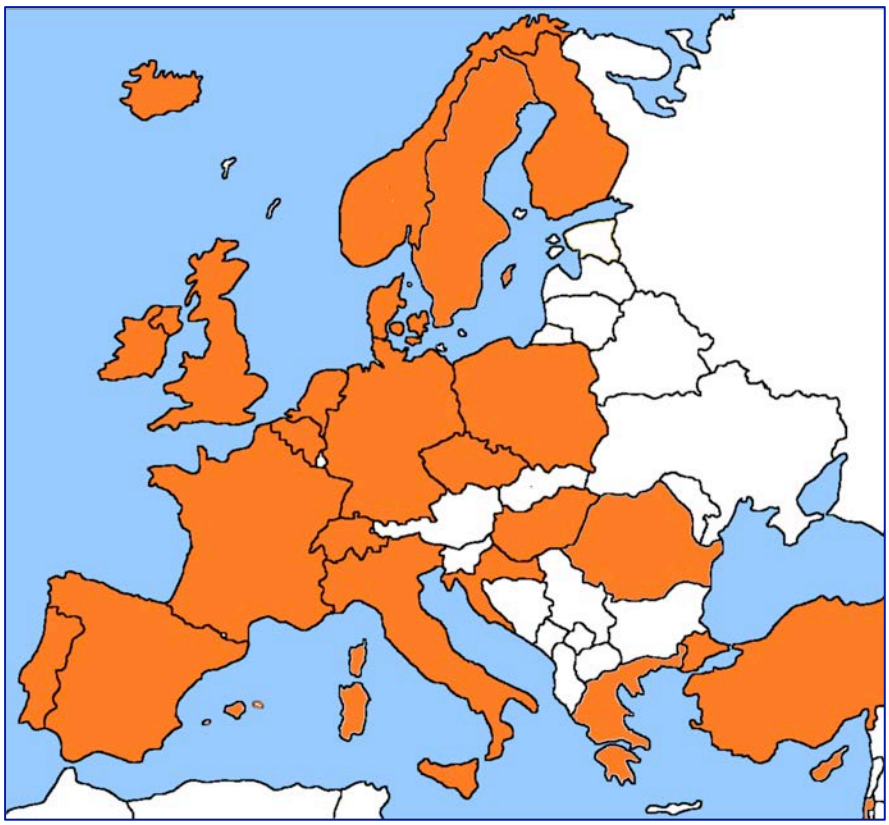
TextLink

Structuring Discourse in Multilingual Europe



First Action Conference

25 – 28 January 2015

Université Catholique de Louvain,

Louvain-la-Neuve, Belgium

Organiser: Liesbeth Degand

cost

Welcome to the first TextLink Conference

At present the TextLink COST Action network counts 25 participating countries with more than 100 members from over 50 academic institutions. The most important objective of the First TextLink Action meeting is to make the Action as "active" as possible! Therefore the focus will first and foremost be on improving and enhancing networking among the members of the many different research teams involved.

The aim of the TextLink Action is to make theoretical and methodological progress in the fields of discourse and corpus annotation, more particularly of discourse structuring devices in no less than 20 different languages. All these languages vary in how discourse relations and structure are signaled, but they also have a number of principles in common. This appears from the many discourse-annotated corpora that are becoming available in individual languages (cf. outcome of the WG1 Meeting in Prague; http://textlinkintranet.wix.com/intranet#!wg1-meeting---prague-october-2014/cr82). But we have still some way to go to interconnect these resources by contrasting, comparing, putting together, discussing, arguing, agreeing and disagreeing. This is the program of the three days in Louvain-la-Neuve! It is our hope that by learning what is already available in some of the languages, what is sharable and what needs urgent development, we will make progress in our collaborative effort.

Oral presentations and poster presentations distributed over general sessions, as well as working group sessions, will give ample opportunity to exchange and learn from one another. We are happy to count on the expertise of our two keynote speakers at this conference, Maite Taboada (Simon Fraser University) and Andy Kehler (UC San Diego). We also wish to warmly thank the working group leaders for their active involvement in preparing this scientific event.

We hope that you will enjoy the talks and the discussions, the questions and the answers and that you will find lots of food for thought and further collaborations.

Liesbeth Degand

Nicky Thrupp

CONTENT

TextLink

Structuring Discourse in Multilingual Europe

FIRST ACTION CONFERENCE
AGENDA

Venue: BATIMENT SOCRATE (lunches and coffeee breaks at COLLEGE ERASME)

Monday 26 January 2015

9.30-10.30	*Coffee and Registration*
10.30-11.00	**Welcome** (Liesbeth Degand)
	Keynote Speakers
11.00-12.00	**Conversational Eliciture** (Andrew Kehler, UC San Diego)
12.00-13.00	**Rhetorical relations are relations of coherence: What discourse coherence means, and how we can find it** (Maite Taboada, Simon Fraser University)
13.00-14.30	*Lunch*
	Collaborative Research Presentations
14.30-15.00	**Crowd-sourcing concurrent relations** (Anna Dickinson, Hannah Rohde, Annie Louis, Chris Clark and Bonnie Webber)
15.00-15.30	**"On the one hand" as a Cue to in the Comprehension of Discourse Structure** (Vera Demberg, Hannah Rohde, Merel Scholman, Chris Cummins, Emily Nicolet)
15.30-16.00	**Detecting Simplex Subordinators in Turkish** (Faruk Acar, Deniz Zeyrek, Ruket Çakici)
16.00 – 16.30	*Coffee*
16.30-17.00	**PDiT-GECCo Scientific mission: towards comprehensive approaches to discourse-structuring devices** (Anna Nedoluzhko, Ekaterina Lapshinova-Koltunski, Kerstin Kunz)
17.00-17.30	**Assessing the validity of annotation guidelines: towards multimodal equivalence of DSDs** (Ludivine Crible, Sandrine Zufferey)
17.30-18.00	**WG1 Report on Prague Meeting** (Jiri Mirovsky, Manfred Stede)
19.00 -	*Conference dinner (Fleur de sel)*

Tuesday 27 January 2015

Joint Poster Sessions	
9.00-9.30	Putting up posters (WG1 & WG2)
9.30 - 10.45	**Poster session WG1&2** (coffee available during session)
10.45 - 11.15	Putting up posters (WG3 & WG4)
11.15 - 12.30	**Poster session WG3&4** (coffee available during session)
12.30-13.30	*Lunch*

WG parallel sessions		
13.30 - 14.00	**WG1: Perspectives for future work**	**WG4: Perspectives for future work**

WG2 & 3 joint session	
14.00 - 17.00	**Annotating and learning full discourse structures for texts and dialogues** (Nicholas Asher)
	Reliable annotation in RST: Segmentation, nuclearity, relations and signalling (Maite Taboada)
	PDTB-style Annotation of Discourse Relations: Principles, Benefits, and New Directions (Bonnie Webber)
	A Neo-Humean Taxonomy of Coherence Relations (Andrew Kehler)
	Cognitive plausibility and a systematic set of relations – Useful for discourse annotation? (Ted Sanders)
	The ISO Semantic Annotation Framework for Discourse Relations (Harry Bunt)

WG parallel sessions		
17.00 - 18.00	**WG2: Perspectives for future work**	**WG3: Perspectives for future work**
18.00 -	*Free Evening*	

Wednesday 28 January 2015

WG Feedback Session	
8.30-10.00	**WG feedback session**
10.00-10.30	*Coffee*

MC Meeting	
10.30 - 13.00	**MC Meeting***
12.00 - 14.00	*Lunch*

**non-MC members are welcome to attend (without voting rights)*

Rhetorical relations are relations of coherence: What discourse coherence means, and how we can find it

Maite Taboada

Simon Fraser University

In this talk, I will do two things. First, I will discuss hierarchies of rhetorical or coherence relations, and the issue of consensus on a common taxonomy. Second, I will talk about signals for coherence relations, and our corpus annotation of a broad set of signals.

First, in terms of hierarchies, and in general considering the nature of rhetorical relations, I propose a top-down examination of rhetorical relations, that is, one that views relations between propositions in discourse as relations that help create coherence. I will review different approaches to rhetorical, coherence and conjunctive relations, and explain where Rhetorical Structure Theory (Mann and Thompson 1988) fits in with other proposals. Coherence is part of texture, and thus related to entity-based coherence or cohesion (Halliday and Hasan 1976) and to general properties of discourse. I will argue that there is a cline of grammaticalization of rhetorical relations, from discourse to syntax, and that differences across theories are sometimes rooted in where in that cline the theory positions itself. For instance, RST is at the end of the cline closer to discourse, and does not make strong claims about the syntactic realization of rhetorical relations. The conjunctive relations of Halliday and Hasan (1976) and Martin (1992), on the other hand, are more clearly syntactic, and have lexical elements as signals of the relation. My optimistic view is that we can probably map relations in different theories if we bear in mind that they may be more or less abstract versions of each other.

In the second part of the talk, I will discuss signalling. In this sense of rhetorical relations as relations of coherence, the relations are present whether signalled by a particular device or not. This is the long-held view within Rhetorical Structure Theory. The concern in RST has been to explain how coherence, and the impression of coherence, is achieved when relations are apparently not signalled. Signalling has traditionally been taken to refer to conjunctions or discourse markers which link propositions. I will propose that signalling is actually quite prevalent, if we broaden our definition of signalling devices. I will report on the results of our annotation (Taboada and Das 2013) of the RST Discourse Treebank (Carlson et al. 2002), which shows that the vast majority of relations are signalled by at least one device. I will describe the taxonomy of signalling devices, including semantic relations, syntactic structure, punctuation and discourse markers, and will provide detail on the types of signalling devices found for various relations.

Selected references

Carlson, L., Marcu, D., & Okurowski, M. E. (2002). RST Discourse Treebank, LDC2002T07 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *An Introduction to Functional Grammar* (4th ed.). London: Arnold.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text, 8*(3), 243-281.

Martin, J. R. (1992). *English Text: System and Structure*. Amsterdam and Philadelphia: John Benjamins.

Taboada, M., & Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse, 4*(2), 249-281.

Conversational Eliciture

Andrew Kehler

University of California San Diego

(Contains joint work with Jonathan Cohen and with Hannah Rohde)

Zipf (1949) famously posited two opposing desiderata in language design: The AUDITOR'S ECONOMY, which dictates that languages should be expressive enough to allow hearers to readily recover the speaker's message, and the SPEAKER'S ECONOMY, which dictates that languages should allow speakers to get their message across efficiently. One way that speakers manage to be efficient while remaining expressive is by designing their utterances to take advantage of the hearer's mental state and capacity for inference to communicate more than what is explicitly said. The sources of such PRAGMATICALLY-DETERMINED ASPECTS OF SENTENCE MEANING have occupied the attention of researchers interested in the semantics and pragmatics of language for many years, and became an industry of its own after the seminal work of Grice (1975).

In this talk, we focus on a form of pragmatically conveyed content that, we claim, fails to fit neatly into any of the types of enrichment typically addressed in the pragmatics literature. Consider:

(1a) A jogger was hit by a car in Palo Alto last night. (Hobbs, 1990)

(1b) A rapper was hit by a car in Palo Alto last night.

(2a) The drug-addled undergrad fell off of the Torrey Pines cliffs. (cf. Webber, 1991)

(2b) The well-liked undergrad fell off of the Torrey Pines cliffs.

(2c) The normally risk-averse undergrad fell off of the Torrey Pines cliffs.

(3a) The boss fired the employee who was embezzling money. (Rohde et al. 2011)

(3b) The boss fired the employee who was hired early last year.

(3c) The boss fired the employee who won numerous corporate awards.

Although not entailed, the indefinite *a jogger* in (1a) strongly invites the defeasible inference that the victim was jogging at the time of the accident. In contrast, the analogous inference for (1b) -- that the rapper was rapping at the time of the accident—is not normally evoked. Similarly, (2a) invites the inference that the drugs caused the undergrad to fall off of the cliff, while (2b) does not invite the corresponding inference that being well liked was a cause of the falling. Further, (2c) yields a counter-to-expectation inference, leading us to be surprised that a normally risk-averse undergrad would fall off of the cliffs. This same three-way distinction characterizes the relative clause modifiers in (3a-c) as well.

We posit that these inferences do not follow directly from the procedures that have been argued to underlie other sorts of pragmatic enrichment, such as from a violation of communicative (e.g., Gricean) norms based on principles of rationality/cooperativity (as in IMPLICATURE), or the need to fill in a value for an otherwise unsaturated grammatical or semantic parameter (as in Bach's IMPLICITURE). We argue instead that they follow from more basic, general cognitive (not specifically linguistic) strategies for building mental models of the world that draw on types of experiential knowledge and associative principles that are already well known to this audience: those that determine the coherence of passages *across* clauses. That is, the inferences at play in (1a)-(3a) are recognizable as those that characterize coherent interpretations of (4)-(6):

(4) A man was jogging in Palo Alto last night. He was hit by a car.

(5) An undergrad fell off of the Torrey Pines cliff. She was on drugs.

(6) The boss fired the employee. He was embezzling money.

A crucial difference, however, is that unlike (4)-(6), there is no requirement to infer a coherence relationship between the constituents in (1a)-(3a), per the felicity of (1b)-(3b).

For want of a term of art, we brand the phenomenon as ELICITURE, selected to capture the fact that a speaker, by choosing a particular form of reference, intends to elicit such inferences on the part of her hearer. The importance of accounting for such inferences goes beyond the recovery of implicit but nonetheless communicated content. It is also crucial for the interpretation of explicit linguistic expressions. To provide an example, we describe a case study examining pronoun interpretation. A passage completion experiment was conducted using stimuli like (3a-b) as context sentences, presented to participants with or without an additional pronoun prompt. Whereas accounts of pronoun interpretation that appeal primarily to surface-level contextual factors find little to distinguish contexts (3a-b), a Bayesian analysis (Kehler et al. 2008; Kehler & Rohde 2013) predicts a difference, through an interconnected chain of referential and coherence-driven dependencies. First, it predicts fewer continuations that explain the context event in (3a) than (3b), since the relative clause in (3a) already provides a cause. Second, this difference is predicted to yield a difference in who gets referred to first: Since implicit causality verbs like 'fire' impute causality to the object, a greater number of explanation continuations for (b) should lead to a greater number of first-mentions of the object. Third, the analysis predicts that pronoun production should not be affected by the relative clause manipulation, and instead only by grammatical role. Finally, the relative clause and prompt manipulations are both expected to affect pronoun interpretation. All of these predictions were confirmed. As such, pronoun interpretation biases, but not production biases, are sensitive to whether or not an implicit explanation can be inferred from context, revealing precisely the asymmetry predicted by the Bayesian analysis.

Finally, as if the annotation of coherence relations between clauses was not difficult enough, the annotation of such relations within the sentence will undoubtedly be more so. We will conclude the talk by discussing some of these issues. Those notwithstanding, there is little doubt that an attempt to perform such an annotation on corpus data, even if not fully successful, would move us far forward in our understanding of the phenomenon of eliciture.

References

Bach, K. (1994). Conversational impliciture. *Mind and Language*, 9(2), 124--162.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics, Volume 3*, pages 41--58. Academic Press, New York.

Hobbs, J. R. (1990). *Literature and Cognition*. CSLI Lecture Notes 21, Stanford, CA.

Kehler, A., Kertz, L., Rohde, H., and Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, 25(1), 1--44.

Kehler, A. and Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics,* 39(1-2), 1--37.

Rohde, H., Levy, R., and Kehler, A. (2011). Anticipating explanations in relative clause processing. *Cognition*, 118, 339--358.

Webber, B. L. (1991). Discourse Modelling: Life at the Bottom. *Proceedings of the AAAI Fall Symposium Series on Discourse Structure in Natural Language Understanding and Generation*, American Association for Artificial Intelligence, Asilomar, CA, 146--151.

Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort. Addison Wesley, Cambridge.

COLLABORATIVE RESEARCH TALKS

Detecting Simplex Subordinators in Turkish

Faruk Acar, Deniz Zeyrek, Ruket Çakıcı

Middle East Technical University, Ankara

We present the initial results for automatically detecting simplex subordinators in Turkish Discourse Bank (TDB). These are referred to as converbs, i.e., nonfinite verb forms with specialized suffixes.

TDB is a ~400K word multi-genre corpus of written texts annotated for discourse relations in the PDTB style (Zeyrek et al., 2013). TDB mainly annotates explicit connectives, their modifiers and two arguments. Discourse connectives may connect clauses or nominalizations by coordinating conjunctions (and, but), subordinating conjunctions (although, because) or discourse adverbials (nevertheless).

In Turkish, subordination is primarily controlled by morphology through subordinating suffixes, most of which have a converbial function. Zeyrek&Webber (2008) identify two kinds of subordinators with a discourse connective role: simplex subordinators (converbs), and complex subordinators, i.e. a postposition co-occurring with a converb. The aim of this study is to 1) identify simplex subordinators in Turkish and annotate them and 2) provide methods to detect them automatically. The extraction and annotation of simplex subordinators will enrich TDB 1.0. We hope to provide a set of meaningful features for automatic discourse connective classification for morphologically rich languages like Turkish.

We used a predetermined list of subordinator suffixes (-AcAğInA 'rather than', AcAğIndAn/dIğIndAn 'due to', -AlI 'since', -ArAk 'by means of', -dIğIndA 'when', -dIkçA 'as long as', -IncA 'when, as per', -Ip 'and then', -ken 'while', -sA 'if'). The list items were then searched in morphologically analyzed TDB texts. A two-level morphological analyzer (Oflazer, 1994) and a morphological disambiguator (Sak et al., 2008) were utilized for this task. The accuracy of the disambiguator is reported as 87.67% by Eryiğit (2012) on the Turkish Dependency Treebank. Then, two independent annotators annotated a sample of TDB converbs categorically: converbs that are simplex subordinators are considered as the true category, converbs with other discursive roles and non-discursive roles are considered as false. Inter-annotator agreement is measured by Cohen's Kappa. Except for –ken and –AlI, we found a good level of agreement between annotators. Disagreements were resolved with an expert and gold standard annotations were created. After finalizing the annotations, a set of intuitive rules were derived. Using this rule based classifier, we obtained precision (0.88), recall (0.91) and accuracy (0.85) values for the true category. Precision (0.73) and recall (0.66) values are relatively low for the false category (complex subordinators, discourse adverbials, alternative lexicalizations and non-discourse usages of subordinators, e.g. headless relative clauses, lexicalized forms of converbs, reduplications). An automatic Naïve-Bayes classifier was also trained by using 60% of the annotated data, leaving 40% as the test set. Precision (0.84), recall (0.95) and accuracy (0.83) metrics for the true category are found to be lower than the rule-based classifier's results.

Error analysis shows that there are mainly two causes of misclassification for both techniques: the lexicalized form of some converbs (bilerek 'intentionally') and the morphological disambiguator errors. The study suggests that rule-based and Bayesian classifiers can be used to determine the converbs' discourse role in Turkish with high accuracy if we can access lexicalized forms of converbs and a gold-standard morphological analysis of the text.

References

Eryiğit, G. (2012). The Impact of Automatic Morphological Analysis & Disambiguation on Dependency Parsing of Turkish. LREC 2012, Istanbul.

Oflazer, K. (1994). Two-level description of Turkish morphology. Literary and linguistic computing, 137-148.

Sak, H., Güng.r, T., & Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. Advances in natural language processing, 417-427.

Zeyrek, D., & Webber, B. (2008). A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. Proceedings, The 6th Workshop on Asian Language Resources, 65-71.

Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., & Çakıcı, R. (2013). Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. Dialogue & Discourse, 174-184.

Assessing the validity of annotation guidelines: towards multimodal equivalence of DSDs

Ludivine Crible[1], Sandrine Zufferey[2]

1Université Catholique de Louvain, 2Université de Fribourg

Cross-linguistic studies of discourse-structuring devices (DSDs) often face various methodological problems regarding the applicability of a single annotation protocol to very diverse data-driven sets of items from several languages. These elements include, in our view, relational, connecting devices (*but, or, so*) and non-relational discourse markers (*you know, well*). Both types can be described as metadiscursive instructions to the hearer on how to interpret an utterance (Hansen 2006, Brinton 2008). Ideally, such a protocol should overcome language-specific preferences and encompass all possible actualizations of DSDs, in different contexts and speech situations.

This presentation reports the outcome of a joint work conducted in the framework of a three-week Short Term Scientific Mission (STSM). The purpose of this STSM was to test and improve an annotation scheme (Crible 2014) that was originally designed for spoken discourse markers (or DSDs) over a variety of situations in French and English. The focus of our common research is threefold: (1) to assess the replicability of the coding scheme by several analysts, (2) to assess its applicability to other languages (with new experiments in German) and (3) its applicability across modalities (from speech to writing).

The protocol consists of an adaptation of the PDTB annotation guidelines (Prasad et al. 2007), and more precisely its revised version by Zufferey et al. (2012), so that it may fit the needs of spoken data analysis. The main modifications concern the grouping of certain discourse relations together to avoid hesitations and discordances in the annotation process, and the addition of functions that are more specific to speech, such as hedging or monitoring. All these functions have been grouped into four functional domains, that partly map existing taxonomies such as Redeker's (1990), Sweetser's (1990) or Gonzalez's (2005).

Although the PDTB framework has been adapted before to other genres (e.g. Prasad et al. 2011) and languages (e.g. Oza et al., 2009 for Hindi, Zeyrek et al., 2013 for Turkish), its applicability to speech is rather innovative. Our contribution will thus serve as a test bed to evaluate the equivalence of different guidelines for speech and writing, as well as to identify possible weaknesses to look for in multilingual and multimodal coding schemes. For this purpose, inter-rater agreement analysis will be performed on different datasets, thus improving in return the

operationality of the protocol. The model in its final stage is thought to offer some experience into assessing and improving annotation guidelines, as well as a proposal for an interoperable taxonomy and its own coding system.

References

Brinton, L. 2008. *The comment clause in English*: Syntactic origins and pragmatic development. Cambridge: CUP.

Crible, L. 2014. "Selection and functional description of DMs in French and English: Towards crosslinguistic and operational categories for contrastive annotation". Paper presented at the International workshop "Pragmatic markers, discourse markers and modal particles : what do we know and where do we go from here?", October 16-17, Como, Italy.

Gonzalez, M. 2005. "Pragmatic markers and discourse coherence relations in English and Catalan oral narrative". *Discourse studies* 7: 53-86.

Hansen, M.-J. M. 2006. "A dynamic polysemy approach to the lexical semantics of discourse markers (with an exemplary analysis of French *toujours*)". In K. Fischer (ed.), *Approaches to discourse particles*: 21-41.

Oza, U., Prasad, R., Kolachina, S., Sharma, D. M. & Joshi, A. 2009. "The Hindi Discourse Relational Bank". In *Proceedings of the Third Linguistic Annotation Workshop*: 158-161. Association for Computational Linguistics.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A. & Joshi, A. 2007. "The Penn Discourse Treebank 2.0 annotation manual". Institute for Research in Cognitive Science.

Prasad, R., McRoy, S., Frid, N., Joshi, A. & Yu, H. 2011. "The biomedical discourse relation bank". *BMC Bioinformatics* 12: 188.

Redeker, G. 1990. "Ideational and pragmatic markers of discourse structure". *Journal of Pragmatics* 14: 367-381.

Sweetser, E. 1990. *From etymology to pragmatics*. Cambridge: CUP.

Zeyrek, D., Demirsahin, I., Sevdik Calli, A. B. & Çakici, R. 2013. "Turkish Discourse Bank : Porting a discourse annotation style to a morphologically rich language". *Dialogue and discourse* 4/2: 174-184.

Zufferey S. Degand L., Popescu-Belis A. & Sanders T. 2012. "Empirical validations of multilingual annotation schemes for discourse relations." *Proceedings of the Eighth Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation*. Pisa, Italy, 77-84.

Crowd-Sourcing Concurrent Relations

Anna Dickinson, Hannah Rohde, Annie Louis, Chris Clark and Bonnie Webber

University of Edinburgh

While discourse relations can be signaled explicitly with conjunctions (Ex. 1) or adverbials (Ex. 2),

(1) "We've started trying just about anything to keep sales moving in the stores," says Kim Renk, a Swank vice president. <u>But</u> there are limits. [wsj0280][1]

[1]References of this form are to files in the 1989 *Wall Street Journal* section of the Penn TreeBank.

(2) They both called it a “welcome home" gathering. Nevertheless, an ANC rally by any other name is still an ANC rally. [wsj0559]

we also find sentences (Ex. 3-5) with both forms of DRD:

(3) If that became public knowledge, the last bit of influence she had over her bank would be gone. So instead she hardened her soul and pretended to be a banker who was working her own will. [COCA]

(4) It's past ten. I could go to bed but instead I crawl out the window onto my little roof with the joint behind my ear. [COCA]

(5) Appealing to a young audience, he scraps an old reference to Ozzie and Harriet and instead quotes the Grateful Dead. [wsj 1615]

In such cases, the conjunction and adverbial can each signal a distinct discourse relation. A previous crowd-sourced study of four adverbials that can co-occur with conjunctions (Jiang, 2013) asked respondents to insert a conjunction that had either been removed from a passage (using examples like 3-5) or never appeared (as in 6-7). The study of 100 passages (each viewed by 52 respondents) showed a strong distinction between different adverbials: With after all, they had an overall preference for because, whereas with instead, they varied passage-by-passage, reliably inserting but in (6) and so in (7):

(6) Logically, she should be dead. Instead, she feels fine, caring for her daughters and walking a pedometer-measured two miles a day. [COCA]

(7) He suspected he shouldn't say that. Instead he lied. [COCA]

We call these *concurrent relations* because, as Jiang's study shows, even without an explicit conjunction, two separate senses are concurrently conveyed. Ours is a study of concurrent relations. We aim to establish which discourse adverbials co-occur with conjunctions and which conjunctions each adverbial licenses and favors. Concurrent relations are a challenge for psycholinguists interested in whether an explicit linguistic signal is only needed when other evidence is too weak (a matter of discourse efficiency). They are also a challenge for language technology, which has heretofore assumed that a discourse relation was either signalled explicitly or inferred based on other clues, but not both at the same time.

We have used Google N-grams to categorize discourse adverbials by frequency (high vs. low) and co-occurence with conjunctions (skewed vs. uniform). Frequency of use is shown in Figure 1, where ‘use’ was restricted to an approximation of adverbial contexts (e.g., sentence-initial overall followed by a comma to alter out adjectival uses). While no adverbial was found to occur most frequently with a conjunction, the majority of adverbials had clear conjunction preferences (Figure 2), while some had a broader distribution (Figure 3). We speculate that adverbials with a strong conjunction bias may have an underlying semantic reason for that choice (perhaps an adverbial that signals consequence can only combine with a conjunction of that type), whereas adverbials that combine more flexibly with a set of conjunctions may permit properties of the linked propositions to determine the relevant conjunction.

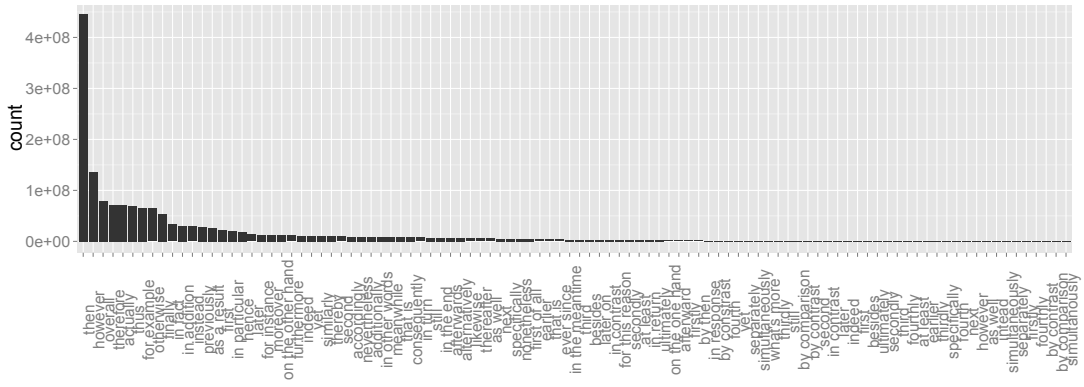
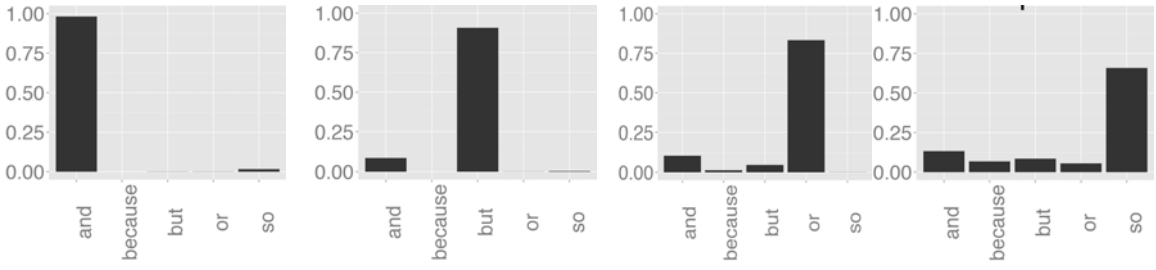


Figure 1: Adverbial ngram frequency: counts consist of instances with or without conjunction

Using a web interface akin to the one shown in Figure 4, the project will crowd-source judgments so as to replicate Jiang's experiment more systematically and with broader coverage. The first dataset targets 20 adverbials, including the four tested by Jiang *(actually, after all, first of all, for example, for instance, however, in fact, in general, in other words, indeed, instead, nevertheless, nonetheless, on the one hand, on the other hand, otherwise, specifically/more specifically, then, therefore, thus*), appearing in 50 passages each. As in Jiang's experiment, in half the passages, an overt conjunction was present in the original text, while in the other half, the adverbial originally appeared alone. Passages have been drawn primarily from the *New York Times Annotated Corpus* (Sandhaus, 2008) and the *Corpus of Contemporary American English*, using web text as needed for rarer conjunction-adverbial combinations.

A pre-test was conducted on the 20 targeted adverbials to pilot the interface and test whether other adverbials show patterns similar to what Jiang (2013) found for her small set. Three of the co-authors, who were naive to the status of each passage (as underlyingly implicit vs. explicit), judged a total of 760 passages, 260 implicit and 500 explicit. Their judgments confirmed Jiang's results: For example, in the implicit after all passages, at least one judge responded with because in all but one case; likewise, in the after all passages which had originally appeared with an explicit conjunction, judges over-estimated the use of because for passages with an original and or but. In contrast, other adverbials were found to be more sensitive to the passage they appeared in: In the instead passages, for example, there was no single preferred conjunction. The example depicted in Figure 4 had an explicit so in the original text, though other senses are available (and and but are both plausible).



a. therefore b. nevertheless c. otherwise d. for example

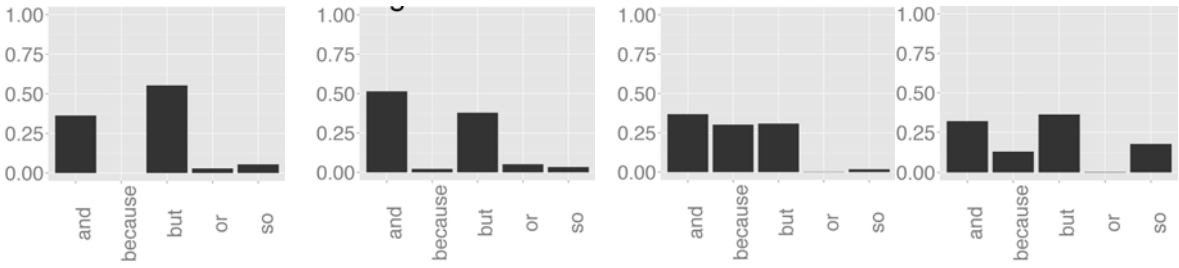
Figure 2: Sample adverbials with skewed distributions across conjunctions

The pre-test pointed to several potential sources of judgment variability that we must be alert to in the crowd-sourced results. First, different readers may interpret the instructions differently. Currently, the instructions read (in part) *Your job is to make explicit the meaning that links the adjacent text spans. You must make a choice even if the insertion leads to an awkward or lengthy sentence, as long as you think the word brings out the meaning that links the adjacent text spans.* It is possible that readers may not understand that they are being encouraged to make a **sense** judgment about the author's intended meaning, rather than a **stylistic** judgement about passage readability. The different specificity of the conjunctions may add further variability: Can we assume that a judgment to insert so is a more specific – but not conflicting – judgement than and, such that a passage for which readers selected a mix of so and and is not necessarily an ambiguous passage? Even if yes, the pre-test has revealed that certain passages permit multiple conjunctions that are not necessarily more or less specific variants of each other. Consider example (8):

(8) You got to be nice to them ____otherwise, they're not going to be nice to you.

While this passage had an explicit or in the original text, pre-test readers assigned a mix of or and because. And while these two conjunctions are typically associated with very different meanings, they seem to convey the same meaning here.

In summary, in contrast to models of DRD usage that assume that if one DRD is present in a passage, it signals a single relationship, our data raise three points: (i) there are many cases in which a DRD does not act alone (because afterall); (ii) a given DRD need not admit the presence of only one additional relationship (instead with but or so or because); and (iii) discourse adverbials differ in the way they combine with possible conjunctions – some demonstrating a clear preference, while others are more flexible.



a. instead b. in general c. after all d. first of all

Figure 3: Sample adverbials with broader distributions across conjunctions

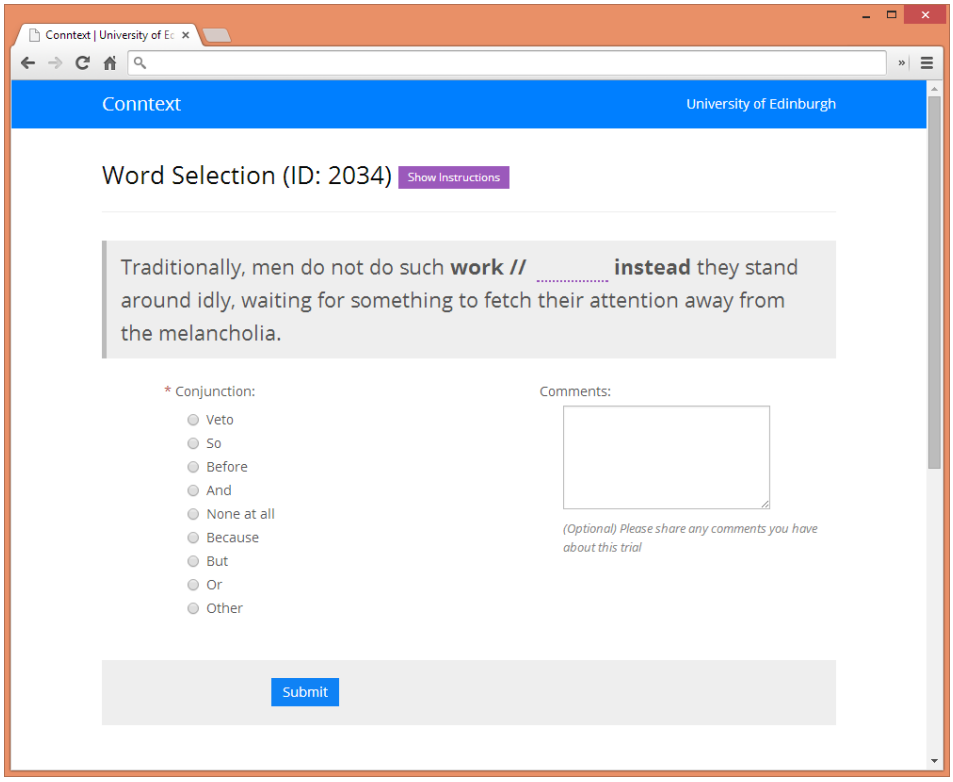


Figure 4: Experiment interface used during the pre-test, permiting either a conjunction to be selected or the example to be 'vetoed'. The passage depicted appeared with so in the original text.

References

Xi Jiang. *Predicting the use and interpretation of implicit and explicit discourse connectives*. PhD thesis, Linguistics and English Language (LEL), University of Edinburgh, 2013. MSc in English Language.

Evan Sandhaus. New York Times Annotated Corpus: Corpus overview. LDC catalogue entry LDC2008T19, 2008.

PDiT-GECCo Scientific mission: towards comprehensive approaches to discourse-structuring devices

Anna Nedoluzhko[1], Ekaterina Lapshinova-Koltunski[2], Kerstin Kunz[3]

1Charles University in Prague, 2Saarland University, 3Heidelberg University

In this presentation, we introduce preliminary results of an experimental comparison of two approaches to discourse analysis: the one within the project 'German-English contrasts in cohesion (GECCo)' at the Saarland University and the other within the Prague Discourse Treebank (PDiT). The comparative analysis proceedes within the Short-Term Scientific Mission of the COST-Textlink initiative, realized as the visit to the Saarland University (January 15 to February 8, 2015).

The analysis in GECCo is based on a large-scale parallel and comparable corpus of English and German, containing various text types annotated for different types of cohesive phenomena, including discourse connectives, coreference relations, ellipses, substitution and chains of lexical cohesion, as well as their structural and functional subtypes. Prague Discourse Treebank (PDiT) contains a detailed annotation of discourse relations (discourse connectives, discourse units linked by them and semantic relations between these units), ellipses, coreference and bridging (associative anaphoric) relations.

Our initial pilot observations have shown that both conceptions of textual phenomena analysis are very close in several respects, and their further comparison can bring interesting results, e.g. compilation of an interoperable annotation scheme applicable for all aspects of discourse phenomena. However, the classification of annotation categories or features is based on different frameworks: in GECCo, explicit linguistic means that signal a discourse relation are taken into account, while PDiT includes rather cognitive / implicit relations of coherence. For instance, the approach to ellipses is different in the Prague and GECCo corpora – different rules are used for ellipses reconstruction in each approach. Other important distinctions are in the field of coreference relations and lexical cohesion/bridging. Moreover, the annotation scheme in PDiT was primarily applied on journalistic texts, whereas the GECCo corpus includes texts belonging to different genres. So, analysing both annotation schemes applied on the same text will show which discourse phenomena can be better captured by which approach, also taking into account differences in genre.

The interoperability of both conceptions will be tested within a pilot experiment which includes comparison of annotation schemes. This will allow us to think of ways of applying both without losing important categories and aspects, and to discuss specific details of annotating discourse relations and other discourse phenomena (e.g. lexical cohesion). Additionally, we will be able to see if both schemes perform well on the same text types. The revealed commonalities and differences will provide us with topics for further discussion and analysis.

"On the one hand" as a Cue to in the Comprehension of Discourse Structure

Vera Demberg[1], Hannah Rohde[2], Merel Scholman[3], Chris Cummins[2], Emily Nicolet[2]

1Saarland University, 2University of Edinburgh, 3Universiteit Utrecht

Given evidence of anticipation within sentences (for upcoming sounds, words, and syntactic structures; Delong, et al. 2005; Kamide, et al., 2003; Levy, 2008), an open question is how comprehenders use cross-sentence cues to anticipate upcoming relationships between sentences. Within sentences, words combine via syntactic rules to determine what structures are possible. Between sentences, however, the resulting discourse structure is less constrained. Models of discourse coherence typically target relations that can be inferred to hold between pairs of propositions (Asher & Lascarides, 2003; Hobbs, 1979; Kehler 2002; Mann & Thompson, 1988; Prasad et al. 2008), with few hard constraints regarding the eventual structure of the discourse (cf. Roberts, 1996). Nevertheless there are cases in which the possible relations that could hold between a current sentence and a subsequent sentence are restricted. Existing work primarily targets local effects (e.g., verb-driven biases for the immediately upcoming sentence; Kehler et al., 2008; Staub & Clifton, 2006). Here we consider the contrast relation between sentences marked with *On the one hand* and *On the other hand*. Based on evidence of syntactic prediction (e.g., dependencies like *either…or*, Staub, 2006), our goal is to test whether comprehenders use *On the one hand* as a cue to anticipate upcoming discourse structure and furthermore how their processing of *On the other hand* is influenced by intervening material.

The expression *On the one hand* signals that a subsequent proposition will provide a contrast and will likely be marked with the expression *On the other hand*. The anticipation of a subsequent contrast can be satisfied immediately (e.g., *Joe was interested in a car. On the one hand, it looks flashy. On the other hand, it doesn't get very good mileage*.). If the expected contrast is delayed, comprehenders are predicted to process *On the other hand* differently depending on the type of intervening material.

Self-paced reading study: Participants (n=60, recruited from Amazon's Mechanical Turk) read sentences phrase-by-phrase via a web-based interface (IbexFarm). The intervening material varied—either leaving the expectation for contrast unfulfilled by mentioning causal information (1a,1b) or providing a contrast that could plausibly resolve the expectation for contrast (1c). Reading times were measured at *On the other hand.*

SentenceA: Joe was interested in a car.

SentenceB:

(a) *On the one hand*, he would like to buy it, because it looks flashy.

(b) *On the one hand*, it looks flashy, so he would like to buy it.

(c) *On the one hand*, he would like to buy it, but he might try leasing it first.

SentenceC: *On the other hand*, it doesn't get very good mileage.

As predicted, *On the other hand* in SentenceC was read faster following conditions with causal information (1a,1b) than contrastive information (1c), suggesting that participants used *On the one hand* as a cue to an upcoming contrast and were surprised (as evidenced by their reading-time slowdown) by *On the other hand* when they had already encountered a plausible contrast. Comprehenders thus use discourse connectors to predict discourse relations and can maintain such predictions across clauses.

References

Asher, N., & Lascarides, A. (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

DeLong, K. A., Urbach, T. A., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nature Neuroscience, 8, 1117-1121.

Hobbs, J. R. (1990). Literature and cognition. Stanford, CA: CSLI. Lecture Notes 21.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language, 49, 33-156.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI Publications.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008*). Coherence and coreference revisited*. Journal of Semantics, 25, 1–44.

Levy, R. (2008). *Expectation-based syntactic comprehension.* Cognition, 106, 1126-1177.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text, 8, 243-281.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. & Webber, B. (2008). *The Penn Discourse Treebank 2.0.* In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC). Marrakech, Morocco.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. OSU Working Papers in Linguistics, 49: Papers in Semantics.

Rohde, H. & Horton, W. (2014). Anticipatory looks reveal expectations about discourse relations. Cognition, 133(3), 667-691.

Staub, A., & Clifton Jr, C. (2006). *Syntactic prediction in language comprehension: Evidence from either... or*. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32(2), 425.

The Development of the Annotation Procedures of DSDs in the HuComTech Corpus

Ágnes Abuczki[1] & Péter Furkó[2]

1MTA-DE Research Group for Theoretical Linguistics, 2Károli Gáspár University of the Reformed Church in Hungary

The multimodal HuComTech corpus (Hunyadi & al. 2011) contains the audio and video material as well as the aligned transcript and multi-level annotation of 118 simulated job interviews and 118 informal conversations in Hungarian between 2 interviewers and 118 interviewees, and altogether adds up to about 50 hours of interaction. The corpus was originally designed to study the interrelation of the multimodal aspects (prosody, visual signals, etc.) of spontaneous human-human interaction.

Verbatim transcriptions involved the labelling of non-lexical/nonverbal vocalisations, such as filled pauses and hesitations, as well. The transcribed dialogues were segmented into utterances. The pitch movements in the intonational phrases of the speakers were automatically annotated using a Prosogram-based application (Mertens 2004) further developed by the HuComTech Project (Hunyadi & al. 2012).

The video contents of the HuComTech corpus were annotated in several tiers in ELAN 4.6.1 (Brugman & Russel 2004), marking several types of facial expressions, gaze directions, eyebrow movements, head movements, hand movements, hand shape types, posture, touch motion types, deictic gestures and emblems. When audio and video annotations were completed, we added the segmentation and functional labelling of a few selected DSDs to 50 recordings between 2012 and 2013. Following the segmentation of the DSDs, their acoustic features (duration, pitch and intensity values, pitch movement and surrounding pause) were automatically extracted using a Praat script (Boersma & Weenink 2007) and were merged and exported into .eaf files and finally queried in ELAN. We also created an additional scheme in ELAN 4.6.1 which entirely describes the functional spectrum of DSDs and covers all the domains of discourse. Within each of these discourse domains we offered mutually exclusive categories so that the annotator can attach only one label/tag at one functional discourse level, but may attach a label at any number of the large functional categories. Therefore, a single DSD can be described in several domains of discourse along the following aspects of interaction:

Own Speech Management: lexical search, reformulation, giving example, explanation

Attitude Marking: approximation, emphasis, PFM_booster, PFM_hedge, rhetorical question

Interpersonal Functions: agreement, emphasis, asking for reassurance, expressing sympathy

Structural Conversation Management: turn-take (distinction of preferred second pair parts and dispreferred second pair parts), turn-keep, turn-give (end-of-turn), (listener's) backchannel

Thematic Control: introducing topic initiation, topic elaboration, topic change, marking concession

Information Management: signalling *new information*, *evidentiality marker*.

The annotation tool, ELAN 4.6.1 enables tagging multiple functions to a single DSD, which is necessary because most oral DSDs simultaneously perform multiple functions.

References

Boersma P. & Weenink, D. 2007. *Praat: doing phonetics by computer 5.0.02.* University of Amsterdam: Institute of Phonetic Sciences. http://www.praat.org

Brugman, H. & Russel, A. 2004. Annotating multi-media / multi-modal resources with elan. In: Lino, M., Xavier, M., Ferreire, F., Costa, R., Silva, R. (Eds). *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)* (pp. 2065–2068). Lisbon: Portugal.

Hunyadi, L, Bertok, K, Nemeth T, E, Szekrenyes, I, Abuczki, A, Nagy, G, Nagy, N, Nemeti, P. & Bodog, A. (2011) The outlines of a theory and technology of human-computer interaction as represented in the model of the HuComTech project. In: *2nd International Conference on Cognitive Infocommunications (CogInfoCom)*. Budapest, 7-9 July, 2011. Budapest: IEEE. http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?reload=true&arnumber=5999491

Hunyadi, L, Szekrényes, I, Borbély, A. & Kiss H. (2012). Annotation of spoken syntax in relation to prosody and multimodal pragmatics. In: *Proceedings of 3rd Cognitive Infocommunications Conference.* Kosice: IEEE Conference Publications. 537–541.

Mertens, P. (2004). Un outil pour la transcription de la prosodie dans les corpus oraoux. *Traitement Automatique des Langues,* 45 (2): 109–130.

The division of text links in the Standard Croatian language

Dijana Ćurković

Institute of Croatian Language and Linguistics

The poster demonstrates the text links or conjunctions in the Standard Croatian language. It presents the division of types of conjunctions relative to the words, phrases, and types of clauses they are connecting. Thus, we can divide the conjunctions into two main types: the coordinate conjunctions and the subordinate conjunctions. The coordinate conjunctions are further divided into conjunctive, disjunctive, contrasting and concluding, while the pronouns will most often be in the role of the subordinate conjunctions, which are relative to the main clause and the type of subordinate clause.

The Louvain Corpus of Annotated Speech – French (LoCAS-F)

Liesbeth Degand, Laurence J. Martin, Anne Catherine Simon

Université Catholique de Louvain

The Louvain Corpus of Annotated Speech – French is a dataset of spoken French segmented into Basic Discourse Units (BDUs) (Degand & Simon 2009ab). A Basic Discourse Unit results from the mapping of a syntactic clause and a major intonation unit, giving rise to different types of discourse units (congruent, syntax-bound, intonation-bound, regulatory) – see Figure 1.

The corpus consists of twelve distinct genres or communicative situations (academic speech, political address, interview, narrative…) categorized in terms of their degree (scale from 1 to 3) of preparation, interaction, and broadcasting. It counts 36,912 words, corresponding to 3:11 hours of speech (Degand, Martin, Simon 2014).

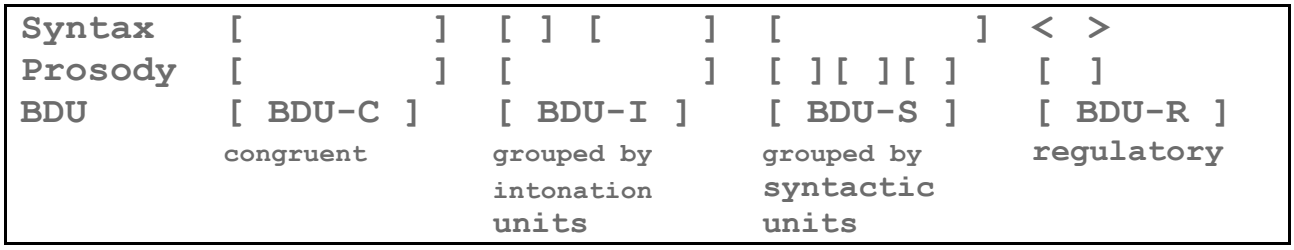


Figure 1 – Four BDU types

Annotations are provided on two distinct levels : the morpho-syntactic and the prosodic level. Starting point for the syntactic segmentation and annotation is the word-based orthographic transcription of the discourse to be analysed in a Praat tire (Boersma & Weeninck 2012), to which two more tires are added: the first for the segmentation into syntactic dependency clauses, the second for the annotation of functional sequences. This syntactic annotation is entirely manual following the theoretical principles of dependency syntax, i.e. a verbal micro- syntax in which the verb (or any other governor) and its governed complements are central. The syntactic analysis leaves us with a number of ungoverned segments, which belong to the macro-syntax rather than to the micro-syntax (Berrendonner 2002). They comprise so-called 'associés' ('adjuncts') and discourse markers, which are not governed by the main clause, but are semantically or pragmatically linked to the whole dependency clause. They have a non- autonomous status in discourse, whilst being syntactically independent. The second and final step in the syntactic annotation process consists in cutting up each dependency clause into so- called "functional sequences", i.e. clausal constituents that occupy a main syntactic function like Verb, Subject, Object, etc. (Bilger & Campione 2002). So far, only discourse markers with weak clause-association (Schourup 1999) have been identified, they were not categorized. The annotation protocol developed by Crible (2014) will be applied to the 1334 identified discourse markers.

The prosodic annotation is semi-automatic following the procedure developed by Mertens & Simon (2009). A major prosodic boundary (///) is established when one of the following cues is detected on the final syllable of a word: a subsequent silent pause longer than 250 ms; an extra-lengthening; a sharp rise of f0. We manually exclude a boundary which coincides with a hesitation mark, since it has been demonstrated that hesitations are not confounded with prosodic breaks in discourse processing. An intermediate prosodic boundary (//) arises when the final syllable of a word is lengthened (the syllable is two times longer than the syllables in the immediate surrounding context), bears a sharp rise of f0 (superior to four semi-tones), or is higher than adjacent syllables (higher than five semi-tones). Minor prosodic boundaries are not taken into consideration because the agreement between automatic detection and manual validation is very low (see Mertens & Simon 2009).

The second step of the prosodic annotation consists in attributing an intonation contour to each prosodic boundary. Four alternatives exist: Continuation (rising f0 movement), Finality (falling or low f0), Focus (sharp falling from high to low contour) and Suspense (flat and lengthened contour).

References

Berrendonner, Alain. 2002. "Les deux syntaxes." *Verbum* XXIV (1-2): 23-35.

Bilger, Mireille and Campione, Estelle. 2002. "Propositions pour un étiquetage en séquences fonctionnelles." *Recherches sur le Français Parlé* 17 : 117–136.

Boersma, Paul and Weenink, Daniel. 2012. Praat: doing phonetics by computer [Computer program]. Version 5.3.32, retrieved 17 October 2012 from http://www.praat.org/

Crible, Ludivine. 2014. Selection and functional description of discourse markers in French and English: towards crosslinguistic and operational categories for contrastive annotation. In: International Workshop - Pragmatic

Markers, Discourse Markers and Modal Particles: What do we know and where do we go from here? October 16-17, Como, Italia.

Degand, Liesbeth, Martin, Laurence J. and Simon, Anne Catherine. 2014. "*LOCAS-F : un corpus oral multigenres annoté*.'' In: *CMLF 2014 - 4 ème Congrès Mondial de Linguistique Française*, EDP Sciences: Berlin.

Degand, Liesbeth and Simon, Anne Catherine. 2009a. "Minimal discourse units in spoken French: On the role of syntactic and prosodic units in discourse segmentation." *Discours* 4. URL : http://discours.revues.org/5852 ; DOI : 10.4000/discours.5852

Degand, Liesbeth and Simon, Anne Catherine. 2009b. "Mapping prosody and syntax as discourse strategies: How Basic Discourse Units vary across genres." In *Where Prosody meets Pragmatics: Research at the Interface.* Anne Wichmann, Dagmar. Barth-Weingarten and Nicole Dehé (eds), 79-105. [Studies in Pragmatics]. Bingley: Emerald.

Mertens, Piet and Simon, Anne Catherine. 2009. "Automatic detection of prosodic boundaries in spoken French" Ms. Katholieke Universiteit Leuven and Université catholique de Louvain (Louvain-la-Neuve).

Schourup, Lawrence. 1999. Discourse markers. *Lingua* 107 (3-4): 227-65.

ANNODIS

Stergos Afantenos, Lydia-Mai Ho-Dac, Nicholas Asher, Myriam Bras and Philippe Muller

Toulouse

The ANNODIS resource is a diversified corpus of written French texts enriched with a manual annotation of discourse structures. It was produced as part of the ANNODIS project (ANNOtation DIScursive), financed by the French National Research Agency (ANR). Two mark-ups are given, corresponding to two distinct approaches to discourse organisation: "rhetorical relations" annotation and "multi-level structures" annotation. The Rhetorical Relations annotation aims at providing a complete structure of a text, starting from the segmentation into Elementary Discourse Units to Complex Discourse Units, by linking each unit with at least one rhetorical relation (e.g. contrast, elaboration, result, attribution, etc.)

Semantically, each Elementary Discourse Units contains at least one eventuality description, and often only one.

The multi-level structures annotation aims at identifying discourse structures which may appear at different granularity levels, including very high levels (from 2 sentences up to several sub-sections), and therefore of interest as building blocks in the construction of text. Two multi-level structures are annotated: enumerative structures ad topical chains.

Annotating multi-level structures consisted both in delimiting the covered text segment and identifying their clues.

As a result, the ANNODIS resource is divided in two parts, corresponding for the rhetorical relation annotation of short texts (a few hundred words each) and excerpts from longer documents and for the multi-level structures annotation, of longer (several thousands words each), complete and more complex documents. The final resource provides on the one hand 3188 Elementary Discourse Units and 1395 Complex Discourse Units linked by 3355 rhetorical relations; and on the other hand 991 Enumerative Structures and 588 Topical Chains with 4649 enumerative clues (sequencers, prospective elements, encapsulation, etc.) and 3456 topical chains clues (redenomination, pronouns, etc.).

References

Afantenos S. D., Asher N., Benamara F., Bras M., Fabre C., Ho-Dac L.-M., Le Draoulec A. Muller P., P?ry-Woodley M.-P., Prévot L., Rebeyrolle J., Tanguy L., Vergez-Couret M., Vieu L. (2012). An empirical resource for discovering cognitive principles of discourse organization: the ANNODIS corpus. LREC 2012, Istanbul, Turkey, July 2012.

GECCo: Corpus to Analyse German-English Contrasts in Cohesion

Kerstin Kunz[1], Ekaterina Lapshinova-Koltunski[2], José Manuel Martínez[2], Katrin Menzel[2], Erich Steiner[2]

1University of Heidelberg, 2Saarland University

GECCo is a German-English corpus containing written and spoken texts, cf. Lapshinova et al. (2012), which was created for a contrastive analysis of cohesion in English and German. The whole corpus contains ca. 1.5 Mio tokens and is structured in six subcorpora: German written originals (GO), English written originals (EO), English spoken originals (EO-SPOKEN) and German spoken originals (GO-SPOKEN), translations of German written originals into English (ETRANS) and translations of English written originals into German (GTRANS). The four written subcorpora (EO, GO, ETRANS and GTRANS) were extracted from CroCo, Hansen-Schirra et al., 2012, and consist of texts from eight registers: popular-scientific texts (POPSCI), tourism leaflets (TOU), prepared speeches (SPEECH), political essays (ESSAYS), fictional texts (FICTION), corporate communication (SHARE), instruction manuals (INSTR) and corporate websites (WEB). The two spoken subcorpora contain academic speeches (ACADEMIC) and interviews (INTERVIEW). Further spoken registers have been currently added to the corpus.

GECCo is annotated with information on cohesive devices and their categories, which include both functional and structural subtypes of co-reference, conjunction, substitution, ellipsis and lexical cohesion. The main classifications are taken from Halliday & Hasan and have been adjusted to cross-linguistic comparison.For co-reference chains, our annotations provides information on the number of antecedents, anaphors, chains, as well as chain length. Currently, the corpus is enriched with information on chains of lexical cohesion. Moreover, information on tokens, lemmas, morpho-syntactic features (e.g. case, number, etc.), parts-of-speech, grammatical chunks along with their syntactic functions, clauses, and sentence boundaries are also available in the corpus. The annotation of the written subcorpora was partly imported from CroCo, whereas for the spoken part, we use the Stanford POS Tagger (Toutanova et al., 2003) and the Stanford Parser (Klein and Manning, 2003). The corpus is encoded in the CWB format (CWB, 2010) and can be queried with Corpus Query Processor (CQP; Evert, 2005). These annotation levels provide us with additional information on cohesive types, i.e. for co-reference or conjunctive relations: morpho-syntactic preferences of antecedents and anaphors, position of coordinating conjunctions and conjunctive adverbials in a clause, etc.

For the annotation of cohesion, semi-automatic procedures were applied, which include a rule-based tagging of cohesive candidates and their manual post-correction by humans. The procedures involve an iterative extraction-annotation process, and are based on the option of the CWB tools to incrementally enhance corpus annotations, as query results deliver not only concordances of the

searched structures but also information on their corpus positions. This permits to import the information on queried data back into the corpus. In this way, we annotate candidates for cohesive categories, which are then corrected manually by human annotators with the help of MMAX2 (Müller and Strube, 2006), as visualisation options of this tool allow annotators to decide whether the candidates tagged by the automatic procedures have a cohesive function and belong to the given category. Manual procedures are also used for annotation of co-reference chains, as human annotators manually identify antecedents and link them to the cohesive referring expressions (anaphoras) which were automatically tagged by our system. A detailed description of the semi-automatic procedures of coreference, substitution and conjunctive relations is given in Lapshinova and Kunz (2014).

The annotated corpus is available in XML format and can be queried with CQP. We also provide a CQP-WEB (cf. Hardie, 2012) version which is available via CLAIN-D project.

References

CWB (2010). The IMS Open Corpus Workbench. http://www.cwb.sourceforge.net.

Evert, S. (2005). *The CQP Query Language Tutorial*. Institut fur Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, April. CWB version 2.2.b90.

Hansen-Schirra, S., S. Neumann, and E. Steiner (2012). Cross-linguistic Corpora for the Study of Translations. Insights from the language pair English – German. Series *Text, Translation, Computational Processing*. Berlin / New York: Mouton de Gruyter.

Halliday, M.A.K. and R. Hasan (1976). *Cohesion in English.* London: Longman.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics, 17(3):380–409.

Klein, D. and C. D. Manning (2003). Accurate Unlexicalized Parsing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – Volume 1, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lapshinova-Koltunski, E., K. Kunz, and M. Amoia (2012). Compiling a Multilingual Spoken Corpus. In Tommaso Raso Heliana Mello, Massimo Pettorino, editor, Proceedings of the VIIth GSCP International Conference: Speech and corpora, pages 79–84, Firenze. Firenze University Press.

Lapshinova-Koltunski, E. and K. Kunz (2014). Annotating Cohesion for Multillingual Analysis. In Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, Reykjavik,May 26, 2014.

Müller, C. and M. Strube (2006). Multi-level annotation oflinguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, pages 197-214. Peter Lang, Frankfurt a.M., Germany.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). Feature-rich part- of-speech tagging with a cyclic dependency network. In: NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 173–180, Morristown, NJ, USA. Association for Computational Linguistics.

Towards a Discourse-Annotated Corpus of Finnish: the Finnish PropBank

Veronika Laippala, Jenna Kanerva, A. Missilä, K. Haverinen, , T. Salakoski, Filip Ginter

University of Turku

The Finnish PropBank (Haverinen et al. 2013a) is a hand-annotated corpus of verbal propositions and arguments created on top of Turku Dependency Treebank (Haverinen et al.

2013b), a corpus containing 204,399 tokens (15,126 sentences) from 10 different text genres in Finnish.

The aim of a PropBank, originally developed for English (Palmer et al. 2005), is to add semantic information on top of the syntax structures by specifying the semantic role of each verb argument. The semantic roles are numbered, Arg0 being generally the prototypical agent and Arg1 the patient or the theme.

In addition to the numbered arguments, the PropBank scheme includes altogether 11 general arguments (ArgMs) labelling *location, extent, general, negation, modality, cause, time, purpose, manner, direction* and *discourse.* As noted by Prasad et al. (2014:22-24), despite some differences, many of these relations correspond to sentence-internal relations in PDTB. In addition, the *discourse* label denotes sentence-initial words referring to previous text and thus linking text segments together.

Although the original purpose of PropBank is not to account for discourse relations, the ArgMs offer a good basis for further work on discourse-level phenomena. Altogether, the Finnish PropBank contains 30,255 ArgM relations, of which 2,254 (7,5%) are labelled *discourse* and signal sentence- external relations. The label offers also a possibility for studying the lexical items used for linking: in the PropBank, 312 different words are used for this, of which the most frequent are *kuitenkin (however)* with 209, *myös (also)* with 148 and *lisäksi (in addition*) with 84 occurrences.

References

Haverinen, K.; Laippala, V.; Kohonen, S.; Missilä, A.; Nyblom, J.; Ojala, S.; Viljanen, T.; Salakoski, T. & Ginter, F. (2013a) Towards a Dependency-based PropBank of General Finnish. 2013. Proceedings of the 19th Nordic Conference on Computational Linguistics (NoDaLiDa'13) , pp. 41-57.

Haverinen, K.; Nyblom, J.; Viljanen, T.; Laippala, V.; Kohonen, S.; Missilä, A.; Ojala, S.; Salakoski, T.; Ginter, F. (2013b) Building the essential resources for Finnish: the Turku Dependency Treebank. Language Resources and Evaluation. 2013. DOI: 10.1007/s10579-013-9244-1

Palmer, M.; Gildea, D.;Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. Computational Linguistics, 31(1):71–106.

Prasad, R.; Webber, B.; Joshi, A. (2014) Reflections on the Penn Discourse TreeBank, Comparable Corpora and Complementary Annotation.. Computational Linguistics, doi:10.1162/COLI_a_00204.

Discourse Annotation in the Prague Dependency Treebank 3.0

Jiří Mírovský, Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Eva Hajičová

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

The Prague Dependency Treebank 3.0 (PDT 3.0; Bejček et al., 2013) is the newest version of the Prague Dependency Treebank series. It is a corpus of Czech, consisting of almost 50 thousand sentences annotated mostly manually on three layers of language description: morphological, analytical (surface syntactic structure), and tectogrammatical (deep syntactic structure). On top of the tectogrammatical layer, explicitly marked discourse relations, both inter- and intra-sentential ones, have been annotated.

Annotation of discourse relations in PDT 3.0 is inspired by the PDTB lexical approach of connective identification (Prasad et al., 2008) but it also takes advantage of the Prague tradition of dependency linguistics (see e.g. Sgall et al., 1986). Only discourse relations indicated by overly present (explicit) discourse connectives with clausal arguments (with a predicate verb) have been annotated. The Prague discourse annotation also includes marking of list structures and marking of smaller text phenomena like article headings, figure captions, metatext etc.

Inter-sentential discourse relations have been annotated completely manually, nevertheless taking advantage of various types of information from the tectogrammatical layer. The annotation of intra-sentential relations proceeded first manually for cases where the tectogrammatical layer did not allow for identifying a discourse relation automatically. Afterwards, using the information (mostly) from the tectogrammatical layer, we were able to identify and mark almost 10 thousand out of more than 12 thousand intra-sentential relations automatically (details in Jínová et al., 2012).

The Prague discourse label set was inspired by the Penn sense tag hierarchy (Prasad et al., 2008) and by the tectogrammatical functors (Mikulová et al., 2005). The four main semantic classes, Temporal, Contingency, Contrast (Comparison) and Expansion are identical to those in PDTB but the hierarchy itself is only two-level (see Poláková et al., 2013). The third level is captured by the direction of the discourse arrow.

References

Bejček, Eduard, Hajičová, Eva, Hajič, Jan et al. (2013). *Prague Dependency Treebank 3.0.* Data/software, Charles University in Prague, MFF, ÚFAL. Available at: http://ufal.mff.cuni.cz/pdt3.0/.

Jínová, Pavlína, Mírovský, Jiří, & Poláková, Lucie (2012). Semi-Automatic Annotation of Intra-Sentential Discourse Relations in PDT. In: *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA) at Coling 2012*, Mumbai, India, pp. 43-58.

Mikulová, Marie et al. (2005). *Annotation on the tectogrammatical layer in the Prague Dependency Treebank. The Annotation Guidelines.* Prague: UFAL MFF. Available at: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html.

Poláková, Lucie, Mírovský, Jiří, Nedoluzhko, Anna et al. (2013). Introducing the Prague Discourse Treebank 1.0. In: *Proceedings of the 6th International Joint Conference on Natural Language Processing,* Asian Federation of Natural Language Processing, Nagoya, Japan, ISBN 978-4-9907348-0-0, pp. 91-99.

Prasad, Rashmi, Dinesh. Nikhil, Lee, Alan et al. (2008). The Penn Discourse Treebank 2.0. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Sgall, Petr, Hajičová, Eva, & Panevová, Jarmila (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Dordrecht: Reidel Publishing Company and Prague: Academia.

PCC 2.0: Annotation for Discourse Research

Arne Neumann & Manfred Stede

University of Potsdam

We present a revised and extended version of the Potsdam Commentary Corpus, a collection of 175 German newspaper commentaries (op-ed pieces) from *Märkische Allgemeine Zeitung*. The corpus is (deliberately) "unbalanced" in two respects:

All texts belong to the same *genre*, because the goal is to build a resource for studying argumentative text in particular.

All texts are from the same newspaper. On the one hand, this is due to licensing issues (only one publisher had to be contacted); on the other hand, the choice of regional daily paper yields a relatively homogeneous selection of relatively simple texts, which supports experiments with automatic text analysis.

All texts have been been annotated with syntax trees and three layers of discourse-level information: nominal coreference, connectives and their arguments (similar to the PDTB, (Prasad et al 2008)), and trees reflecting discourse structure according to Rhetorical Structure Theory (Mann, Thompson 1988).

Syntax trees were produced semi-automatically with the annotate tool (Brants/Plaehn 2000), which suggests a tree to the user, who can then edit and correct it.

Connectives have also been annotated with the help of a semi-automatic tool, Conano (Stede/Heintze 2004), which automatically identifies most connectives and suggests arguments based on their syntactic category. The annotator then has to verify whether the word is in fact used as a connective in the particular context, and if so, whether the arguments are correctly assigned. In contrast to PDTB, this layer does not include any sense relations.

The other two layers have been created manually, also with dedicated annotation tools: MMAX2[i] for coreference, and RSTTool[ii] for rhetorical structure. The corpus is made available on the one hand as a set of original XML files produced with the annotation tools, based on identical tokenization. On the other hand, it will soon be distributed together with the open-source linguistic database ANNIS3 (Chiarcos et al 2008, Zeldes et al. 2009}, which provides multi-layer search functionality and layer-specific visualization modules. This allows for comfortable qualitative evaluation of the correlations between annotation layers.

Thanks to an agreement with the publisher of *Märkische Allgemeine*, the corpus (source texts and annotation files) can be downloaded freely:

http://www.ling.uni-potsdam.de/acl-lab/Forsch/pcc/pcc.html

References

T. Brants, O. Plaehn: Interactive Corpus Annotation. Proc. of the Language Resources and Evaluation Conference (LREC), Athen, 2000

C. Chiarcos, S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz, M. Stede: A flexible framework for integrating annotations from different tools and tagsets. In: *Traitement Automatique des Langues*, 49:217-246, 2008

William C. Mann, Sandra A. Thompson: Rhetorical Structure Theory: Toward a Functional Theory of text Organization. In: *Text* 8 (3), 1988

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

M. Stede, S. Heintze. Machine-assisted Rhetorical Structure Annotation. In: Proc. of the Int'l Conference on Computational Linguistics, COLING-2004, Geneva, 2004.

A. Zeldes, J. Ritz, A. Lüdeling, C. Chiarcos: Annis – a search tool for multi-layer annotated corpora. Proc. of Corpus Linguistics 2009

Acknowledgement: Work on the PCC was partly supported through the project "Linguistic Database" in Sonderforschungsbereich 632 "Information Structure", funded by Deutsche Forschungsgemeinschaft.

The Catalan Discourse GraphBank

Roser Saurí, Teresa Suñol, Toni Badia

Pompeu Fabra University

We present the Catalan Discourse GraphBank, a corpus of Catalan texts (newspaper reports and fiction) marked up for discourse segments and the relations they hold. The annotation scheme is inspired in the Discourse GraphBank corpus for English (Wolf et al. 2005) and is comparable to an equivalent corpus for Spanish (the Spanish Discourse GraphBank) that is currently in preparation.

The Catalan Discourse GraphBank presents two levels of annotation. First, texts are fully segmented into discourse segments, generally at the clause level (along the lines of Asher and Lascarides, 2003; Wolf and Gibson, 2005), but possibly also at smaller linguistic units (e.g., some time- and place-denoting PPs). At the second annotation level, segments are connected through discourse relations following previous work that applied linguistic and reasoning-based criteria (e.g., Hobbs, 1985; Asher and Lascarides, 2003), but also assuming structural considerations, as in Wolf and Gibson (2005). The set of discourse relations in the Catalan Discourse GraphBank splits into two different classes, depending on whether they connect segments through a head-satellite kind of relation or at an equal level of dependency. The first group includes Temporal Sequence, Elaboration, Explanation, Example, Generalization, Violated Expectation, Attribution Condition, Purpose and Background, while the second one contemplates the relations of Parallel, Same, Contrast and Joined. Discourse relations can be established between basic segments, but can also associate pairs or even clusters of segments already connected. Moreover, relations can cross, thus resulting into graph (instead of tree) structures, as in Wolf and Gibson (2005).

The corpus contains a total of 127 texts (48,410 tokens), obtained form the Catalan version of the AnCora corpus (Taulé et al. 2008), which therefore are also marked up with basic linguistic information, e.g., lemma, part of speech, constituent structure, dependency relations, etc. Furthermore, the texts in the Catalan Discourse GraphBank have also been annotated with time and event information as part of the Catalan TimeBank corpus (Saurí and Badia, 2012).

References

Asher, N., A. Lascarides (2003) *Logics of Conversation*. Cambridge University Press, Cambridge. Hobbs, J. (1985) On the coherence and structure of discourse. Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University, CA.

Saurí, R., T. Badia (2012) *Catalan TimeBank 1.0 LDC2012T10*. Philadelphia: Linguistic Data Consortium. URL: https://catalog.ldc.upenn.edu/LDC2012T10

Taulé, M., M.A. Martí, M. Recasens (2008). *AnCora: Multilevel annotated corpora for Catalan and Spanish*. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). Marrakech, Morocco, 2008.

Wolf, F., E. Gibson (2005) *Representing Discourse Coherence: A Corpus-based Study*. Computational Linguistics 31, 249-287.

Wolf, F., E. Gibson, A. Fisher, M. Knight (2005) *Discourse Graphbank*. LDC2005T08. Philadelphia: Linguistic Data Consortium. URL: https://catalog.ldc.upenn.edu/LDC2005T08

DiMLex: A lexicon of German connectives

Tatjana Scheffler & Manfred Stede

University of Potsdam

An immensely valuable resource for information on German connectives is the Handbook by Pasch et al. (2003). It defines quite clear criteria for defining the notion of connective, and provides a wealth of information (predominantly syntactic) on roughly 350 words. Using the definition from Pasch et al, the machine-readable lexicon `DiMLex' provides structured information (in XML) for a subset of connectives that occur with high frequency in written German (newspaper text). The first version, described in (Stede/Umbach 1998), had 150 entries, whereas the current version has 189. The growth resulted from a joint research project with the 'Handbook' team at IDS Mannheim, where causal connectives have been studied in great detail, both from a corpus-linguistic and a computational perspective.

In its present form, DiMLex provides information on the orthographic variants of a connective; on ambiguity with respect to non-connective readings (many words also have readings as discourse particles or pronominal adverbs); on the combinability with focus particles or correlates (items that yield a double signalling of the coherence relation); on the part-of-speech tags that common German taggers assign to the word; on syntax (position in the clause or sentence, etc., following the description scheme of Pasch et al.); on semantic senses and the linking between syntactic and semantic arguments. For illustration, here is an abridged version of the entry for the connective *nämlich,* which exhibits two possible orthographies and is ambiguous between two senses:

```
<entry id="k110">
   <orths>
     <orth type="cont" canonical="1" onr="k110o1">
       <part type="single">n&#228;mlich</part>
     </orth>
     <orth type="cont" canonical="0" onr="k110o2">
       <part type="single">naemlich</part>
     </orth>
   </orths>
   <desambi>
     <conn_d>1</conn_d>
     <sem_d>1</sem_d>
   </desambi>
   <focuspart>0</focuspart>
   <correlate>
     <is_correlate>0</is_correlate>
     <has_correlate>0</has_correlate>
   </correlate>

   <syn>
     <cat>konnadv</cat>
     <integr>
       <vorfeld>0</vorfeld>
       <mittelfeld>1</mittelfeld>
       <nacherst>1</nacherst>
       <nachfeld>0</nachfeld>
       <nullstelle>1</nullstelle>
       <nachnachfeld>0</nachnachfeld>
       <satzklammer>0</satzklammer>
     </integr>
     <ordering>
       <ante>0</ante>
       <post>1</post>
```

```
      <insert>0</insert>
    </ordering>
    <sem>
      <coherence_relations>
        <relation>cause</relation>
      </coherence_relations>
      <role_linking>
        <int>antecedent</int>
        <ext>consequent</ext>
      </role_linking>
    </sem>

    <sem>
      <coherence_relations>
        <relation>elaboration</relation>
      </coherence_relations>
      <role_linking>
        <coordinating/>
      </role_linking>
    </sem>

  </syn>
</entry>
```

One complication results from the fact that connectives can consist of multiple words, which need not be adjacent. We provided an analysis of this problem (suggesting a classification) in (Stede/Irsig 2008), and the base information on multiple parts and the permissible variants of linear order, are also represented in the lexicon.

The XML format was designed to be independent of a particular target application; DiMLex has been integrated into modules for both language analysis (RST-style discourse parsing) and text generation (choice of connective for a given relation).

References

Pasch, Renate/Brauße, Ursula/Breindl, Eva/Waßner, Ulrich Hermann: Handbuch der deutschen Konnektoren. Berlin/New York: de Gruyter, 2003

M. Stede, K. Irsig. *Complex connectives in German: Complications for local coherence analysis*. In: A. Benz, M. Stede, P. Kühnlein: Constraints in Discourse 3 - Representing and Inferring Discourse Structure. Amsterdam: John Benjamins, 2012

M. Stede, C. Umbach. DiMLex: A lexicon of discourse markers for text generation and understanding. In: Proceedings of COLING-ACL '98, Montréal, August 1998

Turkish Discourse Bank (TDB)

Deniz Zeyrek, Ruket Çakıcı, Ayışığı B. Sevdik-Çallı, Işın Demirşahin

Middle East Technical University, Ankara

We describe TDB (Zeyrek et al., 2013), a ~400K-word corpus annotated for discourse relations in the PDTB style (Prasad et al., 2007). As in English, we identify discourse connectives from subordinators, coordinators and discourse adverbials and mainly annotate the connective

together with its two arguments (Arg1, Arg2), supplements to the arguments (Supp1, Supp2) and modifiers of a connective. On the basis of 77 search tokens, 143 discourse connective tokens were identified and annotated with 537 modifiers, 872 Supp1, 342 Supp2, 1176 shared text spans, and 92 supplements to shared text spans, amounting to a total of 8483 annotations. TDB 1.0 is freely available to researchers at: http://medid.ii.metu.edu.tr/. This poster concentrates on the major differences of TDB from PDTB.

The Shared Tag: Turkish is a morphologically rich, variable word order language. Although morphology can be a cue for marking the arguments to a connective, the variable word order is a potential difficulty. The Shared tag (Ex.1) identifies a subject/object or a temporal/locative adjunct shared by the arguments and assists the annotators in easily distinguishing the arguments. In the examples, italics show Arg1, bold indicates Arg2. The discourse connective is underlined.

Kaptandı ama **yüzme bilmezdi** {amcam}.

'(He) *was a captain* but **did not know how to swim**, {my uncle}.'

Marking the shared information is also a potential aid in future NLP applications.

Intra-sentential connectives: In Turkish, suffixes can act as subordinators and commonly function as intra-sentential discourse connectives. Subordinators have the simplex and complex sub-types. While simplex subordinators are merely suffixes on the subordinate verb, complex subordinators involve two parts, viz. a postposition and a suffix on the subordinate verb, e.g. *-na rağmen* 'despite' (Ex.2). The text span where the second part of the postposition appears is necessarily the Arg2 to the connective. This is the sense in which morphology assists in determining the Arg2.

(2) **Gerçeği bilmesi-**ne rağmen *sustu*. 'Despite **knowing the truth**, *she kept quiet*.'

Only complex subordinators are marked in TDB 1.0; work on simplex subordinators has recently started (Acar et al., 2015).

Phrasal expressions: Turkish derives new connective devices on the basis of complex subordinators and deictic pronouns, e.g. *bu-na rağmen* 'despite this', which are categorised as a sub-type of discourse connectives in one of the major grammars of Turkish (Göksel & Kerslake, 2005). They are easily retrieved while searching for the related subordinator connective in the annotation tool (Aktaş, et al. 2010); due to their highly productive nature in the language, they are marked as explicit discourse connectives in TDB while in PDTB, similar connectives are marked as alternative lexicalization devices (Prasad et al. 2010).

The Modifier Tag identifies the modifier of a connective as well as the modifier of the discourse relation (Ex.3). Both modifier types are annotated with the same tag; further research will distinguish the different types and will possibly aid automatic discourse parsing.

/Belki/ **ona karşı çok iyi olduğum** için *bıraktı beni*.

'/Perhaps/ *he left me* because **I treated him too well**.'

We have just started to work on implicit connectives; annotation of senses and attribution is left for further work.

References

Acar, Faruk, Zeyrek, Deniz, Çakıcı, Ruket (2015). *Detecting simplex subordinators in Turkish. Submitted to First Action Conference, Textlink,* January 26-28, 2015, Louvain la Neuve, Belgium.

Aktaş, B., Bozsahin, C., & Zeyrek, D. (2010). Discourse relation configurations in Turkish and an annotation environment. In *Proceedings of the Fourth Linguistic Annotation Workshop, ACL2010* (pp. 202-206),Uppsala, Sweden.

Göksel, A., & Kerslake, C. (2005). *Turkish: A comprehensive grammar*. Psychology Press.

Prasad, Rashmi, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi and Bonnie Webber. 2007. *The Penn Discourse TreeBank 2.0 Annotation Manual*. December 17, 2007. http://www.seas.upenn.edu/~pdtb (retrieved May 29, 2010).

Prasad, R., Joshi, A., & Webber, B. (2010). Realization of discourse relations by other means: alternative lexicalizations. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING2010): Posters* (pp. 1023-1031). Beijing, China.

Zeyrek, Deniz, Demirşahin, Işın, Sevdik Çallı Ayışığı B. and Çakıcı, Ruket. 2013. Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue and Discourse,* 4(2), 2013. pp. 174-184.

Towards the construction of a decision tree for the functional disambiguation of Hungarian DSDs

Ágnes Abuczki[1] & Péter Furkó[2]

1MTA-DE Research Group for Theoretical Linguistics, 2Károli Gáspár University of the Reformed Church in Hungary

The present study aims at identifying the sequential and nonverbal features that typically characterize and best distinguish the different DS uses of Hungarian *mondjuk (~'say')* and *ugye (~'is that so?'*). It is argued in this study that a multimodal approach is indispensable in communication modelling in order to disambiguate the actual meaning of polysemous communicative signals such as oral DSDs. The material of the study is comprised of 6 hours of spontaneous conversation (11 simulated job interviews and 11 informal conversations) from the Hungarian HuComTech corpus (Hunyadi & al. 2011) with a constant agent and 11 different young speakers (university students between 18-25 years of age). The HuComTech corpus is annotated at multiple levels in Praat (Boersma & Weenink 2007) for the audio material and in ELAN 4.5.1 (Brugman & Russel 2004) for the video material. At the discourse level of its annotation, the transcribed dialogue is segmented into dialogue turns. The video annotation of the corpus involves the labelling of facial expressions, gaze directions, eyebrow positions, head movements, handshape types, hand movements, postures, deictic gestures and emblems. The corpus contains 208 tokens of *mondjuk (~'say')* and 70 tokens of *ugye (~'is that so?)*. The following two most salient functional categories of each of the two DSDs will be analyzed and distinguished: (1a) lexical search/approximation (as own speech management functions) versus (1b) contrast/concession (as discourse-pragmatic relations between two segments) expressed by *mondjuk (~'say')*; and (2a) checking information and asking for reassurance as directive acts versus (2b) explanation as a constative act marked by *ugye (~'is that so?'*). After importing and merging audio annotations (Praat TextGrids) into ELAN, the selected DSDs were segmented and functionally indexed. The corpus queries (e.g. *Find overlapping labels, N-gram within annotations*) in ELAN address the analyses of their contextual environment (lexical co-occurrences, presence or absence of surrounding silence), position in the utterance, prosodic features (duration, mean F0, direction of pitch movement) and nonverbal-visual markers (the presence or absence of co-verbal hand movements, gaze direction and facial expression of the speaker). The results of multimodal corpus queries and the statistical tests (*Pearson's chi-square test, Crosstabs test, Fischer's exact test, independent samples t-test, paired t-test, box plot graphs*) suggest that the defining properties distinguishing different functions are the duration of the DSD and the simultaneous performance or cessation of manual gesticulation in both DSDs. The findings of the multimodal queries will be modelled using two decision trees. In the case of the different functions of *ugye (~'is that so?)*, gaze direction is also a distinguishing feature, while in the case of *mondjuk (~'say')*, the facial expression of the speaker also helps to disambiguate the actual function of the DSD. Position has also been found to influence both the actual function and the direction of pitch movement in the DSD and its host unit. In contrast, no relationship has been found either between preceding silence and the function of a DSD or between the mean F0 and the function of a DSD.

References

Boersma, P. & Weenink, D. (2007) *Praat: doing phonetics by computer 5.0.02.* University of Amsterdam: Institute of Phonetic Sciences, http://www.praat.org

Brugman, H. & Russel, A. (2004) Annotating multi-media/ multi-modal resources with elan. In: Lino, M., Xavier, M., Ferreire, F., Costa, R., Silva, R. (Eds.) *Proceedings of the Fourth International Conference on Language*

Resources and Evaluation (LREC). Lisbon: Portugal. 2065–2068. (ELAN 4.5.1 software was used in the study. It can be downloaded at http://tla.mpi.nl/tools/tla-tools/elan/download/).

Hunyadi, L, Bertok, K, Nemeth T, E, Szekrenyes, I, Abuczki, A, Nagy, G, Nagy, N, Nemeti, P. & Bodog, A. (2011) The outlines of a theory and technology of human-computer interaction as represented in the model of the HuComTech project. In: *2nd International Conference on Cognitive Infocommunications (CogInfoCom)*. Budapest, 7-9 July, 2011. Budapest: IEEE. http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?reload=true&arnumber=5999491

FDTB1: the first step in annotating a French corpus for discourse

Laurence Danlos, Margot Colinet, Jacques Steinlin

Université Paris Diderot

Our aim is to develop the French Discourse Tree Bank (FDTB) with a discourse layer on top of the syntactic one which is available in the French Tree Bank (FTB, (Abeillé et al., 2003)) on a journalistic corpus (*Le Monde*). The discourse annotation will be in the vein of the Penn Discourse Tree Bank (PDTB, (Prasad et al., 2008)), however with a different workflow. In the first step, which we called FDTB1 and have just achieved, we have identified all the words or phrases in the corpus that are used as discourse connectives. The methodology was the following: first, we highlighted all the items in the corpus that are recorded in LexConn (Roze et al., 2012), a lexicon of French connectives with 350 items—164 adverbials, 117 subordinating conjunctions, 7 coordinating conjunctions, and 62 prepositions which introduce infinitival clauses —, next we eliminated some of these items with the following criteria:

first, we filtered out the LexConn items that are annotated in FTB with parts of speech incompatible with a connective use, e.g. bref annotated as Adj instead of Adv , en fait annotated as Pro V instead of (compound) Adv ;

second, as we lay down for theoretical and pratical reasons that elementary arguments of connectives must be clauses or VPs, we filtered out e.g. LexConn prepositions that introduce NPs;

last, we filtered out LexConn prepositions and adverbials with a non-discursive function (subordinating and coordinating conjunctions which introduce clauses or VPs are always considered as discourse connectives).

The last criterion requires a manual work contrarily to the two others. Five prepositions, e.g. pour (to), are ambiguous between a connective use (*Fred s'est dépeché pour etre à la gare à 17h (Fred hurried to be at the station at 17h)*) and a preposition introducing a complement (*Fred s'est dépeché pour aller à la gare* (*Fred hurried to go to the station*)), and the disambiguation between the two uses is subtle (Colinet et al., 2014a); see also (Huddleston and Pullum, 2002, p. 1223) for English *to*. 54% adverbials are ambiguous between a discourse connective use and a sentential semantic modifier use. Apart from the general criteria used to determine which adverbials should be part of LexConn (e.g. no compositionality in compound adverbials), we did not find other

criteria to perform adverbial disambiguation between discourse and nondiscourse uses. Instead, the FDTB1 annotation manual (Colinet et al., 2014b) describes on an individual basis for any ambiguous adverbial its modifier and connective uses.

The FDTB corpus contains 1005 articles, 18 535 sentences and about 500 000 words. FDTB1 identifies 9 833 explicit connectives (3200 adverbials, 1925 subordinating conjunctions, 3649 coordinating conjunctions and 1059 prepositions)[2]. A similar annotation enterprise is currently on its way for a (smaller) corpus in other genres (e.g. French Wikipedia) and the numerical results will be available at the beginning of 2015 along with the inter annotator agreement. These annotated corpora will be freely available.

We also plan to discuss the pros and cons of the workflow in the FDTB, i.e. starting with the identification of all the discourse connectives in the corpus.

References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for french. In Abeillé, A., editor, Treebanks. Kluwer Academic Publishers, Dordretch

Colinet, M., Danlos, L., Dargnat, M., and Winterstein, G. (2014a). Emplois de la préposition pour suivie d'une infinitive : description, crit`eres formels et annotation en corpus. In Actes du Congr`es Mondial de Linguistique Francaise (CMLF, 2014) Berlin, Allemagne

Colinet, M., Danlos, L., and Steinlin, J. (2014b). Guide d'annotation du FDTB1. Rapport technique Alpage

Huddleston, R. and Pullum, G. (2002). The Cambridge Grammar of the English Language. Cambridge University Press, Cambridge

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrackech, Maroc

Roze, C., Danlos, L., and Muller, P. (2012). Lexconn: a French lexicon of discourse connectives. Revue Discours [Online] URL : http://discours.revues.org/8645

Annotating implicit coherence relations in parallel corpora

Jet Hoek[1] & Sandrine Zufferey[2]

1Universiteit Utrecht, 2Université de Fribourg

Annotating coherence relations is a difficult task that requires detailed annotation schemes and well-trained annotators. Existing discourse-annotated corpora such as the Penn Discourse Treebank, RST Treebank, and the TüBa-D/Z corpus all have different annotation manuals: not only do these corpora differ in the types of relations they distinguish, but also in their

[2] There are also 529 forms *en V-Ant (by V-ing)*.

segmentation rules and even in their definition of what constitutes a coherence relation. Another problem is caused by the fact that coherence relations can, but need not, be made linguistically explicit by means of connectives (*because*, *if*) or cue phrases (*as a result*, *despite the fact that*). The absence of a connective seems to introduce additional complications to the annotation process. Implicit coherence relations leave annotators with less evidence pointing toward a particular relation and the locating of a coherence relation becomes in itself a potential source of disagreement. In the different discourse-annotated corpora there is even less consensus on how to locate and annotate implicit relations than explicit relations. Although the annotation schemes used in the available discourse-annotated corpora are not identical, similar patterns emerge as to which relations can often be left implicit and which relations are usually expressed with a connective. Temporally backward and negative relations, for instance, are usually explicitly signaled, and it appears to be even rarer to implicitly convey conditional relations.

In this presentation we argue that parallel corpora are useful tools for locating, annotating, and researching the characteristics of implicit coherence relations. We used directional corpora extracted from the Europarl corpus (Koehn 2005; Cartoni, Zufferey & Meyer 2013a) and manually spotted cases of implicit translations using the translation spotting method (Cartoni, Zufferey, & Meyer 2013b) across four target languages (French, German, Dutch and Spanish). We analyzed the discourse relations using a set of basic features based on Sanders, Spooren & Noordman (1992). Our results indicate that the basic features of coherence relations conveyed by connectives helps predict their explicit vs. implicit translation across languages. In particular, we demonstrate that the translation of negative relations, signaled in the source language by *although*, and conditional relations, signaled in the source language by *if*, differs significantly from the translation of positive relations or non-conditional relations when it comes to the level of implicitation, a finding that corresponds to data on the implicitness of discourse relations from available mono-lingual discourse-annotated corpora. Strikingly, this holds for all language pairs, even though some languages appear to be more prone to implicitation in translation than others: in English-Dutch translation, for instance, a lot more connectives are removed than in English-Spanish translation.

The observation that not all coherence relations can be equally well expressed implicitly has received several explanations in the literature. These explanations are related to the assumption that readers or listeners have certain default expectations about the organization of discourse that bias their interpretation (e.g. Asr & Demberg 2012, 2013). Murray (1997) formulated the 'continuity hypothesis,' which supposes that readers expect discourse to unfold in a temporally linear manner and that each new discourse segment will be causally congruent with the preceding context. Under this hypothesis, negative relations do not constitute the default interpretation in a discourse: the discourse segments do not follow logically from each other. Instead, one of the segments functions as a negative counterpart to the other segment (e.g. contrastive cause – consequence). Negative relations are therefore discontinuous and not default. Although the continuity hypothesis can account for the relative explicitness of negative relations, if offers no explanation for the fact that conditional relations are almost always signaled with a connective: conditionals cannot be categorized as either continuous or discontinuous (Asr & Demberg 2012).

We hypothesize that coherence relations that are rarely expressed implicitly are cognitively more complex than the coherence relations that can be easily conveyed without a connective. Specifically in this presentation, we argue that negative relations and conditional relations are cognitively more complex than positive relations and non-conditional relations. Because negative relations and conditional relations do not constitute default interpretations, they will often need to be explicitly marked by means of a connective or cue phrase.

References

Asr, F. & Demberg, V. (2012). Implicitness of discourse relations. In *Proceedings of COLING*. Mumbai, India.

Asr, F. & Demberg, V. (2013). On the information conveyed by discourse markers. *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics* (pp. 84-93). Sofia: Bulgaria.

Cartoni, B., Zufferey, S. & Meyer, T. (2013a). Using the Europarl corpus for linguistic research. *Belgian Journal of Linguistics*, 27, 23-42.

Cartoni B., Zufferey S. & Meyer T. (2013b). "Annotating discourse connectives by looking at their translation: The translation-spotting technique." *Dialogue and Discourse* 4(2), 65–86.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of the 10th Machine Translation Summit* (pp. 79–86). Phuket, Thailand.

Murray, J. (1997). Connectives and narrative text: The role of continuity. *Memory and Cognition*, 25, 227–236.

Sanders, T. Spooren, W. & Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15, 1–35.

RST and its annotation method in the analysis of reflective and argumentative text type

Johanna Komppa & Jyrki Kalliokoski

University of Helsinki

Rhetorical Structure Theory and its annotation method focus on systematic analysis of relations. The definitions of the relations are tested and developed by the RST users, and the annotation tool helps to visualize the relations and the rhetoric spans of the text. Despite of these clear benefits the tree diagram may cause problems if one satellite should be in relation to more than one nucleus (see also Wolf and Gibson 2005, Taboada and Mann 2006).

This poster takes part into the discussion of the suitable framework for annotating DRD´s. It demonstrates some challenges faced when strict limitation of relations (e. g. one relation from one satellite is allowed) and when the RST annotation tool (RST Tool 3.45, O´Donnell 2004) is used in the analysis of non-professional writers writing reflective and argumentative texts.

The data of this study comes from the matriculation examination essays written in Finnish as a second language. The writers of the texts are approx. 17–20 years old students who take part to the examination in the end of the upper secondary school and who have Finnish as L2. The data is comprised of 96 expository essays. Each essay has approximately 281 words. The writers have chosen between two themes, a nice place to meet friends (e. g. coffee house, streets, youth centre) or virtual places to meet friends (e. g. chat, Internet). The data represents reflective and argumentative text type, and the writers have been asked to deal with the chosen theme from different perspectives.

The annotation tool of RST is useful when the target is to illustrate the rhetorical structure of the whole text. The functional parts of the text are visualized with large spans and the spans are based on the analysis of the smallest relations of the text. So every relation has been taken into account.

On the other hand it is found that the annotation tool doesn´t allow to signal the multiple relations one unit might have (see also Komppa 2012). This is evident especially when the writer goes back to the theme presented in a satellite earlier in the essay (e. g. one or two paragraphs ago). Even though the coherence of the text is evident, the RST annotation tool doesn´t allow to signal another relation to the text unit which has been tied to a span already.

Another challenge for the annotation tool is reflective and argumentative text type itself. The text structure and the relations between quite large spans might be difficult to express because of the two-dimensional signaling of the relations is not possible so far.

References

Komppa, Johanna (2012). Retorisen rakenteen teoria suomi toisena kielenä -ylioppilaskokeen kirjoitelman kokonaisrakenteen ja kappalejaon tarkastelussa [Rhetorical structure theory in study of the schematic, rhetorical and paragraph structure of matriculation essays in Finnish as a second language]. PhD thesis, University of Helsinki, Finland. DOI: http://urn.fi/URN:ISBN:978-952-10-8164-4

O´Donnell, Michael 2004: RST-Tool Version 3.45. Annotation tool. http://www.sfu.ca/rst/06tools/index.html

Taboada, Maite, Mann, William C. (2006). *Rhetorical structure theory: looking back and moving ahead.* Discourse Studies 8 (3), s. 423–459.

Wolf, Florian, Gibson, Edward (2005). *Representing discourse coherence: a corpus-based study*. Computational Linguistics 31 (2), s. 249–287.

Finding Nexus in the PDiT and GECCo Annotation Schemes

Ekaterina Lapshinova-Koltunski[1], Anna Nedoluzhko[2], Kerstin Kunz[3], Lucie Poláková[2], Jiří Mírovský[2], Pavlina Jínová[2]

1Saarland University, 2Charles University in Prague, 3University of Heidlberg

In this presentation, we will demonstrate an experiment designed to compare two frameworks for the analysis and annotation of DSDs: the one within the project GECCo (German-English Contrasts in Cohesion) at Saarland University, see Lapshinova & Kunz (2014), and that of the Prague Discourse Treebank (PDiT), see Poláková et al. (2013). The experiment aims at identifying commonalities and/or differences between the two frameworks, with the overarching goal of achieving interoperability and creating an 'all-in-one' scheme which can be applicable to different languages, different genres and registers, including spoken and written dimensions.

Our initial observations have revealed commonalities between both approaches, although the classification of DSDs is based on different frameworks: cohesive relations in GECCo (based on the definition by Halliday & Hasan, 1976) vs. Functional Generative Description (cf. Sgall et al., 1986) and Penn-style discourse annotation (see Prasad et al., 2008) in PDiT. Moreover, annotations in GECCo are applied on comparable and parallel subcorpora of English and German (cf. Hansen-Schirra et al., 2012 and Lapshinova-Koltunski et al. 2012), containing various registers, including written and spoken dimensions, whereas PDiT primarily contains journalistic texts in Czech with further genre classification (see Bejček et al., 2013).

For the sake of convenience, we annotate the same datasets with both annotation schemes. So, we select two different genres – journalistic and fictional texts – and annotate them in accordance with the guidelines of both conceptions. To be able to unify the annotated categories afterwards, we decide to start with English texts only. However, we are planning to work both with German and Czech in the future in order to identify differences between Germanic and Slavic languages in the preferences for explicit and implicit discourse relations. The journalistic samples are texts exported from the Prague English Dependency Treebank (see Cinková et al. 2009), containing around 100 sentences. A sample from fiction of the same size was exported from the written part of GECCo.

The annotation scenarios proceed independently in accordance to the common procedures used in both projects, including automatic pre-annotation and manual annotation. The tools assisting manual annotation are MMAX2 (Müller & Strube, 2006) in GECCo, and TrEd (Pajas & Štěpánek, 2008) in PDiT. They are also used for visualisation of the annotated data in both cases. After the annotation is finished, the resulting double annotations will be unified and compared.

In our poster presentation, we will illustrate both frameworks, show the datasets and the resulting annotated structure, and demonstrate the first comparison results.

References

Bejček Eduard, Hajičová Eva, Hajič Jan, Jínová Pavlína, Kettnerová Václava, Kolářová Veronika, Mikulová Marie, Mírovský Jiří, Nedoluzhko Anna, Panevová Jarmila, Poláková Lucie, Ševčíková Magda, Štěpánek Jan, Zikánová Šárka (2013). Prague Dependency Treebank 3.0. Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, Czech republic, http://ufal.mff.cuni.cz/pdt3.0/, Dec 2013.

Cinková, S., J. Toman, J. Hajič, K. Čermáková, V. Klimeš, L. Mladová, J. Šindlerová, K. Tomšů, and Z. Žabokrtský (2009). Tectogrammatical Annotation of the Wall Street Journal. Prague Bulletin of Mathematical Linguistics, 92.

Halliday, M.A.K. and R. Hasan (1976). Cohesion in English. London: Longman.

Hansen-Schirra, S., S. Neumann, and E. Steiner (2012). Cross-linguistic Corpora for the Study of Translations. Insights from the language pair English – German. Series Text, Translation, Computational Processing. Berlin / New York: Mouton de Gruyter.

Lapshinova-Koltunski, E. and K. Kunz (2014). Annotating Cohesion for Multillingual Analysis. In Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, Reykjavik, May 26, 2014.

Lapshinova-Koltunski, E., K. Kunz, and M. Amoia (2012). Compiling a Multilingual Spoken Corpus. In Tommaso Raso Heliana Mello, Massimo Pettorino, editor, Proceedings of the VIIth GSCP International Conference: Speech and corpora, pages 79–84, Firenze. Firenze University Press.

Müller, C. and M. Strube (2006). Multi-level annotation oflinguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, pages 197-214. Peter Lang, Frankfurt a.M., Germany.

Pajas, P. and J. Štěpánek (2008). Recent Advances in a Feature-Rich Framework for Treebank Annotation. In Proceedings of the 22nd International Conference on Computational Linguistics - Proceedings of the Conference, The Coling 2008 Organizing Committee, Manchester, UK, pp. 673-680.

Poláková, Lucie, Mírovský, Jiří, Nedoluzhko, Anna, Jínová, Pavlína, Zikánová, Šárka, Hajičová, Eva. (2013). Introducing the Prague Discourse Treebank 1.0. In Proceedings of the 6th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, pp. 91-99.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. and Webber, B. (2007). The Penn Discourse TreeBank 2.0 Annotation Manual.

Sgall, P., Hajičová, E. and Panevová, J. (1986). The Meaning of the Sentence in Its Semantic and Pragmatic Aspects, Dordrecht: Reidel Publishing Company, Praha: Academia.

A Translation-based Assessment of PDTB Explicit Connectives in Romanian

Sorina POSTOLEA

"Alexandru Ioan Cuza" University of Iași

This paper intends to explore to what extent the categories of explicit connectives and their senses described in the PDTB Annotation Manual may be used cross-linguistically as a basis to annotate matching categories of Romanian discourse-relational devices (DRDs).

Starting from the premise that the translation process involves the interlingual transference and reconstruction of the source-text discourse relations through the use of equivalent DRDs in the target-language text, the paper analyses the various translations given to the most productive PDTB explicit connectives in a fully aligned, ~550,000-word, English-Romanian parallel corpus built by the author.

Unlike bilingual dictionaries, which provide a relatively small repertory of word- and/or phrase-rank equivalences for a given source connective, the corpus-based analysis of actual translated material may reveal alternative realizations of its underlying discourse meaning/relation in the target language. To give but an example, in the corpus used for this paper six different Romanian connectives are used to translate the English conjunction *because*: *deoarece*, *datorită*, *pentru că*, *din cauza*, *fiindcă*, *întrucât*. Some of these Romanian connectives are sometimes interlaced with other parts of speech, which are not explicit in the English source-sentence, such as nouns – *datorită faptului că* [because of the fact that] – or demonstrative adjectives – *din această cauză* [for this cause/reason]. It is unclear whether these expanded forms may be seen as variants or as alternative lexicalizations according to the PDTB annotation scheme.

Thus, the analysis of the various Romanian translations/equivalents used in our corpus for the PDTB explicit connectives may serve as a basis to infer and analyze DRDs and annotation criteria which were not included in the PDTB for English but may nevertheless prove to be relevant for the Romanian language in a multilingual annotation scheme.

On Definition of Discourse Connectives – Primary vs. Secondary Connectives (Based on a Corpus Probe)

Magdaléna Rysová & Kateřina Rysová

Charles University in Prague

The aim of our presentation is to contribute to the general discussion on discourse connectives, especially on their definition and principles we may hold as boundaries surrounding this class of expressions.

Our theoretic conclusions are based on a practical analysis of large corpus data, i.e. on approximately 50 thousand of Czech sentences from the Prague Dependency Treebank (Bejček et al., 2012), but we think that our statements may be used also for other languages.

The issue of defining discourse connectives in Czech arose mainly during the annotation of multiword expressions like *that is the reason why, the only condition was, this means* etc. On the one hand, these expressions (sometimes called AltLexes – see Prasad et al., 2010) clearly signal discourse relations, on the other hand, they do not belong to the parts of speech generally accepted for connectives (like conjunctions, some types of particles etc.). The problem with these expressions is that they may be inflected (*from this reason – from these reasons*) and may occur in many different forms in the text (cf. *due to this*[3]*, due to this fact, due to this situation* etc.). In this respect, they highly differ from one-word, lexically frozen connectives. But still they function as

[3] We understand the whole structure due to this as a secondary connective, as **due to* itself is an ungrammatical structure and needs to combine with an anaphoric expression to gain a discourse connecting function. At the same time, there are some present-day primary connectives that historically arose from similar combination of a preposition and demonstrative pronoun (e.g. Czech connective *proto* "therefore" from the preposition *pro* "for" and pronoun *to* "this").

indicators of discourse relations – e.g. the expression *that is the reason why* clearly signals a discourse relation of reason and result.

On the basis of practical annotations of authentic Czech texts, we came to several conclusions. Firstly, we define discourse connectives according to two principles: 1. Very generally, according to their function in the text – discourse connectives serve as indicators of discourse relations within the text; 2. Concerning their semantic nature, according to the **universality principle** – **the status of discourse connective must be universal** (cf. the universal connective *this is the reason why* vs. nonuniversal connecting phrase *this increase is the reason why*).

The universality principle evaluates connective structures from the fact whether they have a universal status of connectives, i.e. whether they function as indicators of certain discourse relation universally or occasionally. In other words, we tried to answer – if we have several different contexts with, e.g., the relation of reason and result – whether the given connective structure (with an ability to express this type of relation) fits into each of them (and is therefore universal) or not, see Examples (1), (2) and (3)[4].

(1) *The economy grew. Therefore / Because of this / From this reason / Because of this increase the unemployment dropped.*

(2) *I am ill. Therefore / Because of this / From this reason / *Because of this increase I cannot go to school.*

(3) *I don't like sweets. Therefore / Because of this / From this reason / *Because of this increase I am slim.*

Secondly, within the category of discourse connectives, we define two subclasses according to their lexico-syntactic nature (see Rysová and Rysová, 2014) – **primary connectives** (*like and, but, or, then, therefore* etc.) and **secondary connectives** (*like the result is, the main reason was, this means, because of this* etc.). The differences between primary and secondary connectives are captured in Table 1.

Primary connectives are such expressions whose primary function is to connect two units of a text (they mostly belong to conjunctions and structuring particles). Primary connectives are synsemantic (i.e. grammatical or functional) words and they do not have a role of sentence elements so they do not affect the sentence syntax. Primary connectives are mostly one-word, lexically frozen expressions. The main difference between primary and secondary connectives lies in grammaticalization – i.e. primary connectives are grammaticalized expressions[5] (that arose from other parts of speech and very often from combination of several words[6]) whereas secondary are not. Examples of primary connectives are *but, and, too, only, because, while, or* etc.

Secondary connectives are mainly multiword structures functioning as connectives only in certain collocations. Most of them have a lexical core or key word signaling given type of discourse relation (the cores may be nouns like *condition, reason, difference* etc., verbs like *to mean, to explain, to cause* etc., secondary prepositions like *due to, because of, despite* etc.). Secondary connectives contain (in contrast to primary) some autosemantic (i.e. lexical) word or

[4] We consider the universal structures *therefore / because of this / from this reason* **discourse connectives** (the non-universal phrase *because of this increase* is not the discourse connective despite the fact that it expresses some kind of discourse relation).

[5] Sometimes the grammaticalization is not fully completed, which causes discrepancy among certain parts of speech (especially among conjunctions, adverbs and particles).

[6] E.g. *Because* arose from *bi cause* "by cause", originally a phrase often followed by a subordinate clause, as one word probably from around 1400.

words and have a role of sentence elements or sentence modifiers. Secondary connectives are not grammaticalized, although they exhibit several features typical for the process of grammaticalization (e.g. weakening of singular and plural distinction, gradual loosing of the individual lexical meaning and gaining the primary connecting function as a whole structure etc.). Examples of secondary connectives *are the condition is, this means, this is the reason why, because of this, from these reasons* etc.

To conclude, we distinguish two categories within discourse connectives – (universal) primary connectives (e.g. *therefore, but, and*) and (universal) secondary connectives (e.g. *from this reason*) (as in Rysová and Rysová, 2014).

Table 1

Primary connectives	**Secondary connectives**
synsemantics	structures with autosemantic basis
lexically frozen (grammaticalized)	open or fixed collocations (non-grammaticalized)
non-modifiable (with exceptions)	modifiable (with exceptions)
mainly one-word	mainly multiword
universal	universal
not sentence elements	sentence elements, clause modifiers or separate sentences
	convey anaphoric reference to the 1st argument
	uniqueness of some structures: a) syntactically higher than the 2nd argument b) form of a separate sentence c) nominalization of the 2nd argument

Table 1: Differences between primary and secondary connectives

References

Bejč ek E. et al. (2012). *Prague Dependency Treebank 2.5 -- a revisited version of PDT 2.0*. In: Proceedings of COLING 2012 , Mumbai, India, pp. 231–246.

Prasad, R. et al. (2010). Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In: Coling 2010: Posters , pp. 1023–1031.

Rysová, M. and Rysová, K. (2014). *The Centre and Periphery of Discourse Connectives*. In: Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation. http://www.aclweb.org/anthology/ [to appear]

Multi-layer discourse annotation in the Potsdam Commentary Corpus

Manfred Stede

University of Potsdam

The publicized portion of the Potsdam Commentary Corpus consists of 175 newspaper commentaries annotated with syntax trees, nominal coreference, connectives and their arguments, and rhetorical structure. In addition to this, the non-public portion of the corpus has annotation layers that have so far been developed in smaller experiments only, i.e. not yet on a large selection of texts. Also, we added texts from other sources (Pro & Contra commentaries from *Tagesspiegel* and user-generated reviews from an online forum). We report here on the set of annotation layers, each of which has been thoroughly described by annotation guidelines (so far, only in German). Afterwards we mention a few examples of applications that make use of the data.

Technically, the different layers are being annotated independently (some exceptions will be mentioned below) and with dedicated annotation tools that support the particular annotation style. The resulting data can then be combined with other annotations for specific empirical studies.

(i) Layers that do not pertain to discourse directly, but can be used to support discourse-oriented research

(a) Syntax: We use syntax trees in accordance with the TIGER scheme (Brants et al 2002), which has been designed as a relatively theory-neutral account of syntactic structure.

(b) Negation: Our scheme includes negation operators (certain determiners, adverbs, verbs, nouns) and their scope (Gros/Stede 2013)

(c) Nominal referring expressions: We identify the noun phrases that are interpreted as *referring* to some abstract or physical entity

(ii) Discourse annotations made locally and involving relatively little subjective interpretation

(a) EDUs: We identify spans of text playing the role of an elementary discourse unit. This is implemented as a two-stage annotation that first identifies types of syntactic spans (various kinds of clauses, fragments, parentheticals) and then selects some as EDUs.

(b) Nominal coreference: The units found in (i)(c) are taken as the basis for establishing coreferene links. We annotate only referential identity (i.e., no "bridging").

(c) Information status: For specific studies on information structure, the (i)(c) units are assigned the tags *given, new, accessible*, plus some sub-categories.

(d) Aboutness topic: As a second aspect of information structure, we assign *topics* to certain kinds of EDUs (cf. (ii)(a))

(e) Connectives: Connectives and scopes are handled similarly to the PDTB scheme (Prasad et al. 2008).

(f) Illocutions: We are experimenting with the assignment of illocutionary roles to EDUs (which so far has usually been restricted to dialogue).

(iii) Discourse annotations covering the complete text, involving considerable subjective interpretation

(a) Content zones: A text is segmented into "content zones" that identify portions of fulfilling a particular, genre-specific, function for the text as a whole.

(b) Rhetorical Structure: This annotation uses a slightly adapted version of RST. (c) Argumentation structure: Relations between claims, supporting arguments, rebuttals are being captured in an annotation scheme outlined in (Peldszus/Stede 2013).

Some computational applications

We mention some cases of making use of the data, focusing on the role of connectives. First of all, the sheer presence of connectives (notice that the annotation serves as disambiguation from other kinds of particle readings) in these texts can be exploited for building automatic classifiers that distinguish opinionated/argumentative text (commentary) from more "objective" text. At present we are building such a system for differentiating news stories and opinion pieces in newspapers.

In (Stede/Peldszus 2012) we use the joint annotation of connectives and illocutions to show that certain connectives tend to co-occur with particular types of illocutions, which indicates that the connectives differ in terms of their affiliation with "objective" and "subjective" utterance situations.

We developed a system that uses our connective lexicon DiMLex for automatically disambiguating connectives in text and computing their scope (Küssner/Stede 2011).

In a recent experiment, Scheffler/Stede (submitted) devised an algorithm for computing the correlation between the RST trees and the connective annotation, so that statistics for the mapping can be obtained, and the "signalling behaviour" of connectives can be studied in detail.

References

S. Brants, S. Dipper, P. Eisenberg, S. Hansen, E. König, W. Lezius, C. Rohrer, G. Smith, H. Uszkoreit, Hans (2004): TIGER: Linguistic Interpretation of a German Corpus. Journal of Language and Computation 2:597-620.

O. Gros, M. Stede: Determining negation scope in German and English medical diagnoses. In: M. Taboada, R. Trnavac: *Nonveridicality and Evaluation: Theoretical, Computational and Corpus Approaches.* Leiden/Boston: Brill, 2013

O. Krasavina, C. Chiarcos, D. Zalmanov, D. (2007): Aspects of topicality in the use of demonstrative expressions in German, English and Russian. Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), Lagos/Portugal.

W. Mann, S. Thompson: Rhetorical Structure Theory: Toward a Functional Theory of text Organization. In: *Text* 8 (3), 1988

A. Peldszus, M. Stede: From argument diagrams to argumentation mining in texts: a survey Int'l Journal of Cognitive Informatics and Natural Intelligence (IJCINI) 7(1):1-31, 2013

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, B. Webber: The Penn Discourse Treebank 2.0. Proceedings of LREC, Marrackech, 2008

T. Scheffler, M. Stede: Mapping PDTB-style connective annotation to RST-style discourse structure annotation. Submitted.

M. Stede, U. Kuessner: Kausale Konnektoren in der Automatischen Textanalyse. In: M. Konopka, J. Kubczak, C. Mair, F. Sticha, U. Waßner (Hrsg.): Grammatik und Korpora 2009. Tübingen: Narr, 2011.

M. Stede, A. Peldszus: The role of illocutionary status in the usage conditions of causal connectives and in coherence relations . In: *Journal of Pragmatics*, 44(2):214-229, 2012

Revising the PDTB Sense Annotation Scheme

Bonnie Webber[1], Rashmi Prasad[2], Alan Lee[3], Aravind Joshi[4]

1University of Edinburgh, 2University of Wisconsin (Milwaukee), 3 LexiconTree, 4 University of Pennsylvania

Released to the public in 2008, the Penn Discourse Treebank (PDTB 2.0) remains the world's largest manually annotated corpus of discourse relations (over 40K tokens annotated over 1m words). The PDTB 2.0 annotates for sense, argument span and attribution, discourse relations that are either lexically- grounded in explicit discourse connectives or associated with sentence adjacency. This low-level focus has encouraged not only the use of the PDTB 2.0 in language technology and psycholinguistics, but also annotation of comparable corpora in other languages and genres. A survey and analysis of this work can be found in (Prasad et al., 2014).

Nevertheless, six years of public use of the PDTB 2.0 has shown there to be value in filling in gaps in the annotation and correcting inconsistencies. A recent grant from the U.S. National Science Foundation (NSF) will allow us to develop an enriched and more consistent PDTB 3.0.

Here, we focus on revisions to the sense annotation scheme that will be used in the PDTB 3.0 and what has motivated them. Since the revised scheme will be used not only in annotating additional aspects of the corpus, but also in re-annotating some existing annotation, we will be dealing with standard issues of mapping one annotation scheme to another.

The new scheme retains a three-level structure (class, type and sub-type) and continues to refer to Arg2 as the argument that (syntactically) includes the connective (for explicit relations) or that follows Arg1 for implicit inter-sentential relations. What has changed includes:

- Eliminating fine-grained senses that turned out to be used infrequently and hard for the annotators to apply. This includes all sub-types of Condition and Contrast, as well as Pragmatic Contrast and Pragmatic Concession.
- Restricting sub-types solely to differences in directionality — i.e., whether the sense holds from Arg1 to Arg2, or from Arg2 to Arg1.
- Eliminating sense types that turned out to be un-helpful in back-off, while retaining their sub-types. This includes Restatement and Alternative.
- Including senses that others have needed to annotate corpora in the style of the PDTB 2.0. This includes Purpose (with sub-types Goal and Enablement), Negative Condition (the complement of Condition), Similarity (the complement of Contrast), and Manner.
- Adding directional sub-types needed for annotating intra-sentential discourse relations or to correct inconsistent annotation. This includes sub-types added to Exception and Substitution (the new name for Chosen Alternative).
- Eliminating sense types that differ only in whether their arguments involve an implicit belief or speech act (i.e., Pragmatic Cause and Pragmatic Condition). Instead, a feature will be used to distinguish semantic (the default), epistemic (involving belief) and speech act forms of Cause and Condition.

As of this abstract, we are about to start using this revised annotation scheme. To bring existing annotation in line will involve a combination of simple automatic modification, more complex automatic modification requiring manual review, and manual re-annotation of a well-defined subset of tokens.

References

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. *Reflections on the Penn discourse treebank, comparable corpora and complementary annotation.* Computational Linguistics, 2014. doi: 10.1162/ COLI a 00204.

DRDs for Multilingual Argumentation Analysis

Adam Wyner

University of Aberdeen

Argumentation is a proper subtopic of discourse analysis, where arguments are (minimally) taken to be a pair of statements that are in a contrast relation or a structured passage of statements with discourse relations that associate the statements as premises, exceptions, and a claim. There are other aspects of discourse relevant to argumentation beyond these relations. Argumentation is a particularly interesting and useful subtopic to focus research attention, for there have been recent significant advances in (formal and implemented) computational models of argumentation to reason with inconsistency as well as in analysis of argumentative text to identify, query, and extract arguments from text into the models. Yet to make deeper, more substantive progress, large, richly annotated corpora are needed, especially corpora of arguments across languages, e.g. to better understand and align public arguments on common multilingual EU policy issues such as immigration and culture. In turn, such multilingual corpora depend on having a common, agreed upon, appropriate set of explicit and implicit DRDs, definitions of DRDs, and criteria for identifying them in text. In this work, we consider a range of current proposed DRD annotation frameworks (PDTB, SDRT, RST, and Pragma-Dialectics) and apply them to a bilingual corpus of parallel texts using the GATE tool. We assess the strengths and weaknesses of each framework and outline a way forward to an adequate framework for multilingual annotation of DRDs in argumentative text. In the course of the work, we will explore what DRDs are particularly relevant to argumentation as well as the extent to which discourse relations indicate an extended view on what 'counts as' an argument.

A distributional account of discourse connectives and its effect on fine-grained inferences

Fatemeh Torabi Asr & Vera Demberg

Saarland University

There is a consensus among linguists that *but* is applicable to a wide range of adversative discourse relations, whereas other connectives such as *although* have a narrower usage. Dealing with this property of *but*, Fraser (1999) assigns a core meaning to it that is *simple contrast*; the rest is not encoded in the connective and is obtained from the context. Blakemore (2004) argues against this idea by exemplifying contrastive relations where *but* cannot create the intended *relevance.* Unfortunately, a unified representation of a connective's meaning cannot straightforwardly be obtained from Blakemore's abstract analysis. Asr and Demberg (2012, 2013) propose a distributional account and analyse the distribution of connectives and discourse relations in natural texts. Their account suggests that *but* has similar function to that of other adversative connectives such as *although* when a coarse- grained relational inference is considered, and that they can mark the same set of relations, but that their distribution differs at finer levels. We examine the above arguments by designing stories, in which *but* and *although* can be applied locally to handle the general concessive relation between two clauses of the text. Then we look into the inferential effect of each connective on processing of the global context, i.e., a following sentence. 48 native English speakers on Amazon Mechanical Turk are recruited to score the coherence of variations of 24 stories similar to the following example.

Example:

SENT1: Amy's friends encouraged her to try tanning because her skin was so pale.

SENT2: (a) She thought of going to the beach, *although/but* her friends recommended a salon tan for her skin type.

(b) She thought of going to the tanning salon, *although/but* her friends recommended an outdoor tan for her skin type.

SENT3: She went to a nearby beach to lie in the sun.

In contrast to Fraser's formalism, the behavior of *but* in our experiment indicates that this connective indeed enforces the inference of a specific discourse relation that is different from the one inferred by *although* in the same context (the entire story in the Example is more coherent when sentence (a) is used with *although*, and conversely sentence (b) with *but*), whereas both connectives are equally good in their local context ((a) and (b) are scored equally coherent when SENT3 is excluded from the story). This finding is not compatible with an account where only the core meaning is part of the discourse connective and the rest is dependent on the context. Instead, it reveals that the contribution of the discourse connectives to the meaning of a story goes beyond the interpretations within the boundaries of a directly involved discourse relation; it can also change the reader's expectation of the broader context, by:

- affecting the information structure, e.g., changing the Question Under the Discussion [6] or focus of the story, and
- modulating the truth-conditional state of a possibly present implicature [4] (see the lastsentence in the Example as a confirmation vs. denial of the implied meaning by (a), when *although* vs. *but is utilized*).

References

[1]F. T. Asr and V. Demberg (2012) "Measuring the strength of linguistic cues for discourse relations." Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA):33.

[2]F. T. Asr and V. Demberg (2013) "On the information conveyed by discourse markers." Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL): 84.

[3]D. Blakemore (2004) "Relevance and linguistic meaning: The semantics and pragmatics of discourse markers." Cambridge University Press.

[4]B. Fraser (1999) "What are discourse markers?." Journal of Pragmatics 31.7: 931-952.

[5]H. P. Grice (1975) "Logic and conversation."

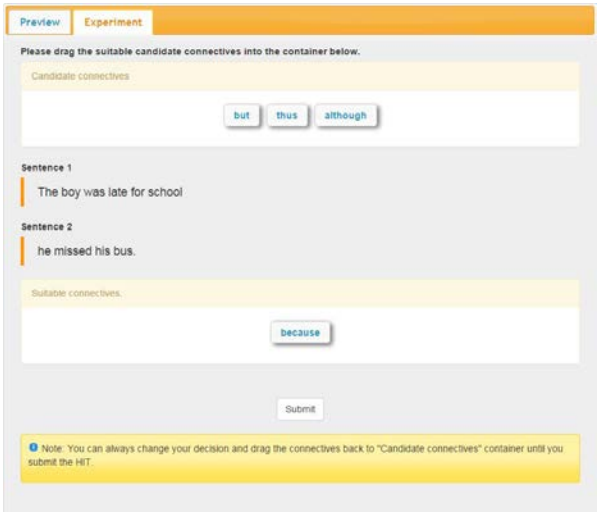
[6]C. Roberts (1996) "Information structure in discourse: Towards an integrated formal theory of pragmatics." Working Papers in Linguistics-Ohio State University Department of Linguistics: 91-136.

Discourse annotation via MechanicalTurk

Vera Demberg, Asad Sayeed, Florian Pusse

Saarland University

Discourse relation annotation is expensive and time consuming if done by trained annotators. We suggest an experimental design that can be used by naïve annotators via crowd-sourcing, e.g., Amazon's Mechanical Turk platform, or crowdflower. The idea is that discourse relations will be annotated entirely by which discourse connectors can occur between the discourse relation. For the annotation of a specific discourse relation, the subject sees the discourse arguments, as well as up to twelve discourse connectors that can be moved between the discourse relations via a drag-and-drop interface. The task is to drag-and-drop exactly those discourse connectors that "fit" for the arguments. Disambiguation is achieved by asking workers to drag-and-drop *all* the connectors that they think fit.



In a first annotation step, a set of 12 discourse connectors is available that together cover the complete space of possible discourse relations we want to annotate (e.g., "and", "but", "when", "before", "after", "because", "therefore", "instead" etc.). Depending on the first choice, the worker then gets a second set of 12 discourse connectors that should be dragged-and-dropped in a similar manner, and which allow us to distinguish between discourse relations in a more fine-grained way. For example, if the discourse connector chosen in the first round was "but", then the second set of discourse connectors (or alternative lexicalizations) to be annotated will include, for example, the connectors "although", "however", "while", "yet", "still", "on the other hand".

Careful and complete connector choice can be encouraged by running a two-people version of the game, where extra points can be earned by choosing the same connectors that were chosen by

other people, or by inserting "bonus" items for which we already know correct connectors and discourse relations (e.g., from PDTB), for which people get a bonus if they get those right (only letting them know after completing an item whether it was a bonus item).

We have the implementation of the drag-and-drop interface in place and are currently in the process of running our first experiments, for which we will present results and experiences on the poster. We will also discuss limitations with this approach, i.e., where full disambiguation with respect to the PDTB annotation could not be achieved, as well as inter-annotator agreement for this method. We will also be happy to share the software.

Validating categories of causal connectives: Converging evidence from corpus-based research and experiments

Jacqueline Evers-Vermeul & Ted Sanders

Utrecht University

Introduction and research question: Several theories have been proposed that make an inventory of the kind of relations that can be found in different types of discourse. Rhetorical Structure Theory (Mann & Thompson 1988; Taboada & Mann 2006) is among the most influential ones.

Sanders and colleagues argue in favour of a cognitive approach to coherence (Sanders et al. 1992; 1993; Sanders & Spooren 2009): if coherence relations are part of the cognitive representation that a reader makes of a text, they should have a cognitive status.

In this paper we address the cognitive validation of three categories that are frequently used to describe causal connectives and coherence relations: Sweetser's (1990) domains of use:

1) content – describing real-world causal relations;
2) epistemic – describing argument-conclusion relations;
3) speech act – giving arguments for performing the speech act.

Our research question is:

How can results from corpus-based and experimental acquisition studies inform us about the categorization of causal connectives and coherence relations?

The general idea behind this approach is that different orders in the acquisition of specific domains are windows on the cognitive categories that children use when producing causally related utterances.

Method: In this paper we review evidence from corpus-based and experimental studies on connective acquisition by 2- to 4-year-olds (Evers-Vermeul & Sanders 2009, 2011; Van Veen et al. 2009, 2013, in press). We focus on positive causal relations in three languages: Dutch, English, and German. We used converging methodologies to investigate when children discover the three domains in the use of causal connectives, and will discuss merits and drawbacks of each method in its role of giving insight in the cognitive categories under investigation.

Results & conclusion: We will show how the acquisition of connectives (*and*, *then*, *because*) can be accounted for by a cumulative complexity approach (Evers-Vermeul & Sanders, 2009). Experiments in which children had to describe causally related events, argue with, and instruct a hand puppet, revealed that even three-year-olds can produce causal connectives in all three

domains. Longitudinal corpus-based studies show that children as young as 2;8 are able to produce causal connectives in the content and the speech act domain, but that the epistemic domain is acquired later. Furthermore, growth curve analyses in which children's language use is related to parental input, reveals that context plays a crucial role in the production of domain types. Our approach of using converging methodologies proves fruitful: corpus-based data show us children's earliest spontaneous use and enable us to track longitudinal developments; experiments enable us to control for context effects. We will discuss implications for the annotation of causal connectives.

Discourse relation annotations, their annotators and how to deal with systematic dependence and response bias

Martin Groen

Universiteit Utrecht

Many annotation studies utilise Cohen's Kappa as a statistic to assess the amount of agreement between annotators correcting for the amount of inter-annotator variability due to chance. There are two concerns. First, Kappa was proposed by Cohen (1960) for cases where there are two observers. Second, Kappa is intended for situations where the different response categories are essentially independent and all disagreements are equally serious. This is very often not the case with discourse relation annotation schemes.

It is argued that we need to propose and adopt statistics that address these issues. Not only will results substantiated with Cohen's Kappa or an alternative be more reliable, they will potentially allow cross-study comparison too investigating heterogeneity in agreement for selected categories (Roberts, 2008).

It has been suggested to use weighted Kappa. This would make it possible to weigh the different categories, making sure that the hierarchical nature of many annotation schemes is taken into account when analysing annotator agreement. The choice of weights is very important, and Cohen (1968) recommends that a team of experts decides them. This last feature makes it unlikely that this will be adopted easily.

For cases where there are more than two annotators, Cohen's Kappa cannot be used either.

For these cases, Fleiss (1971) extended Cohen's Kappa in order to be able to study agreement between many observers. Interestingly, Fleiss proposed a statistic to analyse the probability that given an assignment to category A what is the probability that the same category is chosen for a next item, segment in our case. This could prove to be useful too.

We should take into account the response bias that many people manifest. In short, this is the tendency of humans to select a particular answer category, when they are actually not really sure that this is the appropriate category. For example, annotator A tends to say causal contrastive, even when he is not really sure whether it actually is a concessive relation. Signal detection analysis is the instrument of choice for these cases. A crucial assumption of signal detection analysis (SDA) is that the variances are similar. This is highly unlikely in the case

of discourse relations annotation. For these cases a non-parametric variant has been proposed (Pollack & Norman, 1964).

It is clear that annotating discourse relations is as challenging a task as the original task that SDA was developed for. As discourse relation annotation is such a difficult task, there is some uncertainty involved. Either a discourse relation determinant (DRD) is present (in SDA: signal present) or not (signal absent). Either the annotator correctly identifies the DRD (decision) or not (rejection). There are four outcomes: hit (DRD present and annotator identifies correctly), miss (DRD present and annotator does not identify), false alarm (DRD different as identified) and correct rejection (DRD correctly identified as different). Adopting SDA would take into account the response bias in a more systematic way then possible with Cohen's Kappa or the measures derived.

References

Cohen, J. (1960). *A coefficient of agreement of nominal scales*. Educational and Psychological Measurement, 20, 37–46. doi:10.1177/001316446002000104

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5).

Pollack, I., & Norman, D. A. (1964). *A non-parametric analysis of recognition experiments*. Psychonomic Science, 1(1-12), 125–126. doi:10.3758/BF03342823

Roberts, C. (2008). *Modelling patterns of agreement for nominal scales*. Statistics in Medicine, 27(6), 810–830.

DRDs in a contrastive perspective: a corpus-based cognitive study

Barbara Lewandowska-Tomaszczyk, Piotr Pęzik, Paul A. Wilson, Jerzy Tomaszczyk

University of Lodz

1. Research objectives: This poster presents team research on quantitative and qualitative evidence wrt to functions at the syntactic/semantic interface of Polish and English DRDs. Emphasis is put on the cognitive grounding of markers involving the degree of minimization of the addresses' processing effort in utterance production and comprehension.

2. Materials and Search Engines (Piotr Pęzik): The data comes from the National Corpus of Polish (Pęzik 2012) and the British National Corpus (BNC 2001). The Paralela search engine (http://clarin.pelcra.pl/Paralela/) is used to explore the distribution in a 100 million word collection of Polish-English/English-Polish translations.

3. Methodology: Cognitive corpus analysis, Discourse analysis (Barbara Lewandowska-Tomaszczyk, Piotr Pezik, Paul A. Wilson, Jerzy Tomaszczyk). We employ an interdisciplinary perspective, involving corpus-based cognitive analysis, observational and interpretive studies and discourse analysis. Meaning in discourse is created 'online' and it involves inferential processes and the processes of *frame-modification*, *frame- shifting* and *blending* (Fauconnier and Turner 1996). It is argued that DRDs provide clues as to the type of *mental spaces* in the process of *conceptual integration* and function as *semantic organizers* in discourse comprehension in the *information flow*. They make accessible complex and varied layers of meanings and networks of associations. The meaning of a linguistic unit is characterized here in terms of the *changes* it brings about to a given *discourse domain* (discourse incrementation).

4. Cognitive conditioning of discourse marking (Paul A. Wilson, Barbara Lewandowska-Tomaszczyk): As discourse markers tend to function as the focus of attention in utterances, the addressees' processing effort in utterance production and comprehension are studied in spoken materials of the English and Polish corpora.

5. Negative meaning in Polish and English discourse markers of event reference (Barbara Lewandowska-Tomaszczyk 1996, 2004): The study focuses on the distribution of discourse markers with inherent negative meanings such as English and Polish modal-volitional-evaluative *Why x?*, *Oh no!*, *Not that* (Schmid 2013) and others such as *in fact* (viz. *besides* or *indeed*) and connectors *yet*, *nevertheless* and however.

Negative markers are space builders whose *technical instruction* typically involves expelling part of the material (presuppositional) from the discourse domain. The conceptual material in the *scope of its predication* (cf. Langacker 1987) is then blocked from the discourse domain under construction. We argue for a multi-functional character of DRDs (*cognitive*, *evaluative* and *volitional*), associated with a *scale of modality* and likely *to minimize the addressee's cognitive effort to process the meaning of what is uttered.*

6. New Tools: Piotr Pęzik (2014) explores the hypothesis that functions of recurrent discourse devices in conversational data are systematically associated with distinct prosodic patterns (Crystal 1969). Examples cover polysemy - frame-shifting paradigm in cognitive linguistics terms - e.g., Pol. *daj spokój,*with *stopping* and *stance expressing* functions, associated with distinct prosodic features of pitch contours, intensity (normalised loudness) and duration. The extension of the Spokes search engine (http://spokes.clarin-pl.eu), used for the time-aligned collections and prosodic analysis is planned.

7. Expected results: We aim to arrive at a DRDs taxonomy with a function-, layer- and scale-based annotation system, targeted towards use in cross-linguistic research.

Selected bibliography

BNC, Consortium. 2001. The British National Corpus, Version 2 (BNC World). *Distributed by Oxford University Computing Services.*

Chafe, Wallace, 1994. Discourse, Consciousness, and Time: the flow and displacement of conscious experience in speaking and writing. University of Chicago Press Crystal, David. 1969. Prosodic systems and intonation in English. Cambridge: Cambridge University Press

Fauconnier, Gilles 1985. Mental Spaces: Aspects of Meaning Construction in Natural *Language*. Cambridge: CUP.

Fauconnier, Gilles & Mark Turner 1996. Blending as a central process of grammar. In: Adele E. Goldberg (ed.) *Conceptual Structure, Discourse, and Language*, CA: CSLI Publications. 113-131.

Langacker, Ronald W. 1987, 1991. *Foundations of Cognitive Grammar* Parts 1 and Stanford: Stanford University Press.

Lewandowska-Tomaszczyk, Barbara 1987. Conceptual Structure, Linguistic Meaning, and *Verbal Interaction.* Lodz: Lodz University Press.

Lewandowska-Tomaszczyk, Barbara 1996. Depth of Negation: A Cognitive Semantic *Study.* Lodz: Lodz University Press.

Lewandowska-Tomaszczyk, Barbara 2004. Conceptual blending and discourse functions. The case of no nie 'oh, no'. Research in Language 2. 33–47.

Pęzik, Piotr. 2012. Język mówiony w NKJP. (Spoken Language in the NCP). In: Narodowy Korpus Języka Polskiego, edited by Adam Przepiórkowski, Mirosław Bańko, Rafał Górski, and Barbara Lewandowska-Tomaszczyk, 37–47. Warszawa: Wydawnictwo Naukowe PWN.

Pęzik, Piotr 2014. Spokes – a search and exploration service for conversational corpus data. http:// pelcra.pl/docs/lib/exe/fetch.php?media=spokes_sub_2.pdf.

Scott, Kate 2006. When less is more: Implicit arguments and Relevance Theory. *UCL Working Papers in Linguistics* 18, 139-170.

Schmid, Hans-Joerg 2013. Is usage more than usage after all? The case of English *not that.* Linguistics 1, 51: 75-116.

Seuren, Pieter A. M. 1985. *Discourse Semantics.* Oxford: Blackwell.

Sperber, Dan & Deirdre Wilson 1986/1995. *Relevance Theory: Communication and cognition.* Oxford: Blackwell.

Discourse Structure of Back Covers: A pilot study

Laurent Prévot, Anaïg Pénault, Grégoire Montcheuil, Stéphane Rauzy, Philippe Blache

Aix Marseille Université

Precise and reliable discourse linking between independent syntactic clauses (or elementary discourse units) is the crucial next step for language studies and Natural Language Processing. It can solve many spurious syntactic analysis problems such as the analysis of lengthy realistic written data sentences, analysis of spoken data (in which punctuation is absent), analysis of various DRD that only merely fit into traditional syntactic tradition. However, at this stage, precise discourse relation establishment tend to become very hard to achieve on longer texts due to the multiplication of attachment sites in the left context. In this work we propose to address this question from a practical and experimental angle by :

1. using a corpus of back-covers (that are typically made of 5 to 10 sentences)
2. annotating a small subset (5) of it with some of the main theories available (SDRT, Penn DTB, and others if time allows)
3. performing eye-tracking experiments with these texts

At this stage, we have the eye tracking data of a previous study (Rauzy & Blache, 2012) but for which the texts were much longer (Newspaper Corpus "Le Monde") and discourse annotation for some of the texts used in this experiments. However, given the nature of the texts and our capacity to record more eye-tracking data on these shorter texts, the material for the workshop will be entirely new.

The objective of such crossing between discourse annotation and eye-tracking recordings is twofold: First, we will be able to provide empirical support to the theories considered (perhaps not of the same strength for all theories) by using the structures annotated as a model for explaining fixation times and trajectories. Second, more general principles such as the Right Frontier Constraint (Asher, 2008 ; Prévot & Vieu, 2008) will be evaluated against the eye-tracking data. More precisely, a key element will consist in comparing fixation time for discourse units that have a direct unique attachment to immediately previous discourse unit vs. the one that attach in another location of the discourse structure. Our objective is then to use this methodology also on Mandarin Chinese backcover database (partially matched with the French one). Overall, we believe that backcovers are interesting for multilingual or comparative research.

References

Asher, N. (2008). Troubles on the right frontier. PRAGMATICS AND BEYOND NEW SERIES, 172, 29

Prévot, L., & Vieu, L. (2008). The moving right frontier. PRAGMATICS AND BEYOND NEW SERIES, 172, 53

Stéphane Rauzy & Philippe Blache (2012) "Robustness and processing difficulty models. A pilot study for eye-tracking data on the French Treebank", in proceedings of Eye-tracking and NLP workshop, COLING2012.

Discourse markers and position: consequences for processing

Inés Recio, Laura Nadal, Adriana Cruz

Universität Heidelberg

Key words: information structure, discourse reformulation, *incluso*, connectives, contrast, *sin embargo*, position, eye-tracking, processing effort

One of the challenges for the study of discourse relations lies in describing the relationship between discourse markers and their position (Briz & Pons 2010). Such is the case of Spanish *incluso* (EN. 'even'), which can be employed in either pre-focal or in post-focal position:

(1) Ana ha estado en París, en Londres, incluso *en Tokio.*

'Ana has been to Paris, to London, *incluso* to Tokyo.'

(2) Ana ha estado en París, en Londres, en Tokio incluso*.*

or *sin embargo* (EN. 'however'), which can take an initial, intermediate or final position within its discourse segment:

(3) Estos niños comen mucho dulce. Sin embargo*, están sanos.*

'These children eat a lot of sweets. *Sin embargo*, they are healthy.'

(4) Estos niños comen mucho dulce. Están, sin embargo*, sanos.*

(5) Estos niños comen mucho dulce. Están sanos, sin embargo*.*

Position has been argued to have an impact on the functional meaning of certain discourse markers and to favour polyfunctionality (Loureda & López Serena 2013; Loureda *et alii* (in press)). In a post-focal position, the focus-marking device *incluso* acquires a further discourse function of "discursive reformulation" (Loureda & López Serena 2013). In other words, a function on the level of reformulation is "added" to its prototypical information structuring function. In turn, the counter-argumentative meaning of the connective *sin embargo* might not be affected by positional shifts in exactly the same way as *incluso*, its mobility being rather linked to "aspects such as discourse traditions or register" (Briz & Pons 2010: 283, our translation). These theoretical statements make the following research questions arise:

1. Shifting the position of a discourse marker can lead to differences in discourse processing.
2. The consequences of position shifting could differ if a) the change of position leads to a confluence of several discourse values in the marker (*incluso*); or b) mobility affects primarily other textual features, such as register, and not the connective's function itself (*sin embargo*).

In eye-tracking reading experiments carried out by members of the research group *Diskurspartikeln und Kognition* (DPKog) at Heidelberg University the questions above were addressed with following general results:

- Higher cognitive efforts were found for processing utterances with post-focal *incluso* and for the marked positions of *sin embargo*, to the detriment of other functional areas of the utterance (focus / discourse segment).
- Despite the different functional consequences of shifting the position of *incluso* and *sin embargo*, processing results give account of similar processing patterns.

These first experimental outcomes show the need of deepening research in the interface discourse marking-position, and the advantages of adopting a cognitive approach to do so. This

supports and develops descriptive and corpus-driven accounts of the role of discourse markers in discourse representation.

References

Briz, A., S. Pons & J. Portolés (dirs.), *Diccionario de partículas discursivas del español* (DPDE) [online], www.dpde.es, Servei de Publicacions de la Universitat de València, Valencia. ISBN: 978-84-691-4416-9.

Briz, A. & S. Pons (2010), "Unidades, marcadores discursivos y posición", in Loureda, Ó. & Acín, E. (eds.), *Los estudios sobre marcadores del discurso en español, hoy*, Madrid, Arco/Libros, 327-358.

Loureda, Ó. & A. López Serena (2013), "La reformulación discursiva entre lo oral y lo escrito: una aproximación teórica y experimental", *Oralia* 16, 221-258.

Loureda, Ó., A. Cruz, I. Recio & C. Villalba (in press), "*Incluso* en posición pre y postfocal: un análisis experimental de los costes de procesamiento de escalas pragmáticas", *Revista Española de Lingüística*.

Applying a cognitive approach to coherence relations to discourse annotation: Annotating coherence relations in corpora of language use

Ted J.M. Sanders & Merel C.J. Scholman

Utrecht University

The advent of linguistic corpora is an important stimulant for language use researchers. The focus area of corpora has mainly been on lexical, syntactic and semantic characteristics of language. However, the notion of "discourse", and more specifically the coherence relations between parts of discourse such as *cause-consequence* and *claim-argument,* has become increasingly important in linguistics over the years. This has led to the international tendency in the last decennium to create discourse annotated corpora. Leading examples are the Penn Discourse Treebank (Prasad et al., 2008), the RST Treebank (Carlson et al., 2001) and the Potsdam Commentary Corpus (Stede, 2004). While discourse annotation guidelines generally agree on the idea of coherence relations, a uniform standard for discourse annotation is not yet available. Current annotation methods often lack a systematic order of coherence relations, which increases the difficulty of the annotation task. Due to this, annotators need large manuals and intensive training before they can start annotating.

In this contribution, we investigate the usability of a cognitive approach to coherence relations (Sanders, Spooren & Noordman, 1992) for discourse annotation. The theory proposes a taxonomy of coherence relations in terms of four cognitive categories. On the basis of these categories, a systematic, step-wise annotation scheme was developed, which may facilitate the annotation process. Two annotation experiments are presented which investigated the reliability and validity of this approach in discourse annotation. In these experiments, non-trained, non- expert annotators analyzed fragments using a short manual and an instruction. An implicit and explicit version of the instruction was created to determine whether the type of instruction influences the reliability of the scheme. The implicit instruction relied only on the annotator's knowledge of the categories. The explicit instruction relied on this knowledge, as well as on text- linguistic insights. This instruction contained paraphrase and substitution tests, which were hypothesized to facilitate the decision making process. The results showed that non-trained, non- expert annotators can reach fair to almost perfect agreement using the cognitive approach to coherence relations. Given the little amount of training these annotators received, this amount of agreement is promising. Moreover, annotators using the explicit instruction showed higher agreement than annotators

using the explicit instruction. We will discuss to what extent the categories of coherence relations area applicable in discourse annotation and how the approach can be used in cross-linguistic corpus research.

References

Carlson, L.; Marcu, D. and Okurowski, M.E. (2001). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, Eurospeech 2001, Denmark.

Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A. and Webber, B. (2008). The Penn Discourse Treebank 2.0. Proceedings of the 6th International Conference of Language Resources and Evaluation (LREC 2008), Marrakech.

Sanders, T.J.M.; Spooren, W.P.M.S. and Noordman, L.G.M. (1992). Toward a Taxonomy of Coherence Relations. *Discourse Processes,* 15: 1–35.

Stede, M. (2004). The Potsdam Commentary Corpus. *Proceedings ACL Workshop on Discourse Annotation.* Pennsylvania: ACL.

Three-layer approach towards the cognitive representation and linguistic marking of subjectivity and perspective

Yipu Wei, Jacqueline Evers-Vermeul

Utrecht University

Background: People use linguistic elements to express cognitive features. Some linguistic elements express discourse coherence, such as connectives. Specific connectives also code information in the dimension of subjectivity, such as Dutch *want* 'because'. These connectives function as processing instructions for readers and affect the on-line processing of coherence relations (Canestrelli, Mak and Sanders, 2013). Processing studies suggest that other linguistic elements such as epistemic modals (*perhaps, probably*) and expressions with cognition/communication verbs (*John thinks/says*) also influence on-line processing and interact with the processing effects introduced by connectives (Canestrelli, Mak and Sanders, 2013; Traxler, Sanford, Aked and Moxey, 1997). Works on subjectivity and perspective have termed these expressions as perspective markers (Verhagen, 2005) or mental space builders (Sanders, Sanders and Sweetser, 2012).

Research questions & Methods: Most previous studies look at perspective and subjectivity separately. With this theoretical paper, we aim to provide a unified account towards subjectivity and perspective and answer the following research questions: at the linguistic level, what are the functions of linguistic elements in marking various degrees of subjectivity and different perspectives? How can we categorize these linguistic elements using an integrated approach and how do they interact with one another in the discourse representation? At the cognitive level, what is the relation between subjectivity and perspective? Can perspective and subjectivity be analyzed in an integrated fashion?

We have conducted a literature study on previous research to develop a unified approach. The new model is expected to systematically describe the isolated linguistic phenomena of perspective markers reported in literature and to analyze subjectivity and perspective in an integrated way.

Analysis: We will present a three-layer framework, developed from the three metafunctions by Halliday (1985). Under this framework, information coded by linguistic elements is functional at different layers, i.e. the propositional layer, the relational layer and the interpersonal layer. Perspective markers can be categorized according to the layers they function at. The integrated approach can be applied to account for the phenomena such as scope ambiguity, ambiguity in defining causal domains, etc. From the cognitive aspect, this three-layer model gives insights into the conceptual configuration of subjectivity and perspective: subjectivity in coherence relations is formed at the relational layer, while the extra cognitive effort brought by subjectivity is relieved at the interpersonal layer by the explicit marking of perspectives. Furthermore, this three-layer approach provides a new view point to analyze discourse markers, specifically connectives in relations with other perspective markers in corpora.

Key words: connectives, subjectivity, perspective, perspective markers

Selected references

Canestrelli, A., Mak, W. & Sanders, T. (2013). Causal connectives in discourse processing: How differences in subjectivity are reflected in eye-movements. *Language and Cognitive processes, 28(9)*, 1394-1413.

Halliday, M. (1985). *An Introduction to Functional Grammar*. Edward Arnold Publishers: London.

Sanders, T., Sanders, J., & Sweetser, E. (2012). Responsible subjects and discourse causality.

How mental spaces and perspective help identifying subjectivity in Dutch backward causal connectives. *Journal of Pragmatics, 40*(2), 191-231.

Traxler, M., Sanford, A., Aked, J., & Moxey, L. (1997). Processing causal and diagnostic statements in discourse. *Journal of Experimental Psychology: Learning Memory, and Cognition, 23*(1), 88-101.

Verhagen, A. (2005). Constructions of intersubjectivity. Discourse, syntax, and cognition. Oxford: Oxford University Press.

Annotating the meaning of connectives in multilingual corpora

Sandrine Zufferey[1] & Liesbeth Degand[2]

1Université de Fribourg, 2Université Catholique de Louvain

In this poster, we present three annotation experiments performed on parallel directional corpora of newspaper articles (Zufferey & Degand, 2013). In these experiments, we tested the applicability of the PDTB annotation scheme (Prasad et al., 2008) as a tool to simultaneously annotate connectives in five Indo-European languages (English, French, German, Dutch and Italian). The rationale for conducting these experiments is that even though the PDTB has been adapted to a variety of languages (see Webber & Joshi, 2012 for a review) it had to our knowledge not been used for multilingual annotations.

We discuss the methodological choices that we made in order to perform the annotation experiments; namely the advantages and disadvantages of using parallel corpora instead of comparable ones, the necessity to use of a pivot language in order to compare the cross-linguistic reliability of annotations, and the way inter-annotator agreements were computed.

Based on the results of a first pilot experiment, we identified problematic labels from the PDTB in terms of inter-annotator agreement, and argue that these problems are recurrent across languages. Based on these results, we implemented a number of simplifications to the PDTB tag

set and tested their usefulness in two more extensive experiments: one monolingual experiment in French and one multilingual experiment using the same five Indo-European languages. We demonstrate that these changes led to significant improvements in inter-annotator agreement.

We conclude that the PDTB is a useful tool that can be adapted for cross-linguistic annotations with similar levels of agreement compared to monolingual annotations expect in two cases: translation shifts altering the meaning of connectives in translated data and use of language specific connectives with no real equivalents in the target language system.

References

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, PLACE, 2961–2968.

Webber Bonnie, and Aravind Joshi. 2012. Discourse Structure and Computation: Past, Present and Future. Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, p. 42–54, Jeju, Republic of Korea.

Zufferey, Sandrine and Liesbeth Degand. 2013. Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*. Published online ahead of print.

Turkish Discourse Bank Tools

Ruket Çakıcı[1], Ayışığı Sevdik-Çallı[1], Deniz Zeyrek[1], Berfin Aktaş[1], Utku Şirin[2], Cem Bozşahin[1], Işın Demirşahin[1]

1Middle East Technical University, 2École Polytechnique Fédérale de Lausanne

Turkish Discourse Bank (TDB) is a 400K subcorpus of METU Corpus (Say et al.,2002) annotated for discourse relations (Zeyrek et al., 2008; Zeyrek et al., 2009; Zeyrek et al., 2010). In this study we describe the tools that were created for annotating and browsing the TDB.

TDB was annotated using an annotation tool, DATT, created for the sole purpose of annotating discourse relations in Turkish (Aktaş et al, 2010). DATT adds a layer of annotation in the form of character indices to the existing level of raw text. Annotation files are created in XML format. The choice of characters as marking units is an intentional decision so that the annotation is not limited by the word boundaries, to provide the necessary ground for future annotations of some connectives that are at the sublexical level in Turkish. DATT also includes search functionality supporting allomorphy and different inflections of connectives, which allow different forms of inflectional suffixes affecting the semantics of a connective (e.g. the factive nominal suffix "-dık" has eight different allomorphs).

METU TDB Browser (Şirin et al, 2012) uses the text and the annotation files and displays the relations in the TDB. In addition to search options such as connectives only in certain genre, the browser performs complex searches combining text and regular expression matches on different parts of the annotations such as connectives, arguments, modifiers of connectives, supplements etc.

METU TDB Browser has a built in feature to distinguish discontinuous and adjacent arguments while searching. Furthermore, both DATT and the Browser are designed to incorporate sense once sense is annotated in the TDB. The browser gives simple statistics on the search results and displays the relations in a window of text that contains the text file and the annotations on the file that are the results of the specific search (Şirin et al, 2010).

Both tools are available with the Discourse Bank release downloadable at the following url: http://medid.ii.metu.edu.tr/index_eng.html

References

Berfin Aktaş, Cem Bozşahin, and Deniz Zeyrek. 2010. Discourse Relation Configurations in Turkish and an An- notation Environment. In *Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010,* pages 202–206.

Bilge Say, Deniz Zeyrek, Kemal Oflazer, and Umut Özge. 2002. Development of a Corpus and a Treebank for Present-day Written Turkish. In *Proceedings of 11th International Conference on Turkish Linguistics,* pages 183–192.

Utku Sirin, Ruket Cakici and Deniz Zeyrek. 2012. METU Turkish Discourse Bank Browser , In *Proceedings of LREC(2012*), pages 2808-2812, Istanbul, Turkey.

Deniz Zeyrek, Ümit Turan, Cem Bozşahin, Ruket Çakıcı, Ayışığı Sevdik-Çallı, Işın Demirşahin, Berfin Aktaş, Ihsan Yalçınkaya, and Hale Ögel. 2009. Annotating Subordinators in the Turkish Discourse Bank. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009,* pages 44–47.

Deniz Zeyrek, Işın Demirşahin, Ayışığı Sevdik-Çallı, Hale Ögel Balaban , Ihsan Yalçınkaya , and Ümit Deniz Turan . 2010. The Annotation Scheme of the Turkish Discourse Bank and An Evaluation of Inconsistent Annotations. In *Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010,* pages 282–289.

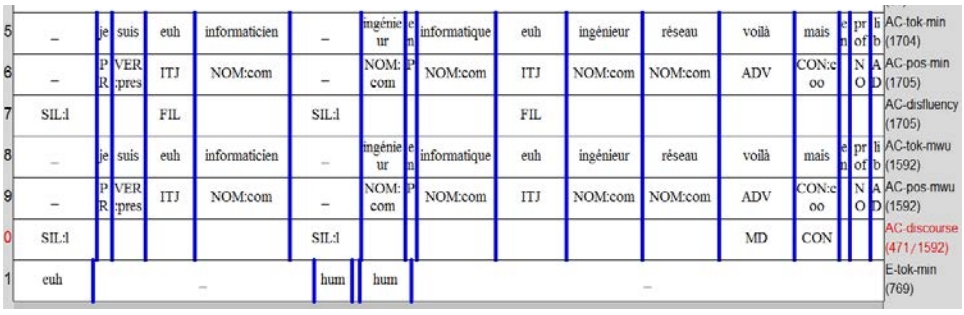
Automatic Detection of Discourse Structuring Devices in French using the *DisMo* Corpus Annotator

George Christodoulides[1], Mathieu Avanzi[1], Giulia Barreca[2]

1Université Catholique de Louvain, 2CNRS, Université Paris Ouest Nanterre La Défense

DisMo is an automatic a multi-level annotator that integrates part-of-speech tagging with disfluency detection and multi-word unit recognition (Christodoulides et al. 2014). It is designed to cope with the specific characteristics of spoken language, and more generally of "non-canonical" text (e.g. informal computer-mediated communication), such as the absence of punctuation and often incomplete syntactic structures. *DisMo* is a hybrid system that uses a combination of lexical resources, rules, and statistical models based on Conditional Random Fields (CRF).

The version for French has been trained on data originating from the PFC Corpus (*Phonologie du Français Contemporain*, Durand et al. 2002): initially, a 60k-token corpus of spontaneous speech (a subset of the corpus presented in Avanzi 2014, including 15 different regional varieties and balanced for speaker age and sex), and subsequently on a 150k-token corpus of spontaneous speech (Barreca et al. 2014). Discourse structuring devices were manually identified by two expert annotators in these two training corpora. An iterative process of manual correction and retraining improved the system's accuracy. *DisMo* supports a multi-level annotation scheme, in which the tokenisation to minimal word units is complemented with multi-word unit groupings (each having associated POS tags), as well as separate levels for annotating disfluencies and discourse phenomena (see figure below, for a timeline-based representation of the three levels of annotation):



In this poster presentation we will focus on the aspects of DisMo that provide the automatic detection of discourse structuring devices, including discourse markers and connectors. We show how the combination of POS tagging (after taking into account disfluencies and multi-word units) coupled with a shallow syntactic parsing into minimal chunks is used to statistically detect DSDs (with a weak clause association, cf. Schourup 1999) and potential DSDs. Among the perspectives of

this research is the large-scale corpus-based automatic detection of potential DSDs, facilitating further research on the phenomenon in the context of the TextLink COST project.

References

Avanzi, M. (2014). A Corpus-Based Approach to French Regional Prosodic Variation. *Nouveaux cahiers de linguistique française*, vol. 31, 2014.

Barreca, G.; Christodoulides, G. (2014). Un concordancier multi-niveaux pour des corpus oraux. *Actes de la 21ème Conférence Traitement Automatique du Langage Naturel (TALN)*, Marseille, France, 1-4 juillet 2014.

Christodoulides, G.; Avanzi, M.; Goldman, J.-Ph. (2014). DisMo: A morphosyntactic, disfluency and multi-word unit annotator. An evaluation on a corpus of French spontaneous and read speech. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 26-31 May 2014, pp. 3902-3907. Available online: www.corpusannotation.org/dismo.

Durand, J., Laks, B., Lyche, C. (2002). La phonologie du français contemporain: usages, variétés et structure. In Pusch et Raible (2002), pp. 93-106.

Schourup, L. (1999). Discourse markers. *Lingua* 107 (3-4), pp. 227-265.

Using Collaborative Tools For Building And Annotating Multilingual Knowledge

Mauro Dragoni

Fondazione Bruno Kessler

In the presented poster, I will propose, in the context of the TextLink Cost Action, the use of a collaborative knowledge management tool, called MoKi, developed in the context of several EU-funded project and used for modeling knowledge resources in a collaborative way.

I will start by highlight and discussing the potential and criticality of using Web 2.0 semantic technologies and tools to enhance participatory knowledge sharing, interoperability, and collaboration in the modeling of complex domains.

Then, I will show why MoKi might be a good technical solution for creating and sharing a common repository of resources in the TextLink context.

In particular, I will present in more detail the features mostly related to the tasks of each Working Group: (i) the possibility of creating corpora using a wiki-based tool, this way people are allowed to work collaboratively by speeding up the building process of metadata vocabularies, knowledge bases, etc.; (ii) connecting the tool with translation services in order to foster the development of multilingual resources, (iii) the definition of alignments between multilingual knowledge resources (vocabularies, ontologies, etc.), and (iv) the possibility of using facilities for text annotation.

Finally, I will show how the tool can be used for the creation of a central resource repository in order to have a unique environment for adding, updating, and planning the share of the produced material.

Computational tools for the representation of discourse structures at the University of Évora

Teresa Gonçalves and Paulo Quaresma

University of Évora

In the NLP group of the Computer Science Department of the University of Évora we have created several computational tools that may be used for the automatic analysis and representation of discourse structures.

The most related tool is AutemaDis [LQC 2006], a framework developed in the context of the PhD work of Ana Luísa Leal, which is able to automatically identify, for the Portuguese language, a subset of the RST discourse relations. AutemaDis uses the output of a syntactic parser (PALAVRAS) to identify text segments and, then, it uses a set of logical rules to infer the relations between the segments. AutemaDis main steps are:

- Identification and classification of textual components (from the parser PALAVRAS output);
- Categorization and segmentation of textual components – segments and subsegments (logical rules written in Prolog);
- Organization of the constituents in a tree format, according to the formal and conceptual hierarchy (logical rules);
- Ascription of rhetorical relations (from RST) between textual components (logical rules);
- Presentation of the summarized structure of the analyzed text.

In the context of the MSc thesis of João Sequeira [SGQ 2012] we have created a corpus for semantic role labelling for the Portuguese language (which we believe may also be used for labelling discourse structures). This corpus is based on Bosque 8.0[7] and includes morphologic, syntactic and semantic role information and includes information from 4416 sentences. Using the same format as the one used in CONLL'2004, it has a word per line and contains the following 7 features: word, lemma, part-of-speech tag, chunks with IOB, semantic roles with IOB, named entities with IOB and clauses. Words, lemmas, chunks and clauses were extracted from the Bosque8.0 corpus; the part-of-speech column uses LABEL-LEX tags having performed a manual review in ambiguous situations; the named entities were obtained an in-house NER tool. In the same context, a preliminary tool was also developed using a machine-learning framework (Minorthird) to learn and predict "predicates", "subjects" and "objects", being able to obtain results similar to the state of the art for the English language.

[7]http://www.linguateca.pt/Floresta/corpus.html

For the juridical domain, Prakash Poudyal, another MSc student, developed a tool for the annotation of some semantic roles [PQ 2012]. With this tool it was possible to automatically identify and annotate distinct juridical entities and the relations between them. The approach was also based on the Minorthird framework. For this work a corpus based on EUR-Lex[8] documents was also created.

As ongoing work with Amália Mendes and Iris Hendrickx we are applying a machine learning approach to predict the modal meaning of a verb in a sentence [QMHG2014]. Modality is the expression of the speaker's (or the subject's) attitude towards the content of the sentences. This tool takes into account features available from the PALAVRAS parser (morphological, syntactic and semantic information) from the word, word's window context and its path to the sentence root node. As golden data set we are using a corpus of 160.000 tokens manually annotated, according to a modality annotation scheme for Portuguese from of the work of [HMM2012].

References

[HMM2012] Iris Hendrickx, Amália Mendes, and Sílvia Mencarelli: Modality in text: a proposal for corpus annotation. In: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) LREC'2012 – Eighth International Conference on Language Resources and Evaluation. pp. 1805–1812. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012)

[LQC 2006] Ana Luísa Varani Leal, Paulo Quaresma, and Rove Chishman. "From syntactical analysis to textual segmentation". In R. Vieira, P. Quaresma, M. G. V. Nunes, N. Mamede, C. Oliveira, & M. C. Dias (eds.) PROPOR, volume 3960 of Lecture Notes in Computer Science, 252–255. Berlin. Springer. 2006.

[SGQ 2012] João Sequeira, Teresa Gonçalves, and Paulo Quaresma. Semantic role labeling for portuguese - a preliminary approach -. In Helena de Medeiros Caseli, Aline Villavicencio, António J. S. Teixeira, and Fernando Perdigão, editors, PROPOR, volume 7243 of Lecture Notes in Computer Science, pages 193–203. Springer, 2012.

[PQ 2012] Prakash Poudyal, and Paulo Quaresma. An hybrid approach for legal information extraction. In Burkhard Schafer, editor, Legal Knowledge and Information Systems - JURIX 2012: The Twenty-Fifth Annual Conference, University of Amsterdam, The Netherlands, 17-19 December 2012, volume 250 of Frontiers in Artificial Intelligence and Applications, pages 115–118. IOS Press, 2012.

[QMHG2014] Paulo Quaresma, Amália Mendes, Iris Hendrickx, and Teresa Gonçalves. Tagging and labelling portuguese modal verbs. In Jorge Baptista, Nuno J. Mamede, Sara Candeias, Ivandré Paraboni, Thiago Alexandre Salgueiro Pardo, and Maria das Graças Volpe Nunes, editors, PROPOR, volume 8775 of Lecture Notes in Computer Science, pages 70–81. Springer, 2014.

[8]http://eur-lex.europa.eu/homepage.html

Prague Dependency Treebank 3.0 and PML-Tree Query

Jiří Mírovský, Eva Hajičová

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

The intended poster/demo will show how to search for discourse relations in the Prague Dependency Treebank using a powerful and user-friendly PML-Tree Query search system.

The Prague Dependency Treebank 3.0 (PDT; Bejček et al., 2013) is a corpus of Czech, consisting of almost 50 thousand sentences annotated mostly manually on three layers of language description: morphological, analytical (surface syntactic structure), and tectogrammatical (deep syntactic structure). On top of the tectogrammatical layer, explicitly marked discourse relations, both inter- and intra-sentential ones, have been annotated.

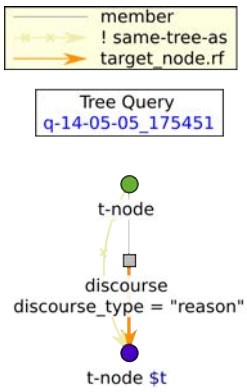
For searching in PDT, a client-server based system called PML-Tree Query has been developed (PML-TQ; Pajas and Štěpánek, 2009). It belongs to the most powerful systems for searching in treebanks. Queries in PML-TQ can be created both in a textual form and in a graphical environment. The query language allows to define properties of tree nodes and relations among them, inside or between sentences and also across layers of annotation. Negation on the tree structure and Boolean expressions over the relations can be used. Results of the corpus search can be viewed along with the context or processed with output filters to produce statistical tables.

The following example query defines two tectogrammatical nodes (t-nodes) connected with a special "member" node that represents a discourse relation between the two nodes. The required type of the discourse relation can be specified at the member node, in this example it is set to "reason". The query also specifies that the start and target nodes of the relation are not from the same tree, i.e. it looks for an inter-sentential discourse relation of the semantic type "reason".

Textual form of the query:

```
t-node
[ !same-tree-as $t,
     member discourse
     [ discourse_type = "reason",
 target_node.rf t-node $t := [  ] ] ];
```

Graphical form of the query:



The following two sentences represent one of the results of the query:

Pronikání do cizích počítačových systémů je podle našich zákonů beztrestné.
Policie **tak** jen bezmocně přihlíží, když v bankách řádí SLÍDILOVÉ.

[Infiltration into other computer systems is according to our laws not a criminal act.
Thus the police only helplessly watches, as SNOOPERS rage in banks.]

Results of queries in PML-TQ can be further processed using output filters. Thanks to an output filter, a result of a query does not consist of individual matching positions in the data but of a tabular summary of all the matching positions, as specified by the output filter. If we modify the previous query by deleting the definition of the discourse type (`discourse_type = "reason"`), naming the member node (`$d :=`) and adding an output filter (the last line with prefix >>):

```
t-node
[ !same-tree-as $t,
     member discourse $d :=
         [ target_node.rf t-node $t := [ ] ] ];
>> for $d.discourse_type give $1, count() sort by $2 desc
```

...the query will search for all inter-sentential discourse relations in the data and – thanks to the output filter – produce the following distribution table of the discourse types, sorted in the descending order by the number of occurrences (only a few selected lines are printed here to save space):

opp	1,800
Conj	1,389
Reason	1,031
…	
grad	204
Restr	172
Explicat	130
…	

Table 1: (Selected) results of the output filter

References

Bejček, Eduard, Hajičová, Eva, Hajič, Jan et al. (2013). *Prague Dependency Treebank 3.0.* Data/software, Charles University in Prague, MFF, ÚFAL. Available at: http://ufal.mff.cuni.cz/pdt3.0/.

Pajas, Petr, & Štěpánek, Jan (2009). System for Querying Syntactically Annotated Corpora. In: *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, Association for Computational Linguistics, Suntec, Singapore, ISBN 1-932432-61-2, pp. 33-36.

Annotating and learning full discourse structures for texts and dialogues

Nicholas Asher

CNRS Toulouse

In my talk I'll talk about some of the successes and failures of 3 efforts to annotate texts and dialogues with full discourse structures. I will also briefly review work that my colleagues and I have done to compare discourse structure annotation schemes from different theoretical frameworks---- RST, SDRT and a dependency tree or graph annotation scheme. I'll discuss the expressive power of these different schemes, and review, time permitting, some of the technical results that permit a translation between these various schemes. I'll then mention how we have used these annotations for automatically extracting discourse structures from text and dialogue.

The ISO Semantic Annotation Framework for Discourse Relations

Harry Bunt

Tilburg University

Within the ISO Semantic Annotation Framework a number of annotation schemes is developed with the aim to support the creation of interoperable corpus resources enriched with semantic annotations. For two areas an ISO standard has been established in 2012: Time and Events (ISO 24617-1) and Dialogue Acts (ISO 24617-2); for two more areas a standard is about to be published: Semantic Roles (ISO 24617-4) and Spatial Information (ISO 224617-7). In the area of discourse relations a project has been started which is still in a relatively early stage and in which it is hoped to work closely together with the TextLink action. So far, a number of choices have been made concerned the scope of the project and its general direction, based on a study of existing analysis frameworks and annotation efforts. An initial set of 'core semantic discourse relations' has been identified and is provided with carefully formulated definitions, following an existing terminology standard. This allows detailed comparisons with for example RST analyses and the annotations in the Penn Discourse Treebank. In this talk I will present the current state of the project.

A Neo-Humean Taxonomy of Coherence Relations

Andrew Kehler

University of California San Diego

In his Inquiry Concerning Human Understanding, David Hume identified three types of associative links by which ideas in the mind are connected: Resemblance, Contiguity in time/space, and Cause-Effect. I will briefly summarize an inventory of coherence relations, mostly due to Hobbs (1990), that instantiate these three categories, and that I have used in linguistic and psycholinguistic analyses over the last two decades. Advantages and limitations will be discussed.

References

Hobbs, J. R. (1990). *Literature and Cognition*. CSLI Lecture Notes 21, Stanford, CA.

Cognitive plausibility and a systematic set of relations – Useful for discourse annotation?

Ted J.M. Sanders

Universiteit Utrecht

Discourse coherence can be characterized in terms of the relations that hold between clauses: coherence relations. In recent years, we have seen how corpora of language use are annotated at the level of coherence relations. Excellent annotation systems exist, such as Penn Treebank and Rhetorical Structure Theory. A major goal of the COST-TEXTLINK-project is to use and develop viable annotation systems of coherence relations, which are empirically and cognitively sound.

I will focus on the issue of cognitive plausibility. I will argue that it is attractive to take a cognitive approach to coherence relations, and that this idea can be corroborated with empirical research, by looking at different types of converging evidence: cross-linguistic analyses of connectives, as well acquisition data and results from studies on discourse processing and representation.

For instance, languages of the world provide their speakers with means to indicate causal relationships. Causal relations can be expressed by connectives and lexical cue phrases, such as *because*, *since*, *so* and *As a result.* Striving for converging evidence, we may ask about these phenomena: What is the system behind the use of such connectives in languages like English, French, Dutch and German, or Mandarin Chinese? How can we describe these systems in a cognitively plausible way? How do children acquire this connective system? And what is the role of these causal relations and connectives in discourse processing? Based on the results, I will suggest

that Causality and Subjectivity should be considered salient categorizing principles. In other coherence relations, similar systematical categories can be distinguished.

The ultimate question then is: are such insights useful in developing and using sound systems for discourse annotation? I will argue why a systematic set may be useful and beneficial, and that even non-expert analysts may be able to use it.

Reliable annotation in RST: Segmentation, nuclearity, relations and signalling

Maite Taboada

Simon Fraser University

Annotating Rhetorical Structure Theory (Mann and Thompson, 1988) relations involves segmenting the text, deciding on nucleus-satellite status, and labelling relations (the latter two being connected). I will discuss the main principles behind a plausible and reliable annotation in RST, and will outline a newly created method for assessing inter-annotator reliability (Iruskieta et al., to appear). In addition, I will provide some detail on a recent annotation effort aimed at including signalling information for RST relations (Taboada and Das, 2013). By signalling we mean indicators that a relation is present, and those include 'classic' discourse markers, but also other types of signals, such as semantic relations and cohesive chains, syntactic structure or punctuation.

References

Iruskieta, M., I. da Cunha and M. Taboada (to appear) Principles of a qualitative method for rhetorical analysis evaluation: A contrastive analysis English-Spanish-Basque. *Language Resources and Evaluation*.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text, 8*(3), 243-281.

Taboada, M. and D. Das (2013) Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse* 4 (2): 249-281.

PDTB-style Annotation of Discourse Relations: Principles, Benefits, and New Directions

Bonnie Webber

University of Edinburgh

Two principles underlie PDTB-style annotation of discourse relations: (1) No commitment to any structure above individual discourse relations, and (2) a commitment to discourse relations being grounded in particular lexical items (words or phrases) or syntactic constructions. Each principle has

its benefits. The first invites experimentation with various approaches to higher-level discourse structure. The second (lexico-syntactic grounding) was aimed at making annotation more reliable, but has brought benefits in terms of interesting new questions (about the possibility of multiple concurrent discourse relations, about constraints on the arguments to particular discourse relations) and in terms of features available to Language Technology developers working on automated discourse relation identification or on machine translation.

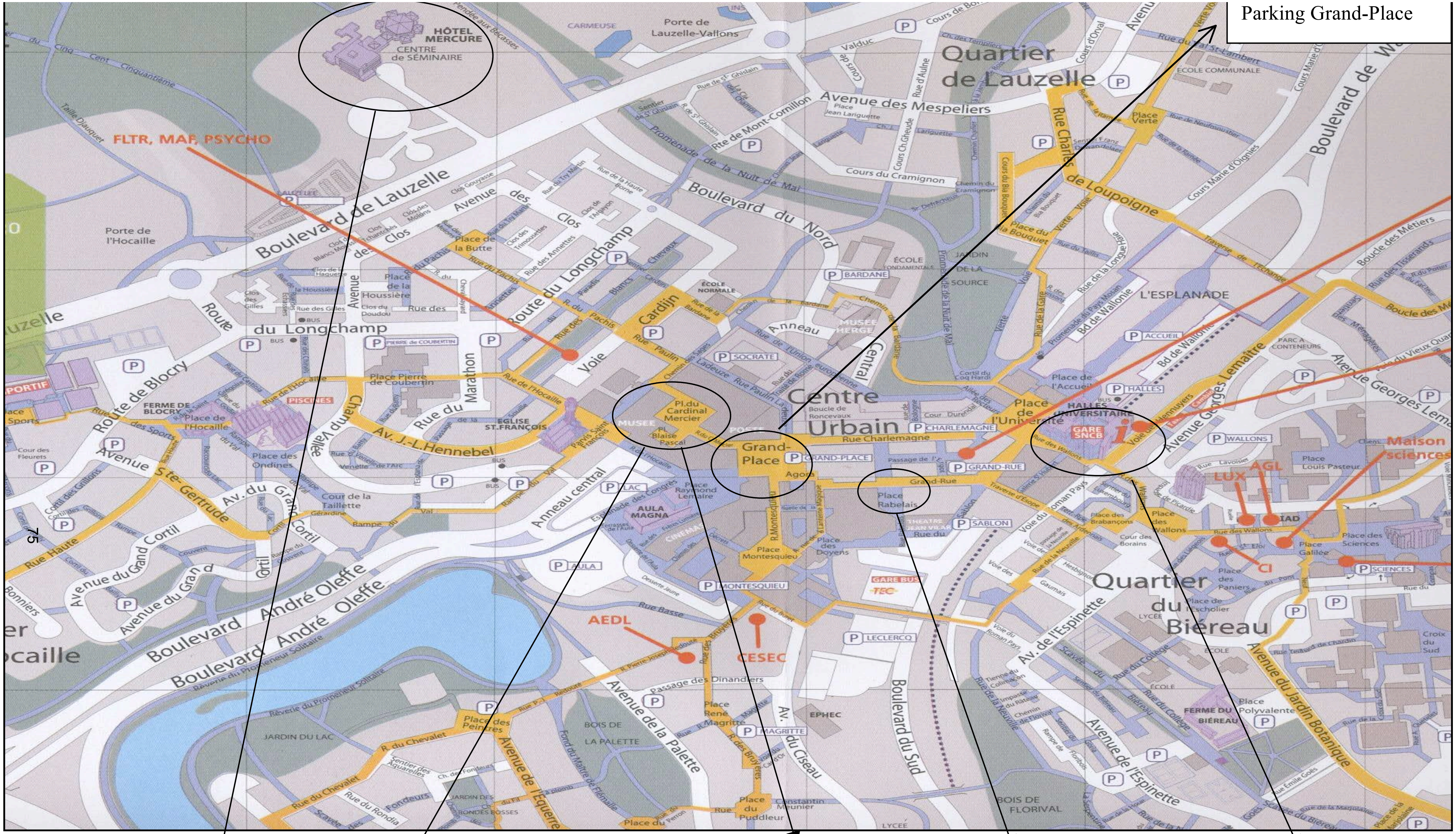
I will elaborate on these points in my presentation, including how the second principle has been adapted to the needs of languages other than English. I will describe some new directions we are taking, and conclude by revisiting our aim of reliable discourse relation annotation, in order to note some new work we are starting on detecting potential inconsistencies in discourse annotation.

List of Participants

Name	Affiliation	Email
Adam Wyner	University of Aberdeen	adam@wyner.info
Agnes Abuczki	Hungarian Academy of Sciences	abuczki.agnes@gmail.com
Ahmet Faruk Acar	Informatics Institute, Ankara	acarahmetfaruk@gmail.com
Alan Lee	Lexicon Tree	alanlee@lexicontree.org
Alexandra Fodor	Budapest Business School	sanfodor@gmail.com
Amalia Mendes	Universidade de Lisboa	amalia.mendes@clul.ul.pt
Anaig Penault	Aix Marseille Université	anaig.penault@gmail.com
Andy Kehler	UC San Diego	akehler@ucsd.edu
Anna Nedoluzhko	Charles University in Prague	nedoluzko@ufal.mff.cuni.cz
Antigoni Parmaxi	Cyprus University of Technology	antigoni.parmaxi@gmail.com
Ariadna Stefanescu	Universitatea Bucuresti	ariadna.stefanescu@gmail.com
Asad Sayeed	Saarland University	asayeed@mbl.ca
Ayisigi Basak Sevdik Calli	Informatics Institute, Ankara	asevdik@gmail.com
Balint Peter Furko	Karoli Gaspar University of the Reformed Church in Hungary	furko.peter@gmail.com
Barbara Lewandowska-Tomaszczyk	University of Lodz	blt@uni.lodz.pl
Bonnie Webber	University of Edinburgh	bonnie.webber@ed.ac.uk
Carl Vogel	Trinity College Dublin	vogel@tcd.ie
Catherine Bolly	Université Catholique de Louvain	catherine.bolly@uclouvain.be
Daniel Hardt	Copenhagen Business School	dh.itm@cbs.dk
Deniz Zeyrek	Informatics Institute, Ankara	dezeyrek@metu.edu.tr
Dijana Curkovic	Institute Institute of Croatian language and linguistics	dcurkov@ihjj.hr
Ekaterina Lapshinova-Koltunski	Saarland University	e.lapshinova@mx.uni-saarland.de
Eva Hajicova	Charles University in Prague	hajicova@ufal.mff.cuni.cz
Farah Benamara	IRIT, Toulouse	farah.benamara@irit.fr
Fatemeh Torabi Asr	Saarland University	fatemeh@coli.uni-saarland.de
Filip Ginter	University of Turku	figint@utu.fi
Florian Pusse	Saarland University	florian.pusse@freenet.de
George Christodoulides	Université Catholique de Louvain	george@mycontent.gr
Gulia Barreca	University of Paris Ouest Nanterre La Défense	barreca.gl@gmail.com

Name	Affiliation	Email
Hannah Rohde	University of Edinburgh	hannah.rohde@ed.ac.uk
Harry Bunt	Tilburg University	harry.bunt@uvt.nl
Hilde Hasselgard	University of Oslo	hilde.hasselgard@ilos.uio.no
Ines Recio Fernandez	Universität Heidelberg	ines.recio@iued.uni-heidelberg.de
Jacqueline Evers-Vermeul	Universiteit Utrecht	j.evers@uu.nl
Jacqueline Visconti	University of Genoa	j.visconti@unige.it
Jacques Steinlin	Inria, Paris	jacques.steinlin@gmail.com
Jenna Kanerva	University of Turku	jmnybl@utu.fi
Jet Hoek	Universiteit Utrecht	j.hoek@uu.nl
Jiři Mirovsky	Charles University in Prague	mirovsky@ufal.mff.cuni.cz
Jyrki Kalliokoski	Helsingin Yliopisto	jyrki.kalliokoski@helsinki.fi
Kateřina Rysova	Charles University in Prague	rysova@ufal.mff.cuni.cz
Kerstin Kunz	Universität Heidelberg	kerstin.kunz@iued.uni-heidelberg.de
Laurence Danlos	Université Paris Diderot	laurence.danlos@inria.fr
Laurence Meurant	University of Namur	laurence.meurant@unamur.be
Laurent Prevot	Aix Marseille Université	laurent.prevot@lpl-aix.fr
Liesbeth Degand	Université Catholique de Louvain	liesbeth.degand@uclouvain.be
Luciana Avila	Centro de Ciências Humanas, Letras e Artes	lucianabeatrizavila@gmail.com
Lucie Polakova	Charles University in Prague	polakova@ufal.mff.cuni.cz
Ludivine Crible	Université Catholique de Louvain	ludivine.crible@uclouvain.be
Lydia-Mai Ho-Dac	University of Toulouse Jean Jaurès	hodaclm@gmail.com
Maciej Ogrodniczuk	Polish Academy of Sciences	maciej.ogrodniczuk@ipipan.waw.pl
Magdalena Rysova	Charles University in Prague	magdalena.rysova@ufal.mff.cuni.cz
Maite Dupont	Université Catholique de Louvain	maite.dupont@uclouvain.be
Maite Taboada	Simon Fraser University	mtaboada@sfu.ca
Manfred Stede	Universität Potsdam	stede@uni-potsdam.de
Margot Colinet	Université Paris Diderot	margotcolinet@gmail.com
Maria Josep Cuenca	Facultat de Filologia, Traducció i Interpretació	maria.j.cuenca@uv.es
Maria Lada	CSRI, Athens	mlada@csri.gr

Name	Affiliation	Email
Marie-Francine Moens	Katholieke Universiteit Leuven	sien.moens@cs.kuleuven.be
Martin Groen	Universiteit Utrecht	m.g.m.groen@uu.nl
Mauro Dragoni	Fondazione Bruno Kessler, Trento	dragoni@fbk.eu
Merel Scholman	Universiteit Utrecht	m.c.j.scholman@uu.nl
Mihai Dascalu	University Politehnica of Bucharest	mihai.dascalu@cs.pub.ro
Nicholas Asher	CNRS Toulouse	nicholas.asher@irit.fr
Nicola Thrupp	Université Catholique de Louvain	nicola.thrupp@uclouvain.be
Paraskevi Giouli	Institute for Language and Speech Processing, R.C. "Athena"	voula@ilsp.athena-innovation.gr
Paul Wilson	University of Lodz	p.wilson@psychology.bbk.ac.uk
Paulo Quaresma	University of Evora	pq@uevora.pt
Pavlina Jinova	Charles University in Prague	jinova@ufal.mff.cuni.cz
Philippe Muller	Toulouse University	philippe.muller@irit.fr
Pierre Lejeune	Universidade de Lisboa	lejeunepierre@hotmail.com
Piotr Pezik	University of Lodz	piotr.pezik@gmail.com
Ruken Cakici	Middle East Technical University	ruken@ceng.metu.edu.tr
Sandrine Zufferey	Université de Fribourg	sandrine.zufferey@unifr.ch
Silvia Gabarro-Lopez	University of Namur	silvia.gabarro@unamur.be
Sorina Postolea	Al.I. Cuza University of Iasi	sorinapostolea@gmail.com
Stefan Trausan-Matu	Romanian Academy Research Institute for Artificial Intelligence	trausan@gmail.com
Stergos Afantenos	IRIT, Toulouse	stergos.afantenos@irit.fr
Tatjana Scheffler	Universität Potsdam	tatjana.scheffler@uni-potsdam.de
Ted Sanders	Universiteit Utrecht	t.j.m.sanders@uu.nl
Teresa Goncalves	University of Evora	tcg@uevora.pt
Teresa Sunol	Universitat Pompeu Fabra	tsunol@yahoo.es
Utku Sirin	Ecole Polytechnique Federale de Lausanne	utkusirin@gmail.com
Vera Demberg	Saarland University	vera@coli.uni-saarland.de
Veronika Laippala	University of Turku	veronika.laippala@utu.fi
Volker Gast	Friedrich Schiller University	volker.gast@uni-jena.de
Yael Maschler	University of Haifa	maschler@research.haifa.ac.il
Yannick Versley	Ruprecht-Karls-Universität Heidelberg	versley@cl.uni-heidelberg.de
Yipu Wei	Universiteit Utrecht	y.wei1@uu.nl



Hôtel Ibis Styles
Boulevard de Lauzelle 61
1348 Louvain-la-Neuve
010/45 07 51

Bâtiment SOCRATE
Collège Mercier (SOCR 25, 26, 27 et 28)
Collège Michotte (SOCR 11)
Place Cardinal Mercier 14 - 10
1348 Louvain-La-Neuve

Faculté de philosophie, arts et lettres
Collège Érasme (Salle du conseil)
Place Blaise Pascal 1
1348 Louvain-La-Neuve
Café et Lunch

La Fleur de Sel
Place Rabelais
1348 Louvain-La-Neuve
Conference dinner

Train station : Gare SNCB
Louvain-la-Neuve-Université

This conference has been funded and supported by: