourse connectives as lexical anchors of local coherence relations. The connective signals the sense of the discourse relation; if it is absent, the relation is called implicit. The list of types of discourse relations in the Prague scheme is close to the list of senses used in the PDTB (especially to the PDTB 3.0 hierarchy), slightly adopted according to the Czech syntactic tradition (there is e.g. a relation of gradation). Contrary to other approaches, the annotation was carried out directly on top of deep syntax dependency trees. Whereas discourse relations according to the PDTB can be embedded and form hierarchical structures, there

is no claim about the shape of the overall structure of the text, that is why it is referred to as a framework for "shallow" discourse analysis.

For browsing, editing and searching in the data, the customizable tree editor TrEd [8] and the advanced search tool PML-Tree Query (PML-TQ; [9]) were used. The PML-TQ provides a powerful query language and as a query result offers not only individual positions in the data for a detailed inspection, but also complex statistical summaries defined by a system of output filters.

3 Anaphoric Connectives with a Non-Adjacent External Argument

Overall, out of the 18,072 discourse relations in the Prague Dependency Treebank 3.0[4] (out of which 5 455 relations are inter-sentential), 636 relations (11.7% of inter-sentential relations and 3.5% of all discourse relations) were detected where the external argument of a connective is non-adjacent to the internal argument. Detailed figures for the most frequent connectives in long-distance relations (Table 1) show that the individual proportions range up to 47% in all inter-sentential relations.[5]

3.1 Anaphoric Connectives and PoS

Surprisingly, among the 20 most frequent Czech connectives with a non-adjacent external argument, 10 are coordinating conjunctions,[6] which are structural connectives and should not accept non-adjacent external arguments.

There are several possible explanations for this behaviour. First, the issue may lie in the definition of a coordinating conjunction itself in different languages. There is a well-known tendency in the diachronic development of some adverbs, possibly in connection with demonstrative pronouns, towards sentence adverbs and gradually to conjunctions (see e.g. [18], p. 153–155).[7] In contrast to English grammar, where the strict coordinating conjunction category only contains *and, but* and *or* (e.g. [14], p. 920), the tradition of Czech PoS categorization also includes historically adverbial/pronominal expressions, the syntactical behaviour of which is nevertheless in contemporary Czech equal to those of conjunctions.

Second, for the task of Arg1 detection in [11], the sentence-initial *But*-adverbial was introduced, as also the annotations confirm long-distance relations for even the basic coordinating conjunctions. In the PDT 3.0, a very frequent coordinating conjunction *však* [*but, however*] is to our surprise more frequently

[4] All reported numbers correspond to 9/10 of the whole PDT 3.0 data (i.e. 44 thousand sentences), as the last 1/10 of the data has been designated as evaluation test data.

[5] and 36% in all relations

[6] 3 of those 10 in fact function as connectives with two different PoS labels, according to the PDT tagging.

[7] traceable by their position in the sentence – moving from right to left, writing (separate vs. as one word), loss of original meaning etc.

used as an inter-sentential (1,120) than intra-sentential (236) connective. Moreover, in absolute numbers it is the most frequent connective with non-adjacent external argument (113 tokens) in the corpus.

Third, according to [17], structural discourse connectives allow "stretching", similarly as syntactic dependencies within a sentence allow long-distance by embedding constituents. The interpretation may also be that structural connectives allow non-adjacent external arguments via (syntactic) stretching, not via anaphora resolution. Also another study of (German) ACs reports that the absence of an explicitly-anaphoric morpheme in the connective does not exclude its anaphoric behaviour [16].

As a practical application here, we suggest (and the more for experiments with non-English data) to also work with coordinating conjunctions as possible anaphoric connectives and to be critical to the outcomes of a PoS tagger. Also, detection of such inter-sententially used conjunctions might be not trivial, as, at least in Czech, they may not stand at the sentence-initial position, see Example 3.

(3) "Já to nevyhrál za svých šestnáct let závodění, *já totiž žádné peníze nikdy nedostával.* Za různé prémie a etapová vítězství jsem ovšem měl tolik aktovek a necesérů, že bych je mohl prodávat. Také nějaké ty tepláky jsem vyhrál," vzpomněl Veselý. **"Na jednu stovku si ale přece jen dobře pamatuji**.

["I did not win it in my sixteen years of racing, *I never got any money at all.* For various bonuses and stage victories, I have won so many briefcases and washbags that I could sell them. I've also won some sweatpants," Veselý remembers. **"But those hundred crowns, I still remember them well.**]

Lit.: On one hundred reflex.pron but still well I-remember.

3.2 Types of "Gaps"

For a more detailed insight, we analyzed 245 tokens of the most frequent connectives with non-adjacent discourse arguments manually (70 tokens of *však* and all tokens of *ani, dále, také, ale, přesto, proto* and *přitom*), according to their relative frequencies and across semantic classes. We concentrated on their positions with respect to paragraph boundaries, reported speech zones and we classified the nature of the "gaps", i.e. the text segments left out of the relation. Our observations are displayed in Table 2.[8] The detailed corpus analysis reveals that long-distance relations in the PDT 3.0 can be divided into two general groups of thematic patterns (or progressions): First, it is mostly a general statement/claim in the external argument, a certain type of *elaboration* in the gap, and a return or strong link to the first topic in the internal argument. Often, the elaboration in the gap zooms in to a specific detail or background information or gives an example.

The second group are *digressions* in the gaps. It is marked parentheses (in brackets, dashes), but much more often unmarked, and so difficult to detect,

[8] The figures for *však* contain all its 113 occurences.

Connective	Type(s)	PI (PI→PI)	PNI	Other
ani [neither]	conj	2 (2)	4	2
dále [next]	conj	19 (13)	16	2
také [also]	conj	15 (10)	35	4
však [however]	opp	39 (21)	55	19
ale [but]	opp	9 (3)	16	12
přesto [yet]	conc	3 (1)	2	2
proto [therefore]	reason	5 (2)	9	8
přitom [and/yet]	conj/opp	1 (1)	6	3

Table 2. Selected connectives with non-adjacent arguments: their *prevalent* semantic types, position in a paragraph-initial (PI) sentence, external arguments also in a PI sentence (PI→PI), in a paragraph-non-initial (PNI) sentence and in other settings (technical digressions, errors in annotation etc.)

comments on the topic by the writer or other person, switching between the plan of the writer and the plan of reported content (reported speech appears in the journalistic data of the PDT often without quotation marks), and also technical digressions like author names, photo captions, subheadings.

The practical difference between these two types of gapping is their referential linkage to their closest text environment. For digressions, less coreference and associative anaphora is expected, sometimes even none (see Section 3.4 below).

3.3 Local Coherence and Higher Discourse Structure

It can be supposed that arguments of connectives in paragraph-initial sentences are more likely to be distant, but also to be represented by larger blocks. For the long-distant relations in the PDT 3.0, a connective in paragraph-initial (ParInit) sentence takes another ParInit sentence as its argument in 15.1% (96/636), and 18.4% (53/288) in the subset described in Table 2. In these specific cases it can be very difficult to decide, whether they are indeed long-distance discourse relations or whether to interpret them as relations between higher discourse segments (paragraphs) that are in fact adjacent. The issue in the local coherence annotation in the PDT may be the annotation rule called *the minimality principle*: annotators were instructed to include in an argument as many clauses and/or sentences as are *minimally required* and *sufficient* for the interpretation of the relation. In the PDT, no supplementary information was annotated (compare [13], p. 14), which could potentially lead to misinterpretation of cases of paragraph coherence. It is nevertheless a problem of analytical perspective, a point where local and global discourse analyses clash and each such case should be judged individually. In the studied dataset, at least 9 relations had both relevant interpretations and there may be more.

3.4 Non-Adjacency across Semantic Classes

The distribution of the four main semantic classes (Temporal, Expansion, Comparison, Contingency) in long-distance relations in the PDT 3.0 is very uneven. There are only 42 (6.6%) Temporal and 71 (11.2%) Contingency relations, whereas the relations of Expansion and Comparison with 261 occurrences (41%) and 262 (41.2%) are much more frequent. Additive and contrastive connectives are thus much more likely to take part in these relations, but also, from the viewpoint of a global analysis (e.g. RST, [7]), these types of connectives can be expected more often in ParInit positions or even relating individual paragraphs. These findings correspond to the nature of the relations: causal, conditional or temporal relations require proximity of their arguments. This is often secured by syntax and by the use of subordinating conjunctions, and inter-sententially by adjacency. Although long-distance is also possible, these relations appear, at least in the studied data, less flexible to embedded contents. Furthermore, arguments of the additive connectives in our survey *dále* [*next, further*] and *také* [*also, too*] show specific patterns which relate to semantics of the relation: they have parallel syntactic patterns with referential identity of subjects (that might be interrupted in the gap) or they contain identical or synonymous verbs forms. *Dále* takes part in 14 cases of the type *He said – He further commented* and in 13 enumerative-like structures with sequences like *First – then – next*.

4 Conclusions

In the texts of Prague Dependency Treebank 3.0, long-distance discourse relations represent 11.7% of inter-sentential relations. In order to contribute to the automatic identification of their external arguments, we have provided a detailed linguistic analysis of connectives, arguments and semantic types in these relations and of the gaps, i.e. text segments left out of the relation. We have addressed the adverbial (anaphorical) behaviour of coordinating conjunctions, as they regularly take non-adjacent arguments (more than 290 tokens in our data). There is also no correlation in Czech between the anaphoricity of a connective and explicitly present demonstrative morpheme in its form. Further, we have classified the gaps as either *elaborations* – giving details, examples, diverting gradually from the original topic; *digressions* – outside comments, parentheses, technicalities; or, in case of both arguments located in two paragraph-inital sentences, possibly not gaps at all. The nature of the gap can be (apart from interpunction signs) traced by different coreferential enviroment and thematic progressions. Additive connectives moreover show a clear tendency to syntactic parallelism in their arguments, with referential identity of subjects, verb synonymy and high occurrence in enumerative structures. Contingency and Temporal relations (and connectives) are non-adjacent only rarely (6.6 and 11.2%). In future research, we want to focus on unmarked elaborations and comments (reported speech segments) in more detail and implement a more complex heuristics for coreference and associative anaphora in non-adjacent arguments.

Acknowledgments

This work has been supported by projects GA17-06123S and GA19-03490S of the Czech Science Foundation. The work has been using language resources and tools distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16 013/0001781).

References

1. Asher, N.: Reference to Abstract Objects in Discourse. Kluwer Academic Publishers, Norwell (1993)
2. 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 3.0. Data/software (2013), charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague. Available from ¡http://www.lindat.cz¿
3. Creswell, C., Forbes, K., Miltsakaki, E., Prasad, R., Joshi, A., Webber, B.: The discourse anaphoric properties of connectives. In: Proceedings of DAARC. vol. 4, pp. 45–50 (2002)
4. Hajič, J., Hajičová, E., Mikulová, M., Mírovský, J.: Prague Dependency Treebank, chap. 21, pp. 555–594. Springer Handbooks, Springer Verlag, Berlin, Germany (2017)
5. Kolhatkar, V., Roussel, A., Dipper, S., Zinsmeister, H.: Anaphora with non-nominal antecedents in computational linguistics: A survey. Computational Linguistics **44**(3), 547–612 (2018)
6. Lee, A., Prasad, R., Joshi, A., Dinesh, N., Webber, B.: Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax? In: Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories. pp. 79–90. Prague, Czech Republic (2006)
7. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse **8**(3), 243–281 (1988)
8. Pajas, P., Štěpánek, J.: Recent advances in a feature-rich framework for treebank annotation. In: Scott, D., Uszkoreit, H. (eds.) Proceedings of the 22nd International Conference on Computational Linguistics. pp. 673–680. The Coling 2008 Organizing Committee, Manchester (2008)
9. Pajas, P., Štěpánek, J.: System for querying syntactically annotated corpora. In: Lee, G., im Walde, S.S. (eds.) Proceedings of the ACL–IJCNLP 2009 Software Demonstrations. pp. 33–36. Association for Computational Linguistics, Suntec (2009)
10. Poláková, L., Jínová, P., Zikánová, Š., Bedřichová, Z., Mírovský, J., Rysová, M., Zdeňková, J., Pavlíková, V., Hajičová, E.: Manual for annotation of discourse relations in Prague Dependency Treebank. Tech. Rep. 47, Prague, Czech Republic (2012)
11. Prasad, R., Joshi, A., Webber, B.: Exploiting scope for shallow discourse parsing. In: Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (May 2010)

12. Prasad, R., Lee, A., Dinesh, N., Miltsakaki, E., Campion, G., Joshi, A., Webber, B.: Penn Discourse Treebank Version 2.0. Data/software (2008), university of Pennsylvania, Linguistic Data Consortium, Philadelphia. LDC2008T05
13. Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., Webber, B.L.: The Penn Discourse Treebank 2.0 Annotation Manual. Tech. rep., University of Pennsylvania, Philadelphia (2007)
14. Quirk, R., Crystal, D., Education, P.: A comprehensive grammar of the English language. Longman, London (2004)
15. Rysová, M., Rysová, K.: The centre and periphery of discourse connectives. In: Aroonmanakun, W., Boonkwan, P., Supnithi, T. (eds.) Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing. pp. 452–459. Department of Linguistics, Faculty of Arts, Chulalongkorn University, Department of Linguistics, Faculty of Arts, Chulalongkorn University, Bangkok, Thailand (2014)
16. Stede, M., Grishina, Y.: Anaphoricity in connectives: A case study on German. In: Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016). pp. 41–46 (2016)
17. Webber, B., Stone, M., Joshi, A., Knott, A.: Anaphora and discourse structure. Computational Linguistics **29**(4), 545–587 (2003)
18. Zikánová, Š., Hajičová, E., Hladká, B., Jínová, P., Mírovský, J., Nedoluzhko, A., Poláková, L., Rysová, K., Rysová, M., Václ, J.: Discourse and Coherence. From the Sentence Structure to Relations in Text. Studies in Computational and Theoretical Linguistics, ÚFAL, Praha, Czechia (2015)