Annotation of Discourse Connectives for the Prague Dependency Treebank

Lucie Mladová
Charles University Prague, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. The paper presents a preliminary study on discourse connectives (DC) in Czech. Aiming to build a computerized language corpus capturing discourse relations in Czech, we base our observations on current foreign projects with the same purpose. In this study, first, the different methods of linguistic analysis of the discourse structure and discourse connectives are described, next, the nature and properties of the group of DCs are analyzed and, finally, the procedure of the annotation of discourse connectives in Prague is presented.

1. Introduction

It is widely accepted among discourse researchers that discourse connectives, i.e. language expressions that connect discourse units such as clauses or sentences, for instance *and, but, also, however, therefore, on the other hand* etc., function as a primary (because the most apparent) source for identifying and describing syntactico-semantic structure of a discourse, both for humans and machines. A text which lacks connective words could be treated as less coherent than one that is rich on them. Also, the researchers agree, the richer the annotated linguistic information available above certain discourse data, the better the discourse analysis. As the Prague Dependency Treebank 2.0 (PDT, *Hajič et al.*, 2006) offers such information for the Czech data already (PoS tagging, shallow and deep syntactic analysis, information structure and coreference relations) at the state of the art, a thorough linguistic description of the properties of discourse connectives is needed for extending the relevance of this corpus for automatic discourse analysis and modeling.

2. Previous research on discourse structure and DCs

Since the first discourse-annotated data collections appeared (e.g. *Carlson and Marcu*, 2001), there have been a number of approaches developed on how to arrange such a project suitably for the different purposes of NLP research. It is very important for every subsequent discourse project to learn from the results and difficulties of the previous projects and their frameworks. Therefore, in this section, we give a brief selective overview of the recent approaches that concentrate on analyzing/modeling discourse and we take a closer look at those analyzing discourse via discourse connectives.

The discourse projects differ, apart from the various sets of meanings assigned to relations in discourse, in the basic idea, whether the structure of a text (text or discourse understood as a coherent string of utterances, written or spoken) is representable with a tree structure or not. So, the RST-Treebank (*Carlson and Marcu*, 2001), manually annotated according to the Rhetorical Structure Theory of Mann and Thompson (1988), builds a tree structure for every document. Other approaches such as Discourse Graphbank (*Wolf et al.,* 2005) claim that a tree representation for the whole structure of a text is descriptively not adequate, and thus it is necessary to build net-like graphs.

Other projects, like Penn Discourse Treebank 2.0 (PDTB, *Prasad et al.*, 2008a), focus on description of lexical markers of discourse relations and their scopes: a discourse connective is considered to be a predicate of a binary relation; it takes two text units (mainly clauses or sentences) as its arguments (see Figure 1). The annotation proceeded in three steps: 1, a connective is found at first, 2, its two arguments, (i.e. their extent) set, and 3, to each relation represented by a connective a discourse-semantic label assigned (PDTB Annotation Manual, 2007).

(1) *[John eats porridge for breakfast], <u>while</u> [Mary eats muesli].*

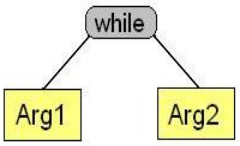

Figure 1. Discourse connective and its argument in the PDTB

For the purpose of Czech annotations this procedure of annotation was adopted from the Penn research team and the guidelines adjusted for Czech in correspondence with the Prague functional generative description (*Sgall, Hajičová and Panevová*, 1986). As there are projects being developed also for other languages based on the Penn method, e.g. Hindi (*Prasad et al.*, 2008b) or Turkish (*Zeyrek and Webber*, 2008), some cross-linguistic research on discourse structure, semantics and on language-universal properties of the connectives will be possible.

The most recent empirical studies on discourse coherence proved, however, that building tree-like structures in the nature of RST alone, or identifying and annotating connectives and their scopes alone does not perform well enough when modeling discourse structure (compare *Stede* 2004). The conclusion is that the richer the linguistic information added, (i.e. the annotations of the same documents including various types of linguistic information – morphology, syntax, coreference, topic-focus articulation etc. – on different annotation levels) the better the performance of automatic procedures. Hence, a multilayer annotation of various linguistic phenomena should be the goal of future discourse projects. However, at the same time, a balanced setting of attributes and their values is required, to avoid sparse appearance of certain values leading to lower relevance of the training data for machine learning.

3. What is a discourse connective?

The group of discourse connectives is determined functionally. There are several, generally shared and accepted criteria for delimiting the group, none of them, however, is really strict.

First, as in the Penn approach, the most important property of DCs is that they take two (or, in some cases more, see Section 3.1) text units as their arguments. They connect these units syntactically to larger ones while signaling a semantic relation between them at the same time (similarly conjunctions within a compound sentence).[1]

Secondly, DCs are morphologically inflexible and they never act as a part of the syntactic structure of a sentence. Like modality markers, they are "above" or "outside" of the proposition. They are represented by coordinating conjunctions, some subordinating conjunctions[2], some particles and sentence adverbials, and marginally also by some other parts-of-speech – mainly in case of fixed compound connectives like *in other words* or *on the contrary*.

Discourse connectives can connect texts units of different discourse levels. The units can be clauses, sentences, and even larger text spans such as paragraphs[3]. So, there is a hierarchy in the discourse structure, a "lower" relation can be embedded in an "upper" one as its argument. Also, going "up" in such a hierarchy, we often move from syntactically motivated relations (typically dependency relations and different types of coordination of adjacent clauses or sentences) to the so called rhetorical relations (*Asher and Lascarides*, 2003) – the organization or composition of the text itself. Thus, a paragraph-initial connective usually marks a hierarchically higher discourse relation (namely

[1] This ability to relate to two text units does not apply to the whole wider group of discourse markers. Discourse markers such as, for instance, time and space anchors of a discourse do not take two text spans as their arguments.

[2] In many approaches some of the dependent clauses, and so the subordinating expressions related to them, are not treated as discourse level arguments, typically dependent clauses representing some of the valency members of the verb or relative clauses.

[3] More on the nature of discourse arguments in the Section 3.2

the relation of the actual paragraph to the previous one (or more)) than a connective within a single sentence. Drawing an exact boundary between the "syntactic" and "rhetorical" nature of the discourse relations indicated by the connectives is difficult; we can only follow the clue that some of the relations are normally not present at the sentence syntactic level: it concerns all kinds of restatements, like correction (*in other words*, *or*, *also* etc.), generalization (*overall, ultimately, indeed, in short* etc.) and specification (*in fact, specifically, in particular*), instantiation (*for example, for instance, in particular*) and similar (compare *Mladová*, 2008a).

As we said, discourse connectives indicate a semantic relation between the units they connect and they are the most visible markers of such a relation. Although in many cases ambiguous when on their own, when specified through a given context discourse connectives "label" the discourse relation quite unambiguously (as shown in examples below). However, when left out, it can be difficult, even with a larger context, to capture the type of the semantic relation between the text units. The PDTB project calls such non-present connectives implicit connectives and in the PDTB annotation scheme, annotators tried to insert the most appropriate relation between every two adjacent sentences (PDTB Annotation Manual, 2007).

For Czech, a survey on semantics of some ambiguous connectives has been carried out. As we found out, a single connective can represent various semantic relations depending on its actual context and its connectability to multiword connective expressions, see examples (2) – (7):

(2) *Pršelo, ale deštník si nevzal.* Concession
(It rained but he didn't take an umbrella.)[4]

(3) *Nespal, ale vymýšlel plán na zítřek.* Opposition
(He didn't sleep but he created the plan for the next day.)

(4) *Nesportuji, ale na plovárnu si občas zajdu.* Exception
(I don't do any sports but I go swimming time to time.)

(5) *Dal si nejen hlavní jídlo, ale objednal si i zákusek.* Gradation[5]
(He not only had the main dish, but he also ordered a dessert.)

(6) *To je ale krásně!* NOT a discourse connective
(What [but] a nice weather!)

(7) *Byl teplý, ale zamračený den.* NOT a discourse connective
(It was a warm but cloudy day.)

3.1. Binary nature of discourse connectives

Many discourse approaches like the Penn Discourse Treebank treat discourse connective relations as exclusively binary (PDTB Annotation Manual, 2007). That means, every connective is a predicate of a binary relation and thus takes exactly two arguments. Although this perspective is certainly more convenient for formal description and automatic analysis, we argue that there are at least two types of connective relations where more arguments than two can be seen (*Mladová*, 2008a). First, it concerns multiple coordinations of the following type (see Figure 2):

(8) *Buď si hráči prosadí svou, nebo zvítězí správní rada, nebo dojde k nějakému kompromisu*.

(Either the players will enforce their view, or the management board will win, or there will be a compromise.)

[4] Some of the English translations of the Czech examples will sound strange but we need to preserve the *but*-connective to illustrate the homonymy of this connective in Czech.

[5] compound connective *nejen – ale i*

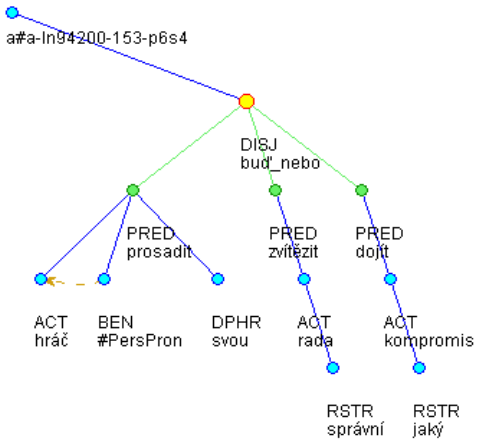

Figure 2. An example of a non-binary discourse relation

The second type is the list relation, where the individual segments start with a number, letter or expressions such as *first, second(ly), next, last* etc. In this type of discourse relation every item of the list is related to the preceding item and to the introductory statement for the whole list. Thus, a connective *secondly* relates the clause where it appeared both to the preceding clause and to the introduction of the listing, if any introduction is present, compare the Czech examples (9) – (11).

(9) *K tomu, aby zaměstnavatel pracovníkovi za škodu opravdu odpovídal, musí být splněny tyto podmínky:*

(10) *1. Zaměstnanci musí vzniknout škoda, tj. musí dojít k určitému snížení hodnoty jeho majetku (v některých případech mu vzniká i právo na náhradu ušlého zisku).*

(11) *2. Zaměstnavatel nebo jiná fyzická či právnická osoba, která jedná jeho jménem, musí porušit své právní povinnosti.*

(The employer is factually responsible for the damage towards the employee when the following conditions are fulfilled:
1. There must be a damage caused to the employee, i.e. there must be some reduction in value of employee's property (in some cases, there arises the right to compensation for loss of employee's profits).
2. The employer or other physical or legal person acting on his behalf must violate their legal obligations.)

3.2. Limits of the group delimitation

The group of discourse connectives, as described so far, still lacks one important, rather practical restriction. It is impossible to define what a discourse connective is without defining precisely what a discourse unit (or argument) is. Generally it is considered that discourse arguments are propositions, *abstract objects* (*Asher*, 1993), i.e. states and events, expressed mainly by a finite verb predicate structure, by a clause. But what are abstract objects? In a strict view, there are many non-clausal arguments such as nominalizations, answers to questions, deictic expressions etc. If we do not mind the formal difference between *protože odešel* – "because he left" and *kvůli jeho odchodu* – "because of his leaving", also some prepositions (here *kvůli* – "because of") could become discourse connectives. In fact, all lexical expressions with certain discourse-connecting meaning could be extracted and annotated for their role in making a discourse coherent, e. g. verb constructions like *it follows from, that implies, to summarize, to conclude,* or non-verbal phrases with similar meaning (*the reason for it, in consequence of, in any case, under these conditions, in the first instance*) and so on. Hence, it is mainly for practical purposes (for keeping both annotations and computational experiments at some level of integrity) that the notion of a discourse connective is restricted to a more or less fixed list of atomic expressions and the notion of its arguments to clausal arguments.

There is one more difficult aspect of identifying the group of discourse connectives for the purposes of the annotation: Some expressions with other primary function also take two text units as their arguments. For instance, some focus particles or rhematizers (*Sgall, Hajičová and Buráňová*, 1980) are functionally homonymous with discourse connectives. See the example sentences (12) and (13).

(12) *Peter wrote a birthday card for his mother. Besides, he bought a bunch of roses for her.*

(13) *Peter wrote a birthday card for his mother. He also bought a bunch of roses for her.*

We argue that in (13), the particle *also* combines two functions in this particular context – it is highlighting the focus part of the sentence *bought a bunch of roses for her,* and it functions as a discourse connective, similarly to the word *besides* in (12). Both words, having their place in the second sentence, imply the existence of a preceding sentence, saying "there has been some other action of Peter". For Czech, this polyfunctionality mainly concerns the particles *také, též, i, rovněž, zároveň, spíše, nejspíš, zase, jen, naopak* in specific contexts (compare *Mladová*, 2008b).

3.3. Pragmatic use of the connectives

In the PDTB, pragmatic sense tags are assigned to connectives in those cases when the actual meaning deviates from the semantics of the connective (PDTB Annotation Manual, 2007). A pragmatic condition then appears with the conditional subordinating conjunction *if* in the following example, with the first clause not being the real (semantic) condition of the second clause:

(14) *If you are thirsty, there is beer in the fridge.*

The potential thirst of the addressee of this utterance is not causally bound with the existence of the beer in the fridge. The truth value of the implication does not hold. Similarly, there is a pragmatic rather then semantic relation between the following two clauses:

(15) *John is home because the lights are on in the house.*

Here, the second clause is only a justification for the claim in the first clause, it is not a cause. In both examples and other such cases, there is some kind of ellipsis – but we all understand well: "*I assume* that John is home *because I see* that the lights in the house are on". The real causal relation exists between the speaker's assumption and his seeing. In the PDTB, four pragmatic meanings are distinguished and annotated: pragmatic cause, condition, contrast and concession. We find it important not to overlook this property of discourse connectives and to distinguish the semantic and the pragmatic aspects of the text structure in the annotation.

4. Annotation of Discourse Relations in Prague

As we mentioned above, in its present shape, the multilayer annotation of the Prague Dependency Treebank 2.0 already marks some of the phenomena relevant for discourse analysis and modeling. In our discourse annotation project, we take advantage of these. The future "discourse layer" of annotation[6] adopts a part of the underlying syntactic annotation – namely some of the dependency relations, the coordinating relations between clauses (not those between lower units) and expressions marked with the semantic label PREC (reference to PREceding Context) – and further the coreference annotation. Then, there are also annotations of the extended coreference (not only pronominal

[6] Discourse layer of annotation is not treated as the next, higher level of language system description according to the relation of forms and functions in the functional generative description (FGD); it is rather a side-step into the communicative aspect of the language. *Discourse* in FGD means usage of the language as a system in the process of communication (compare e.g. Mladová 2008a).

anaphora) and bridging relations[7] in progress and, last but not least, the annotation of discourse connectives and their arguments.

As for the annotation of the connectives, the observations described above led to following decisions: Currently, we annotate only explicit connectives, their scopes and semantics, the annotation of the implicits has been postponed. We annotate only relations of clausal arguments, i.e. neither nominalizations nor deictic expressions. We focus on the annotation of the PREC-labeled expressions but the texts are also being checked for other possible DCs. A survey is being carried out on pragmatic use of the Czech connectives and the Prague set of discourse sense tags is being verified on the first data. After evaluating current testing annotations, the full manual annotations will be initiated later this year.

5. Conclusion

We reported on a corpus project concerning the annotation of the Czech discourse connectives. We introduced the properties of the group of DCs and we argued that for research purposes, their annotation should be accompanied by other discourse-relevant annotations. In the Prague Dependency Treebank, within its deep syntactic and future discourse layers, various means of a natural language are captured that play a role in connecting discourse units into a coherent, meaningful sequence of information. Such a large-scale database (the whole PDT contains approx. 50 000 Czech sentences) will be of use for both linguistic research grounded on language corpora – after all, it is the first "discourse corpus" for Czech! – and for NLP applications such as automatic text summarization, information retrieval, automatic annotation of further data, or, following up the Penn approach and comparing their annotations of English data with the Czech data, it can build a parallel resource for sophisticated machine translation systems.

Acknowledgments. The present work was supported by the grants GAUK 103609 of the Grant Agency of the Charles University Prague, and GA201/09/H057 and GA405/09/0729 of the Grant Agency of the Czech Republic.

References

Asher, N., *Reference to Abstract Objects in Discourse,* Kluwer Academic Publishers, Dordrecht, 1993.

Asher N., Lascarides A., *Logics of Conversation*, Cambridge University Press, 2003.

Carlson, L., Marcu, D., Discourse Tagging Reference Manual, Ms., University of Southern California / Information Sciences Institute, 2001.

Hajič, J. et al., *Prague Dependency Treebank 2.0*, Linguistic Data Consortium, Philadelphia, 2006.

Mann, W., Thompson, S., Rhetorical Structure Theory. A Theory of Text Organisation, In *TEXT 8* (3): 243-281, 1988.

Mladová, L., *Od hloubkové struktury věty k diskurzním vztahům (Diskurzní vztahy v češtině a jejich zachycení v anotovaném korpusu)*, master's thesis, Charles University Prague, Czech Republic, 2008a

Mladová, L., K problematice vztahu rematizátorů a textových konektorů, in: *Čeština doma a ve světě*, vol. 16, 2008b.

Nedoluzhko, A., *Zpráva k anotování rozšířené textové koreference a bridging vztahů v Pražském závislostním korpusu. [Report about the Annotation of the Extended Text-Coreference and Bridging Relations in the Prague Dependency Treebank.],* technical report, Institute of Formal and Applied Linguistics, Charles University, Prague, 2007.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. and Webber, B., The Penn Discourse Treebank 2.0, in: *Proceedings of the 6th International Conference on Language Resources and Evaluation,* 2008a.

Prasad R., Husain, S., Sharma, D. M., Joshi, A., Towards an Annotated Corpus of Discourse Relations in Hindi, in: *Proceedings of the IJCNLP-08 Workshop on Asian Language Resources*, Hyderabad, India, 2008b.

The Penn Discourse Treebank 2.0 Annotation Manual, 2007, http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf

Sgall, P., Hajičová, E., Buráňová, E., *Aktuální členění věty v češtině,* Praha, Academia, 1980.

[7] a project under the leadership of Anja Nedoluzhko, UFAL MFF

Sgall, P., Hajičová, E., Panevová, J., *The Meaning of the Sentence and its Semantic and Pragmatic Aspects*, Dordrecht, Reidel Publishing Company, Praha, Academia, 1986.

Stede, M., The Potsdam Commentary Corpus, in: *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, Barcelona, 2004.

Wolf, F. et al., *Discourse Graphbank,* Linguistic Data Consortium, Philadelphia, 2005.

Zeyrek, D., Webber, B., A Discourse Resourse for Turkish: Annotating Discourse Connectives in the METU Corpus, in: *Proceedings of the IJCNLP-08,* Hyderabad, India*,* 2008.