Connectives with Both Arguments External: A Survey on 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. Determining a relative position of a discourse connective and the two arguments (text segments) it connects is an important part of a full discourse parsing task. This paper investigates discourse connectives whose position in a text deviates from the usual setting – namely connectives that occur in neither of the two arguments – and as such present a challenge for discourse parsers. We find syntactic patterns for this phenomenon and describe it linguistically on the basis of Czech discourse-annotated corpus material, with the aim to facilitate an automatic detection of such connectives and a correct localization of their arguments.

Keywords: Discourse connectives · Text coherence · Attribution · Syntactic structure · Text mining

1 Introduction

Any task concerned with identifying discourse relations presupposes a solid segmentation of a text into discourse units. Syntax plays a crucial helpful role in this, but it is also specific syntactic constructions that cause trouble in the segmentation. It has been argued previously that the mismatch in alignment of syntactic and discourse units is caused largely by attribution, i.e. ascription of contents to the agents who uttered them [2]. In this paper, we arrive at a similar observation from a very different perspective: we investigate the position of discourse connectives, the most apparent markers of discourse relations, with regard to the two discourse segments they connect (discourse arguments). In particular, we study connective tokens that have both arguments external, that means, the connective is not included in either of the arguments, compare the basic graphical scheme in Figure 2 and Examples 1 and 2 below.[1] The existence of such cases is documented at least in two manually annotated discourse corpora for two languages, but since searching for this type of connective-argument configuration is quite complex, there may be more in other discourse-annotated corpora.

[1] In all examples in the paper, the left-sided argument of a discourse relation is highlighted in italics, the other one in bold, the connective is underlined.

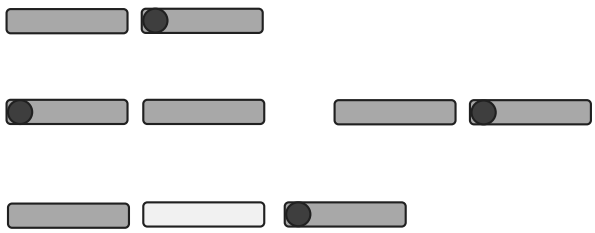

Fig. 1. Typical syntactic positions of connectives with regard to their two arguments: Line 1: coordinating conjunctions, line 2: subordinating conjunctions, line 3: adverbs

The study is based on one of the most prominent approaches to discourse coherence in the recent decade, the analysis of local coherence in the Penn Discourse Treebank (henceforth PDTB, version 2.0 [9] and lately version 3.0 [11] and [13]). This approach was also adopted for annotation of discourse relations in Czech in the Prague Dependency Treebank (henceforth PDT, [1]), the primary data source for our survey.

(1) *Compaq*, which said it discovered the bugs, *still plans to announce new 486 products on Nov. 6.* Because of the glitch, however, the company said **it doesn't know when its machine will be commercially available**. (PDTB)

Our study first discusses basic syntactic configurations of discourse connectives with regard to their PoS characteristics (1.1) and then focuses on the non-typical positions of connectives (1.2). Section 2 describes the frameworks, data a tools used and the way of detection of the targeted connectives and relations. The linguistic analysis itself follows in Section 3, with introduction of testing criteria for connective "scopes" (3.2) and corpus evidence (3.3). The implications of our findings are summed up in the concluding Section (4).

1.1 Part of Speech and the Position of Connectives

Discourse connectives, the core of which includes coordinating and subordinating conjunctions (e. g. *and, but, because*) and adverbs (e. g. *moreover, then, otherwise*), are typically located within one of the two discourse segments (arguments) they connect.[2] Coordinating conjunctions are typically placed at the beginning of the second argument in the linear order, compare the first line of Figure 1, the connective is represented by a dark dot. Subordinating conjunctions are to be found within the argument represented by a dependent clause, either preceding

[2] In this study, we do not address secondary discourse connectives like *that is why* or prepositional connectives with nominalized arguments like *after his arrival.*

the main clause (the other argument) or following it, compare line two of Figure 1. Sentence adverbs occur within the second argument in the linear order; the first argument can be non-adjacent, they are also referred to as anaphoric connectives [14], see line three in Figure 1 – the colorless box represents a (facultative) text segment that does not take part in the discourse relation in question. Claiming this, we refer to the properties of discourse connectives in Czech and also draw on our experience with English and German connectives. Intuitively, these syntactic settings would apply also for other European languages, with a possible minor variation. However, in this paper, we do not want to make claims about syntactic behaviour of connectives in other languages than Czech, this is yet a topic to be researched.

Except for intra-sentential connectives in dependent clauses, these regularities make the task of determining the arguments of a connective a search for the external (left-sided) argument [8], or Arg 1 in the terminology of the authors. The non-adjacency of the left-sided argument of anaphoric connectives has been a known issue in discourse analysis and parsing and it is being dealed with[14, 4][3].

1.2 Connectives with Both Arguments External

The aim of our study, though, are connectives with **both arguments external**, that means, connectives that are located outside both the arguments they connect, in a third clause, compare the scheme in Figure 2. Such cases are not very frequent, but they still could be documented in Czech in the annotated data of the Prague Dependency Treebank 3.0 (Example 2) and also in the Penn Discourse Treebank 2.0 for English, as Example 1 above shows.

(2) *Ministerstvo vnitra považuje před dnešním jednáním o charakteristice rozpočtu na příští rok za předčasné poskytovat detailní informace*, sdělila tisková mluvčí ministerstva. Řekla nicméně také, **že i ministerstvo vnitra počítá v porovnání s letošním rokem s posílením investic.**

Before the negotiations on the budget profile, the ministry of the interior considers it premature to provide detailed information, the ministry spokeswoman announced. She also said, however, **that even the ministry of the interior expects strengthening of investments.**

[Lit: She_said however also that...][4]

Fig. 2. A connective with both arguments external – general scheme

2 Method, Data and Tools

Our method for analysis of discourse connectives follows (i) the Penn Discourse Treebank 2.0 [9] style of annotation – a discourse connective is defined as a predicate of a binary relation between abstract objects (events, states, actions...) called discourse arguments [10], and (ii) dependency syntax of the Prague Functional Generative Description [12]. For this study, we especially use the framework for looking at inclusion/exclusion of a connective in syntactically different types of clauses. The primary source for our survey was the data of the Prague Dependency Treebank 3.0 [1] which comprises approx. 50 thousand sentences of newspaper text with manual syntactic annotation, manual annotation of discourse relations and their semantic types, connectives and arguments. The treebank provided interesting evidence but it proved too small for our study. That is why we subsequently also used the SYN V3 collection of the Czech National Corpus (CNC [3]), a reference corpus of approx. 178.5 million sentences of written contemporary Czech, automatically lemmatized and tagged, and queried it as described in Section 3.3 using the KonText query engine [5]. Evidence for the connectives with external arguments in the PDT was collected via search engine PML-Tree Query [7].[5]

[3] Some frameworks for discourse analysis, e.g. [6], though, do not allow for a non-adjacent interpretation, such a long-distance relation is non-existent.

[4] Note that in Czech, the *nicméně*–connective (*however*) is undoubtedly located within the main clause. The English translations of Czech examples are the best possible approximations to the original sentences given the more relaxed word order in Czech, even for the *ale*–connective (*but*). Where needed, we use literal translations.

[5] The best approximation to finding the relevant connectives was achieved with the following PML-TQ query:

```
t-node $t :=
 [ !descendant $n4, !sibling $n3, !parent $n4,
   member discourse
   [ t-connectors.rf t-node $n4 :=
     [ functor != "RHEM",
       0x parent t-node
         [ nodetype = "coap" ] ],
     target_node.rf t-node $n3 :=
       [ !descendant $n4, !parent $n4 ] ] ];
```

Fig. 3. Argument scheme for sentences in Example 2

3 Analysis

The distant connective position is a setting that clearly deviates from the default connective positions, compare the general scheme in Figure 2 to the previous typical settings in Figure 1. According to the corpus figures, it is a rare setting, yet the evidence texts make full sense and are not in any respect ungrammatical or stylistically incorrect. We were able to document 72 such connectives in the data of PDT which had to be manually sorted out for irrelevant material, double hits etc. with the final 48 relevant connective tokens. In majority of the detected cases, a connective placed outside both its arguments syntactically belongs to the governing clause but, from the semantic viewpoint, it is interpreted in the dependent clause. We illustrate the situation on Example 2 from above: in the main clauses plan, there is the *také*–connective (*also*) anchoring a relation of *conjunction* between the two main clauses (here: attribution clauses with verbs of saying). The contrastive meaning expressed by the *nicméně*–connective (*however/but*) can only be inferred from the meanings of reported contents (lower syntactic level), not from the attribution spans. The argument configuration is schematized in Figure 3 and the sentences can be paraphrased as follows:

– Main clauses plan:

  The spokeswoman announced A. **She <u>also</u> said B.**
  *She announced A. <u>However/but</u> she said B.

– Subordinate clauses plan (reported contents):

  The ministry does not want to reveal any information too early. **<u>However</u>, it reveals expectations on strengthening of investments.**

Over 70% of the relevant connective tokens appear in structures with verbs of saying and thinking (i.e. with governing attribution clauses) or with verbs of existence and general meanings. A closer exploration revealed that structures with attribution clauses are more likely to involve connectives interpreted lower at the level of content clauses, as just demonstrated, but also that some of the cases are difficult to judge as the phenomenon is quite complex, see Table 1. That is why we introduced a set of transformation tests that could partially sort out the different interpretations. In the rest of the paper, we focus solely on structures with verbs of saying and connectives with meanings of comparison/contrast,

as they represent the most numerous and distinctive patterns of the studied phenomenon.

	Structures with verbs of saying and thinking	Structures with verbs of general meaning, verbs of existence etc.
Connective interpreted lower	18	11
Possibly both interpretations	8	11

Table 1. Corpus evidence for connectives with both arguments external (PDT)

3.1 Wide and narrow connective "scopes"

If a connective can be interpreted in a distant clause, lower in the syntactic structure, it is similar to the transfer of negation known from sentences like 3, which is a typical form of hedging:

(3) I don't think Mary will come. = I think Mary will not come.

In the first sentence here, the negation is located in the main clause but it scopes over the dependent clause only, as the paraphrased second sentence shows. This is what we call the narrow scope. Analogous structures with narrow connective "scope" were just described on Example 2 above. These cases are not to be confused with other cases where the connective in fact syntactically and semantically relates two clauses of attribution (wide "scope", including their respective content clauses). This was demonstrated by [2] with the English example from PDTB (4) and is also visible on the Czech Example 5 from the PDT.

(4) *Advocates said the 90-cent-an-hour rise, to $4.25 an hour by April 1991, is too small for the working poor*, while **opponents argued that the increase will still hurt small business and cost many thousands of jobs.** (PDTB)

(5) *Prezident Havel při odchodu z jednání vlády uvedl, že v diskusi zvítězil názor, aby na oslavy byli přizváni i zástupci Německa.* **Premiér Václav Klaus** však **po skončení jednání LN řekl, že vláda o této věci včera ještě nerozhodla.**

When leaving the government meeting President Havel stated that the view prevailed in the debate that the representatives from Germany should also be invited to the celebrations. **Prime Minister Václav Klaus,** however**, said to the LN after the meeting that yesterday the government had not yet decided on the matter.**

3.2 Testing criteria

Determining whether the connective actually semantically relates the attribution clauses or whether it operates at a lower level between the attributed contents can be a problematic issue. We introduce two transformation tests (that can be used independently and should provide the same results) based on substitution and eliding for determining the connective scope in attribution clauses:

It is (i) **moving the connective lower**, to the content clause and (ii) **omitting the attribution clause** in the second sentence in the surface order while preserving the connective in question. After the application of any of the two tests, if the connection of content clauses related with an (originally distant) connective preserves the original meaning, we can speak about the narrow connective scope. On the other hand, if such a connection gets semantically weird or incomplete, the connective will relate the governing, attribution clauses and take the wide scope. Compare the situations under (i1-ii1) and (i2-ii2), where the tests are applied for the original corpus examples 2 and 5 from above:

(i1) Lowering of the connective: the connective operates at the level of attributed contents

(6) *Ministerstvo vnitra považuje před dnešním jednáním o charakteristice rozpočtu na příští rok za předčasné poskytovat detailní informace*, sdělila tisková mluvčí ministerstva. Řekla také, **že nicméně i ministerstvo vnitra počítá v porovnání s letošním rokem s posílením investic.**

Before the negotiations on the budget profile, the ministry considers it premature to provide detailed information, the ministry spokeswoman announced. She also said that, however, **even the ministry expects strengthening of investments.**

[Lit: She_said also, that however even the ministry...]

(ii1) Omitting the attribution clause: the connective operates at the level of attributed contents

(7) *Ministerstvo vnitra považuje před dnešním jednáním o charakteristice rozpočtu na příští rok za předčasné poskytovat detailní informace*, sdělila tisková mluvčí ministerstva. **Nicméně i ministerstvo vnitra počítá v porovnání s letošním rokem s posílením investic.**

Before the negotiations on the budget profile, the ministry considers it premature to provide detailed information, the ministry spokeswoman announced. **However, even the ministry expects strengthening of investments.**

[Lit: However even the_ministry...]

(i2) Lowering of the connective: awkward. The connective operates most likely at the level of main clauses.

(8) *?Prezident Havel při odchodu z jednání vlády uvedl, že v diskusi zvítězil názor, aby na oslavy byli přizváni i zástupci Německa.* **Premiér Václav Klaus po skončení jednání LN řekl, že vláda však o této věci včera ještě nerozhodla.**

?When leaving the government meeting President Havel stated that the view prevailed in the debate that the representatives from Germany should also be invited to the celebrations. **Prime Minister Václav Klaus said to the LN after the meeting that yesterday the government, however, had not yet decided on the matter.**

The placement of the connective *však* (*however*) in Example 8 appears disruptive in the dependent clause. This is most likely because we expected it to come earlier in the sentence. The contrast here apparently relates to the person of Prime Minister claiming something else than the President, the contrastive connective tends to stand in close proximity of both these agents.

(ii2) Omitting the attribution clause: the connective operates at the level of attributed contents, but with a loss of important information. This transformation is thus not possible.

(9) *Prezident Havel při odchodu z jednání vlády uvedl, že *v diskusi zvítězil názor, aby na oslavy byli přizváni i zástupci Německa.* **Vláda však o této věci včera ještě nerozhodla.**

*When leaving the government meeting President Havel stated that *the view prevailed in the debate that the representatives from Germany should also be invited to the celebrations.* **However, the government had not yet decided on the matter yesterday.**

Example 9 makes perfect sense even after the transformation but the meaning has shifted: the omitted information that the second argument is a (contradictory) statement of the Prime Minister, not the President, is an important one. In this case, the connective *však* (*however*) operates between the attribution clauses, which means, the syntactic and the discourse interpretations go hand in hand.

As far as we could observe, a contrastive connective syntactically connecting two main attribution clauses takes the wide scope, that means the syntactic and the discourse interpretations match, in structures with non-identical sources of attribution. In other words, the connective must take a wide scope, if the two contents in dependent clauses are expressed by two different agents (as in 4 and 5 above).

If we accept the fact that a (contrastive) connective scopes over a clause dependent to the one in which it appears, it can be treated as "raised". But the connective does not always "climb" in the syntactic structure like the negation in

Example 3 above. If we switched the order of the clauses, so that the attribution clause is sentence-last, the connective would be most likely located in the content clause (in this case the first clause in linear order). Disregarding the ordering of the attribution clause and the content clause, a contrastive connective **tends to stand as far to the left as possible**, i.e. as close as possible to its first, left-sided argument. If we now consider the previous setting – the attribution clause containing the connective at the sentence-initial position – we can say that the connective has moved left.[6]

3.3 Corpus Evidence

To confirm or disprove these assumptions about preferential positions of connectives in various settings with attribution, we used two different corpora of Czech. The results from the PDT are displayed above in Table 1. In the much larger CNC, we searched for the two following patterns and their variation. In the first pattern, the contrastive connective appears in the main clause, whereas in the second one, it appears in the dependent (content) clause introduced in Czech obligatorily by *že (that)*:

1. [The beginning of the sentence – verb[7] – (0-3 positions) – connective *však/ale* – comma – subordinating conjunction *že*]

Example (lit.): *He_said but/however*(...)*, that*

2. [The beginning of the sentence – verb – (0-3 positions) – comma – subordinating conjunction *že* – connective *však/ale*]:

Example (lit.): *He_said* (...)*, that but/however*

The results are summarized in Table 2.

Structures with *ale/však*–connectives in the governing clause are much more frequent in Czech. We have divided the results to structures with one to three arbitrary positions between the verb and the connective/comma, which allows for inclusion of most multi-word verb forms: first 6 rows of Table 2; and to fixed single-verb structures: rows 7 – 10. Rows 5 and 6 represent structures where both in the main clause and in the dependent clause one to three arbitrary positions are allowed – we added these two rows to the table in order to reflect the possible positions of Czech clitics (= in front of the connective). For both patterns, it is clearly visible that structures with connectives in the governing clause are more frequent than structures with lower placement of the connective. This supports the claim that (contrastive) connectives tend to be located close to their first argument in the linear order. A secondary test is represented by the fact that

[6] In an earlier phase of this research, we called the phenomenon *connective movement*. A colleague later pointed out the possible confusing connection of this term with generative grammar in sense of N. Chomsky, which we do not want to make here, so we abandoned the use of this term.

[7] As Czech is a pro-drop language, we did not need to search for a pronominal subject in this case.

Pattern No.	Pattern variation	Total occurrences	Instances per million tokens
1	Verb (...) but, that	41,842	15.58
1	Verb (...) however, that	50,162	18.68
2	Verb (...), that but	606	0.23
2	Verb (...), that however	294	0.11
2	Verb (...), that (...) but	3,594	1.34
2	Verb (...), that (...) however	1,570	0.58
1	Verb but, that	24,477	9.12
1	Verb however, that	31,822	11.85
2	Verb, that but	415	0.15
2	Verb, that however	235	0.10

Table 2. Connectives in main clauses and dependent "*that*" clauses (CNC)

although the semantics of the verb in the query was not restricted in any way, the query returned almost exclusively verbs of saying.

4 Conclusion

In the annotation of discourse relations and connectives in Czech written texts, externally placed connectives were found (connectives located outside both their arguments). These cases occur in vast majority in connection with attribution (reporting). According to our findings, this pattern usually involves a contrastive connective located in an attribution clause, which is at the same time the governing clause. The connective expresses contrast between two reported contents, but only so, if a single speaker utters them both; involvement of more speakers suggests a "wide connective scope", in other words, a regular interpretation (contrast between the main clauses).

In these structures, a connective expressing contrast between two reported contents **moves left** to the sentence-initial attribution clause in order to stand closer to its left (first in the linear order) argument. In opposite cases, i.e. where the attribution clause is sentence-last, the connective is unlikely to appear that far right. Testing criteria for Czech were partly helpful, although not always univocally decisive. With the meaning of addition (conjunction), it is difficult to decide about the scope of the connective, for other meaning classes, there was no sufficient material found. We provided corpus evidence from a small manually annotated treebank with syntactic and discourse information (PDT): the phenomenon is rare but traceable; results from a large reference corpus of Czech (CNC) show that contrastive connectives *ale/však* are much more frequent in attribution (main) clauses than in the dependent content clauses.

Connectives with both arguments external and discourse relations expressed by them present a challenge for discourse parsers. In our study, we attempted

to define conditions under which connective movement occurs. Our results and testing criteria may help devise rules for recognition of such cases and may contribute to improving the specification of argument extent and the precision of the automatic discourse parsing.

Acknowledgments

This work has been supported by projects GA17-06123S and GA17-03461S 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. 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
2. Dinesh, N., Lee, A., Miltsakaki, E., Prasad, R., Joshi, A., Webber, B.: Attribution and the (non-) alignment of syntactic and discourse arguments of connectives. In: Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky. pp. 29–36. Association for Computational Linguistics (2005)
3. Křen, M., Čermák, F., Hlaváčová, J., Hnátková, M., Jelínek, T., Kocek, J., Kopřivová, M., Novotná, R., Petkevič, V., Procházka, P., Schmiedtová, V., Skoumalová, H., Šulc, M.: Czech National Corpus – SYN, version 3. Data/software (2014), Institute of the Czech National Corpus, Charles University, Faculty of Arts, Prague. Available from http://www.korpus.cz
4. 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)
5. Machálek, T., Křen, M.: Query interface for diverse corpus types. Natural Language Processing, Corpus Linguistics, E-learning pp. 166–173 (2013)
6. 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**, 243–281 (1988)
7. 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)
8. 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)

9. 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
10. 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)
11. Prasad, R., Webber, B., Lee, A., Joshi, A.: The Penn Discourse Treebank 3.0 (in prep)
12. Sgall, P., Nebeský, L., Goralčíková, A., Hajičová, E.: A Functional Approach to Syntax in Generative Description of Language. American Elsevier Pub. Co., New York (1969)
13. Webber, B., Prasad, R., Lee, A., Joshi, A.: The Penn Discourse Treebank 3.0 Annotation Manual. Technical Report (2018)
14. Webber, B., Stone, M., Joshi, A., Knott, A.: Anaphora and discourse structure. Computational Linguistics **29**(4), 545–587 (2003)