CONFERENCE HANDBOOK

TextLink – Structuring Discourse in Multilingual Europe
Second Action Conference
Károli Gáspár University of the Reformed Church in Hungary
Budapest, 11–14 April, 2016

Debreceni Egyetemi Kiadó
Debrecen University Press
2016

Editors:
Liesbeth Degand,
Csilla Dér,
Péter Furkó,
Bonnie Webber

This volume is based upon work from COST Action IS1312 (TextLink), supported by COST (European Cooperation in Science and Technology). COST is a pan-European intergovernmental framework. Its mission is to enable break-through scientific and technological developments leading to new concepts and products and thereby contribute to strengthening Europe's research and innovation capacities.

ISBN 978-963-318-563-6

Kiadta a Debreceni Egyetemi Kiadó, az 1795-ben alapított
Magyar Könyvkiadók és Könyvterjesztők Egyesülésének a tagja
www.dupress.hu
Felelős kiadó: Karácsony Gyöngyi
Műszaki vezető: M. Szabó Monika
Készült a Debreceni Egyetemi Kiadó nyomdaüzemében, 2016-ban

Contents

GENERAL INFORMATION ABOUT THE CONFERENCE

The TextLink COST Action addresses Discourse Relational Devices (DRDs) in terms of resources, annotation models (including their comparability), and tools both for annotating DRDs and interconnecting annotated data. With a network covering research on no less than 20 different languages, written as well as spoken discourse, in a variety of genres and registers, and corpora that range from „in construction" to „fully annotated", the second Action Conference will constitute a milestone in (re)defining the objectives that need to be reached in view of constructing a TextLink Portal that make these resources and tools more widely known and available.

The meeting is open to everyone, both current members of TextLink and other researchers and practitioners working in the area.

In addition to the main meeting, there will be a WorkGroup2/WG3 workshop on 13–14 April 2016, focusing on the comparison of annotation schemes and their interfacing.

Invited speakers:

Andrei Popescu-Belis	(IDIAP Research institute, Martigny, Switzerland)
Nina Vyatkina	(University of Kansas, Lawrence KS, United States)

Program Committee:

Maria Josep Cuenca	(University of Valencia)
Liesbeth Degand	(Université catholique de Louvain)
Peter Furkó	(Károli Gáspár University of the Reformed Church in Hungary)
Daniel Hardt	(Copenhagen Business School)
Jiří Mírovský	(Charles University in Prague)
Philippe Muller	(University of Toulouse)
Piotr Pezik	(University of Łódź)
Hannah Rohde	(University of Edinburgh)
Ted Sanders	(Utrecht University)

Manfred Stede	(Potsdam University)
Jacqueline Visconti	(University of Genoa)
Bonnie Webber	(University of Edinburgh)
Deniz Zeyrek	(Middle East Technical University)
Sandrine Zufferey	(University of Bern)

Organisers:

Péter Furkó	(Károli Gáspár University of the Reformed Church in Hungary, Budapest)
Liesbeth Degand	(Université catholique de Louvain)
Csilla Dér	(Károli Gáspár University of the Reformed Church in Hungary, Budapest)
Judit Nagy	(Károli Gáspár University of the Reformed Church in Hungary, Budapest)
Alexandra Fodor	(Budapest Business School–University of Applied Sciences)
Ágnes Abuczki	(MTA-DE Research Group)
Anna Nagy	(University of Debrecen)

Date:
Monday, 11 April, 2016 to Thursday, 14 April, 2016

Venues:

Main Building	25–27 Dózsa György Street, Budapest, Hungary, H–1146
Synod Building	16 Szabó József Street, Budapest, Hungary, H–1146

Getting to the Synod from the main building (see also Appendix)

Sessions marked as SYNOD (all plenaries and some of the talks) will be held in an off-site building of the Synod of the Hungarian Reformed Church (16 Szabó József Street). Turn left as you leave the main building (25 Dózsa György Street) and head south-east towards Abonyi utca. Turn left onto Abonyi utca and walk about 400 ms. When you reach Szabó József utca turn right, and you'll find the entrance of the building at number 16. The auditorium is on the first floor.

CONFERENCE SCHEDULE

Monday 11 April 2016

08.30–10.30 **Steering Committee meeting – Main Building Room 24**

11.00–13.00 **MC meeting – Main Building Room 24**
(non-MC members are welcome to attend without voting rights)

13.00–14.15 **Lunch – Main Building Rooms 9 and 11**

14.15–14.30 **Welcome / Conference Opening – Synod Building**
Prof. Liesbeth Degand, Chair of TextLink Action
Dr. Enikő Sepsi, Dean, Károli Gáspár University of the Reformed Church in Hungary

14.30–15.30 **Plenary 1 – Synod Building**
Manual and Automatic Labeling of Discourse Connectives for Machine Translation
Andrei Popescu-Belis, Idiap Research Institute

15.30–16.00 **Coffee – Main Building Rooms 9 and 11**

16.00–18.00 **Poster session 1 – Main Building Rooms 24 and 20**
Automatic prediction of implicit discourse relations in Turkish
Murathan Kurfalı, Deniz Zeyrek and Teresa Gonçalves

Analyzing discourse relational devices: quantitative and qualitative perspectives
Johannes Angermüller and Péter Furkó

Discourse Treebanks in a Graph Database
Laurence Danlos and Pierre Magistry

A new approach to merging and querying parallel text annotations
Arne Neumann, Uladzimir Sidarenka and Manfred Stede

Sentence-initial discourse markers in the Finnish Internet
Veronika Laippala, Aki-Juhani Kyröläinen, Johanna Komppa, Maria Vilkuna and Jyrki Kalliokoski, Filip Ginter

Towards Discourse Parsing in Spanish
Iria Da Cunha

Comparing Discourse Annotation Schemes from an NLP Perspective
Chloé Braud

Markers of mirativity in Hungarian
Ildikó Vaskó

19.00– **Social dinner – Építészpince (Builders' Vault)**
1088 Ötpacsirta utca 2.

Tuesday 12 April 2016

09.00–11.00 **Synod Building**
Oral presentations to kick-off discussion on: Annotation Dimensions and their Inter-dependencies

Interdependence of annotation levels in a functional taxonomy for discourse markers in spoken corpora
Ludivine Crible, Liesbeth Degand and Anne-Catherine Simon

On temporality in discourse annotation
Jacqueline Evers-Vermeul, Jet Hoek and Merel Scholman

How can we relate various annotation schemes? Unifying Dimensions in Discourse Relations
Ted Sanders, Vera Demberg, Jacqueline Evers-Vermeul, Jet Hoek, Merel Scholman and Sandrine Zufferey

11.00–11.45 **Coffee – Synod Building**

11.45–12.45 **Plenary 2 – Synod Building**
What can multilingual discourse-annotated corpora do for language learning and teaching
Nina Vyatkina, University of Kansas

12.45–14.00 **Lunch – Main Building Rooms 9 and 11**

14.00–16.00 **Poster session 2 – Main Building Room 24**

Putting things together: Correlating discourse relations with other types of linguistic data
Eva Hajicova, Barbora Hladka, Pavlina Jinova and Sarka Zikanova

Referring expressions as cohesive devices in multiple languages
Yulia Grishina and Manfred Stede

How do expectations based on contextual signals guide the processing of additive and causal relations?
Merel Scholman, Pim Mak and Ted Sanders

Discourse segmentation and ambiguity in discourse structure
Jet Hoek, Jacqueline Evers-Vermeul and Ted Sanders

Automatic Prosodic Annotation for DRD Analysis
István Szekrényes

Some considerations on the use of main verbs to express rhetorical relations
Amália Mendes, Pierre Lejeune and Nuno Martins

An annotation by speech act labels for dialogue discourse analysis in Japanese, French and English
Mutsuko Tomokiyo

Categories and Annotation of Negative Emotionality Discourse Markers in Spoken Language
Barbara Lewandowska-Tomaszczyk and Paul A. Wilson

16.00–16.30 **Coffee – Main Building Rooms 9 and 11**

16.30–18.30 **Synod Building**

Oral presentations to kick-off discussion on: DRDs and (other) Evidence for Discourse Relations

Several Observations from the Annotation of Discourse Connectives in the Prague Dependency Treebank
Katerina Rysova, Eva Hajicova, Magdalena Rysova and Jiri Mirovsky

A Discourse Annotated Corpus of Conjoined VPs
Bonnie Webber, Rashmi Prasad, Alan Lee and Aravind Joshi

Data-driven discourse markers representation and classification
Juliette Conrath, Philippe Muller, Stergos Afantenos and Nicholas Asher

20.00– **Social program: cruise along the Danube – Port (harbour) at Jászai Mari tér**

Wednesday 13 April 2016

09.00–11.00 **Poster Session 3 – Main Building Room 215**

Studying the position of Discourse Relational Devices in signed languages: adapting the Basic Discourse Units Model to the signed modality
Sílvia Gabarró-López and Laurence Meurant

Discourse Structuring Devices on Twitter
Tatjana Scheffler, Rike Schlüter and Manfred Stede

Semantic weakening of discourse structuring devices
Šárka Zikánová, Liesbeth Degand, Péter Furkó, Sandrine Zufferey and Ágnes Abuczki

LDM-PT A Portuguese Lexicon of Discourse Markers
Amália Mendes and Pierre Lejeune

From Monolingual Annotations towards Cross-lingual Resources: An Interoperable Approach to the Analysis of Discourse
Ekaterina Lapshinova-Koltunski, Kerstin Anna Kunz and Anna Nedoluzhko

Discourse features for computational stylometry
Ben Verhoeven and Walter Daelemans

Annotating metadiscourse markers in the English-Spanish MULTINOT corpus: preliminary steps
Julia Lavid and Lara Moraton

Annotation of discourse units and speech ruptures in spontaneous conversations within a segmentation system Elena Pascual Aliaga

11.00–11.30 **Coffee – Main Building Rooms 9 and 11**

11.30–13.00 **Main Building Room 215**
Special focus group discussion on constructing and enriching DRD lexicons

13.00–14.30 **Closing and lunch – Main Building Rooms 9 and 11**

14.30–16.00 **WG2 and WG3 workshop – Main Building Room 215**

16.00–16.30 **Coffee – Main Building Rooms 9 and 11**

16.30–18.00 **WG2 and WG3 workshop – Main Building Room 215**

19.30– **Social program: Budapest Spring Festival Concert**
Concert hall, Liszt Ferenc tér 8.

Thursday 14 April 2016

09.00–10.30 **WG2 and WG3 workshop – Main Building Room 24**

10.30–11.00 **Coffee – Main Building Room 9**

11.00–13.00 **WG2 and WG3 workshop – Main Building Room 24**

KEYNOTE PAPERS

Andrei Popescu-Belis

Manual and Automatic Labeling of Discourse Connectives for Machine Translation

Idiap Research Institute
Rue Marconi 19, CH-1920 Martigny, Switzerland
andrei.popescu-belis@idiap.ch

Introduction

The automatic translation of discourse connectives must cope with the fact that discourse connectives are often multi-functional, in other words, a connective in a source language may signal different discourse relations upon different occurrences. In translation, these relations may be expressed by different connectives in the target language. To generate an accurate output, a machine translation (MT) system must be able to deal with such translation divergencies. The knowledge that a system can leverage to solve these translation divergencies depends on its architecture, but the pragmatic functions of discourse connectives suggest that examining only their local context cannot always enable a system to generate a correct translation. Instead, knowledge of the discourse relation signaled by a connective is the key information enabling a system to reliably generate a correct translation for each occurrence of a discourse connective.

Exploiting knowledge from discourse relations for MT requires answering three main questions. First, how should one define the range of possible discourse relations that are needed to enable MT systems to reliably produce correct outputs? Then, how can a system identify the correct relation for each occurrence of a connective, without the need for human intervention? And finally, how could this information be integrated with the other features that govern the generation of the target sentence in an MT system?

In my presentation, I will answer these questions by presenting the research results of two large collaborative projects,[1] with participants from Idiap and the Universities of Geneva, Utrecht, and Zürich. In these projects, we designed and evaluated the first operational method to improve the automatic translation of discourse connectives by leveraging text-level features to identify discourse relations, and conveying this information to a statistical MT system from English to four target languages (French, German, Italian and Arabic). In the process, the issue of defining a range of discourse relations has been addressed from a practical, application-oriented perspective.

Classification and labeling of discourse connectives

The main linguistic challenge was the definition of a range of possible discourse relations signaled by discourse connectives, bearing in mind several requirements: theoretical grounding; availability of corpora annotated with the corresponding labels (at least for testing our systems, but preferably also for training them) or, alternatively, feasibility of manual annotation within an acceptable time frame; applicability to the languages of the project (English, French, German, Italian, with priority to the first one); and tractability of automatic annotation with acceptable accuracy.

In the first phase of the project, we have examined the existing classifications of discourse relations and the corresponding connectives. While they all appeared to offer sufficient granularity among the relations that we needed, only few annotated corpora were available with these annotations (or none at all), and manual annotation was too costly to be done within our project (due to the required level of detail). The major annotated resource, in English, was the Penn Discourse Treebank (PDTB), but this resource still raised two problems: being monolingual, it did not enable MT experiments (which require parallel corpora); and the automatic annotation with PDTB labels, although already studied in the literature, was quite challenging. We even considered the possibility of annotating only translation divergencies, in other words, observing the discourse connectives which have several possible translations in parallel corpora, and annotate each target-side occurrence simply with its translation.

[1] The COMTIS and MODERN Sinergia projects, supported by the Swiss National Science Foundation.

The solution that we finally adopted[2] was halfway between a theoretically-oriented annotation scheme and a fully empirical one based on observed divergencies. First, using translation spotting by human coders, we annotated all the observed translations of several discourse connective types in parallel corpora (excerpts from Europarl, with EN/FR and EN/DE, using only texts originally written in English). The observed translations were either target connectives or other constructions. This annotation gave us a precise view of the range of possible translations of seven highly multi-functional English connectives: *although*, *however*, *meanwhile*, *since*, *(even) though*, *while* and *yet*. These translations were clustered a posteriori into "senses" (or labels) inspired from the second level of the PDTB (e.g. for *since*: 'causal', 'temporal', or 'causal/temporal'). About 2000 instances have thus been annotated and are made available.[3]

Automatic annotation of English discourse connectives

The annotated resources were used to train and test systems for automatic labeling of connectives using the set of labels we designed. The systems had access to a set of surface features for each English connective, including non-local ones, which can be extracted automatically so that the labeling can be done without human intervention. The features included state-of-the-art positional, word-level and syntactic features, as well as semantically-oriented and contextual features, such as the detection of pairs of synonyms or antonyms on each side of a connective, features from an RST discourse parser, and TimeML labels – all from the sentence containing the connective and the previous one. Several types of classifiers were trained on the Europarl and PDTB data (with an appropriate mapping of labels in the latter case), and tested on subsets that were not seen during training. The best performing classifier, the Maximum Entropy one, reached scores above or comparable to those of previous studies: the accuracies (F1 scores) vary between 0.5 and 0.9 depending on the connective and on the testing set.[4] All features were shown to be beneficial to automatic discourse relation labeling.

---

[2] Cartoni et al (2013). See also: Danlos & Roze (2011).

[3] Popescu-Belis et al. (2012). The data is available from http://www.idiap.ch/dataset/Disco-Annotation.

[4] Meyer (2015), Chapter 5; Meyer et al. (2015).

Machine translation of labeled discourse connectives and its evaluation

However, our final goal was not the labeling of discourse connectives *per se*, but the use of the automatically assigned labels in an end-to-end MT system. We determined by experimentation, using the Moses toolkit,[5] that factored translation models were the most effective solution. We trained four MT systems for translating from English into, respectively, French, German, Italian, and Arabic, and provided them, for the test data, with the labels of each connective as hypothesized by our system.[6]

To evaluate the improvement brought by the (imperfect) connective labels that were automatically assigned, the frequently-used n-gram based BLEU metric is not sensitive enough to measure the changes. Therefore, one must examine whether the translations of connectives by the modified MT system are better, or (as a proxy) closer to the reference translation, than those of an unmodified system. We therefore designed the ACT metric (standing for Accuracy of Connective Translation) to automatically compare the connectives translated by a system with those of a reference translation. Although identity with the reference is not compulsory for a translation to be correct, similarity with the reference over a large number of occurrences is a reliable indicator of quality. Depending on the test data and language pair, we found that our systems improved up to 7% of the discourse connectives in translation.

Conclusion and perspectives

We have presented an end-to-end solution to improve the automatic translation of discourse connectives, from English to four target languages. The solution relies on the automatic labeling of discourse connectives in source texts, by a classifier trained using machine learning. Its design required the definition of a practical set of labels characterizing the "senses" of discourse connectives, and the annotation of a sizeable corpus with these labels.

The method presented here is one of the first attempts to use discourse-level information to improve statistical MT. Within our consortium, we have also explored the case of verb tenses, focusing on translation divergencies of Eng-

[5] Available from www.stamt.org/moses/.

[6] Meyer (2015), Chapter 7; Meyer et al. (2015).

lish Simple Past into to French, which can be addressed through a similar approach (using French tenses as labels). We are currently exploring the use of co-reference information, either to improve pronoun choice, or more generally to constrain the translation of all noun phrases. However, the relational nature of co-referent expressions appears to be more challenging to model in order to observe an improvement over a pure statistical MT system.

References

- Cartoni, B., Zufferey, S. & Meyer, T. 2013. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue & Discourse* 4 (2): 65–86.
- Danlos, L. & Roze, C. 2011. Traduction (automatique) des connecteurs de discours. In: *Actes de la 18ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France.
- Meyer, T. 2015. *Discourse-level features for statistical machine translation*, PhD thesis, École polytechnique fédérale de Lausanne (EPFL), n. 6501.
- Meyer, T., Hajlaoui, N. & Popescu-Belis, A. 2015. Disambiguating Discourse Connectives for Statistical Machine Translation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 23 (7): 1184–1197.
- Popescu-Belis, A., Meyer, T., Liyanapathirana, J., Cartoni, B. & Zufferey, S. 2012. Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns. In: *Proceedings of the 8th Language Resources and Evaluation Conference (LREC)*, Istanbul, Turkey.

Nina Vyatkina

What can multilingual discourse-annotated corpora do for language learning and teaching?

University of Kansas, Lawrence KS, United States
vyatkina@ku.edu

Introduction

This talk explores existing and potential interfaces between Learner Corpus Research (LCR), corpus-based teaching (a.k.a. Data-Driven Learning, or DDL), discourse studies, and the TextLink annotation efforts. First, I will present the state-of-the-art of LCR that focuses on learner use of cohesive devices in their Second Language (L2). After that, I will review DDL applications for teaching cohesive devices. I will conclude by outlining future research directions in which multilingual discourse-annotated corpora can help advance LCR and DDL.

Learner corpus research on cohesive devices

The advent of learner corpora has enriched the data landscape for empirical L2 research tremendously. Most LCR studies employ the Contrastive Interlanguage Analysis method (Granger 2015). Using this method, researchers compare and contrast learner and native speaker corpora to find patterns of overuse, underuse, or misuse of target language phenomena by learners in comparison with native speakers or with other learners. Most of this research involves relatively advanced L2 learners, and most found patterns exemplify stylistically infelicitous language use rather than grammatical errors. This is especially relevant to discourse studies, as "[i]t has been demonstrated that discourse is a crucial aspect for L2 learners of a language, especially at more advanced levels" (Neff-van Aertselaer 2015: 255). Although still small in number, LCR discourse studies have addressed a wide array of phenomena, including coherence, cohesion, thematic progression, and textual rhetorical features. Cohesive devices

have attracted the most attention of LCR scholars, which is not surprising given that they feature prominently in L2 pedagogy and assessment.

Most studies conducted on cohesive devices in **written** learner corpora found that there is no uniform pattern of underuse or overuse but that their use is modulated by four main variables. The first of these variables is connector types: learners primarily use generic connectors (e.g., those expressing addition, exemplification, or emphasis) but, compared to native speakers, they underuse genre-specific connectors such as contrastive and grounding connectors used for argument. The second is the First Language (L1) transfer effect, where learners use language patterns typical of their L1. The third variable is L2 proficiency, with more proficient learners exhibiting more native-like trends, and the fourth is overall writing competence. All studies that used both expert and novice L1 corpora as baseline showed that the overuse of connectors was a general feature of novice writing. Moreover, studies that explored cohesion using Coh-Metrix, a tool that computes multiple cohesion indices for English texts (Graesser et al. 2004), found that both L1 and L2 essays that were rated as having higher quality by expert raters had lower cohesion but more modification and embedding that contributed to implicit coherence.

The majority of discourse LCR studies conducted on **spoken** corpora have focused on discourse markers. These studies invariably found that learners used both fewer tokens and fewer types of discourse markers compared to native speakers or higher proficiency learners. In addition to studies that focused exclusively on frequency, there have also been a few studies that illuminated qualitative differences between native and non-native discourse marker use. Finally, high inter-individual variation has been found in learner corpora. Within the same cohort, some learners will overuse while others will underuse one and the same discourse marker.

Corpora and language teaching

Whereas LCR studies pinpoint learner problems with cohesive devices, another research strand is the proposal and testing of pedagogical solutions for these problems. In the early indirect pedagogical applications, textbook and reference grammar authors used corpus research results for material selection. More recent years have seen an exponential growth of suggestions for direct corpus use by language teachers and learners (DDL). The first recently published meta-analysis of DDL studies (Cobb & Boulton 2015) has shown that this innovative teaching method is generally effective (i.e., it leads to significant

learning gains), and is more efficient than traditional teaching methods for certain L2 targets.

Although DDL use for instruction in discourse-pragmatic phenomena is still rare, several recent studies have applied this method for teaching English linking adverbials (LAs) to L2 learners. Authors of these studies made their decision to develop DDL interventions for LAs due to limitations of available pedagogical materials such as simplified models and mismatches between corpus and textbook frequencies of different LA types. These studies have shown that: 1) learners acquire LAs better while using printed concordance lines as reference materials than while using a bilingual dictionary or a grammar manual; 2) learners who directly search corpora for LAs for several weeks use them with higher frequency and accuracy than learners who spend an equal amount of time learning with traditional methods; and 3) learners who compare the use of LAs in a native speaker corpus and a learner corpus of their own writing improve LA frequency, diversity, and accuracy more than learners who only work with native corpora.

Avenues for future interdisciplinary collaboration

In summary, LCR and DDL studies provide valuable insights into L2 discourse characteristics, problems, and solutions. However, there are a number of salient limitations evident from this brief overview that can serve as a sketch of promising avenues for future collaboration between L1 and L2 discourse and corpus analysts.

First, all discourse LCR studies have so far been 'word-based' rather than 'category-based' (Hunston 2002) because none of the analyzed corpora were annotated for discourse-pragmatic categories. Researchers habitually use the Longman Grammar (Biber et al. 1999) list of connectors that fulfill different cohesive functions to identify focal phenomena in their corpus data. This also entails that the scope of the available research has been limited to rather crudely defined discourse categories. If robust fine-grained taxonomies for manual discourse annotation are developed for multilingual L1 corpora, they can serve as models for annotation of learner corpora.

In the same vein, if automatic discourse taggers and corpus analysis tools are developed for L1 corpora, they can be applied to learner corpora as well. As an example, the Coh-Metrix tool has been successfully applied to both L1 and L2 English corpora, so similar tools can and should be developed for other languages. Such applications could work even with texts produced by less pro-

ficient learners if learner corpora are error-corrected. The discrepancies in the tagger output from the learner language layer and the error correction layer would then provide rich material for analyzing unique characteristics of learner language – which can be interpreted as either deficiencies or manifestations of learner creativity and agency, or both.

Next, access (especially public and free access) to multilingual corpora would provide an unprecedented boost to multidimensional analyses of learner texts to compare them not only to the target language texts but also to texts in the learners' native language (Neff-van Aertselaer 2015) produced by both expert and novice speakers to gain valuable insights into L1 transfer effects as well as overall writing and speaking competence effects.

Finally, discourse-annotated corpora – especially corpora of languages other than English and oral and multimedia corpora - would present an invaluable supplement to the currently insufficient and inadequate teaching materials on L2 discourse that can be used in DDL applications. Such applications would help reduce teaching-induced errors and help the learners become "not only more native-like, but also more expert-like" (Leńko-Szymańska 2008: 106).

References

- Biber, D., Johansson, St., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.
- Cobb, Th. & Boulton, A. 2015. Classroom Applications of Corpus Analysis, in D. Biber–R. Reppen (ed.), The Cambridge Handbook of English Corpus Linguistics. Cambridge, Cambridge University Press, 2015, 478–497.
- Graesser, A. C.–McNamara, D. S., Louwerse, M. M. & Cai, Zh. 2004. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, & Computers* 36: 193–202.
- Granger, S. 2015. Contrastive Interlanguage Analysis: A Reappraisal. *International Journal of Learner Corpus Research* 1: 7–24.
- Hunston, S. 2002. Corpora in Applied Linguistics, Cambridge: Cambridge University Press.
- Leńko-Symańska, A. 2008. Non-Native or Non-Expert? The Use of Connectors in Native and Foreign Language Learners' Texts. *Acquisition et Interaction en Langue Étrangère* 27: 91–108.
- Neff-van Aertselaer, J. 2015. Learner Corpora and Discourse. In: Granger, S. Gilquin, G. & Meunier, F. (eds) *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press. 255–279.

REGULAR PAPERS

Johannes Angermüller[a], Péter Furkó[b]

Analyzing discourse relational devices: quantitative and qualitative perspectives

[a]University of Warwick; [b]Károli Gáspár University of the Reformed Church in Hungary
J.Angermuller@warwick.ac.uk; furko.peter@gmail.com

If discourse research accounts for the (social) uses of text and talk, corpus analysis is one of the preferred methodologies in this field at the crossroads of language and society. Corpus analysts typically aim to analyse large collections of mostly written texts with the aid of computers. In the research, quantitative aspects need to be articulated with qualitative aspects. A number of solutions have been proposed to deal with typical problems of such an integration.

Accordingly, in the first part of the paper, we will discuss some of the questions, problems and solutions that discourse researchers from various disciplinary background have addressed against the background of the DiscourseNet network (cf. http://www.discourseanalysis.net) and Discourse Studies more generally.

In the second part of the paper, we will propose a methodology for the analysis of reformulation markers which combines qualitative and quantitative approaches.

The most challenging task for discourse annotators is to tag a set of highly frequent DRDs such as *well, you know, I mean, I think* etc., which are used in a wide range of contexts with numerous (discourse-relational as well as interpersonal) functions, rather than DRDs such as *in other words, or rather, in short,* etc. which mark more explicit relations between discourse segments and are used with higher type/token ratios. Our paper will focus on *I mean* and its German and Hungarian counterparts. There is general agreement in the literature that a contrastive analysis can help tease out the diversity of meaning relations that semantically bleached DMs such as *I mean* mark (cf. Mortier & Degand, 2009), while there is also an increasing awareness of the "indexically rich" situ-

ational meaning of DMs (Aijmer, 2013) and the resulting need to analyse DMs across a variety of speech situations and genres.

Moreover, Fetzer (2014: 70) argues for an integrated analysis where quantitative methodology is combined with qualitatively oriented interactional and pragmatic approaches (e.g. conversation analysis, speech act theory and interactional sociolinguistics), thereby "supplementing frequency and distribution of linguistic form with patterned co-occurrences and pragmatic function" (ibid.).

Our methodology also aims at clearing up some of the confusion between discourse relational and interpersonal functions characterised by many ESL textbooks. For example, in Crystal and Davy's (1975) classical account, the main function of *I mean* is in clarifying the meaning of the speaker's immediately preceding expression, marking a restatement of the previous utterance, providing extra information and/or a fresh angle about a previous topic as well as marking a change of mind. Swan (1997) argues that *I mean* introduces explanations, additional details, expressions of opinion and corrections, while it can also serve softening functions or as "a general-purpose connector of 'filler' with little real meaning" (1997: 159).

In order to meet the above demands, our study will map the functional spectrum of 92 tokens of *I mean* in a one-million-word English-German-Hungarian parallel corpus which is based on the dramatized dialogues in *House, M. D.* season one.

The source language DMs have been aligned with the target language lexical items (alternatively, the absence of a translation equivalent has been noted), while individual tokens of *I mean* have been annotated for the following formal properties:

- DM clusters, collocations, lexical co-occurrence patterns in both left/right co-text;
- co-occurrence with pauses and false starts;
- type of host unit (default structure, focus structure, imperative, question);
- position in utterance (initial, medial, final);
- position in the turn (initial, medial, final);
- the host unit's position in conversational structure (first / second part of an adjacency pair, embedded sequence).

In terms of functional properties, the following features have been annotated:

- coherence relation between host unit and previous discourse unit (1 implicit, 2 explicit);

- primary (discourse-relational) functions (1 addition, 2 contrast/concession, 3 explanation / elaboration, 4 reformulation);
- secondary (interpersonal and interactional) functions (0 not clear, 1 emphasis, 2 hedge, 3 irony);
- translation equivalents in Hungarian and German (1 lexical item, 2 morpho-syntactic form, 3 not translated);
- scope of *I mean* (1 wide scope – whole utterance/turn, 2 narrow scope – single word or phrase).

The results suggest that DM clusters, scope and target language items are often reliable indicators of the textual and interpersonal functions *I mean* marks, while positional features, genre, and type of host unit and/or adjacent DU might also help us differentiate between discourse relational and interpersonal functions.

References

- Aijmer, K. 2013. *Understanding Pragmatic Markers – A Variational Pragmatic Approach*. Edinburgh: Edinburgh University Press.
- Crystal, D. & Davy, D. 1975. *Advanced Conversational English*. London: Longman.
- Fetzer, A. 2014. *I think, I mean* and *I believe* in political discourse: Collocates, functions and distribution. *Functions of Language* 21(1): 67–94.
- Mortier, L. & Degand L. 2009. Adversative discourse markers in contrast. *International Journal of Corpus Linguistics* 14(3): 339–66.
- Swan, M. 1997. *Practical English Usage*. Oxford: Oxford University Press.

Chloé Braud

Comparing Discourse Annotation Schemes from an NLP Perspective

University of Copenhagen
chloe.braud@gmail.com

The segments of text in a document are linked by discourse relations (such as *Explanation or Contrast*) in order to form a coherent ensemble. A current NLP challenge is to design systems that automatically perform discourse analysis. Recent efforts have been focused on building systems from three corpora: the RST Discourse Treebank (RST DT, Carlson et al. 2001) and the Penn Discourse Treebank (PDTB, Rashmi et al. 2008) for English, and Annodis (Stergos et al. 2012) for French. They represent the three main annotation schemes, leading to new projects for other languages or domains. They differ on various aspects because of their primary objectives, underlying frameworks, or some specific choices. In this paper, we want to discuss differences that prevent from any straightforward combination of these corpora and highlight the impact of their specific features on automatic systems.

A first difference lies in the relation set they defined: there are 78 relations in the RST DT, 18 in Annodis and, in the PDTB, a three-level hierarchy with 4 classes, 16 types and 23 subtypes – the levels 1 and 2 being the most used. The difference in terms of size mostly comes from the criteria used to define the relations, linked to the underlying framework for the RST DT and Annodis (resp. RST and SDRT, Mann & Thompson 1988, Asher & Lascarides 2003). The recent parsers (Joty et al. 2012, Feng & Hirst 2014, Ji & Eisenstein 2014) developed on the RST DT, however, use a smaller set of 18 relations (Carlson & Marcu 2001), thus of similar size, because the statistical approaches implemented can hardly distinguish between so many relations, with possibly skewness issue, using a rather small dataset. There are commonalities: all these relation sets involve temporal, causal, conditional, additive and comparative relations. Therefore, it would be possible to construct a system combining the datasets that only uses these coarse-grained classes. But such a system could be not as useful for downstream applications, since it would miss likely crucial fine-grained distinctions.

There are, however, fundamental differences for finer grained relations. Some relations are specific: only the RST DT set includes the relations Topic-Change and Topic-Comment – that could in fact be viewed as another layer of discourse annotation (Webber et al. 2011) – or the relation Same-Unit – that is not rhetorical; only Annodis uses Frame, a relation linked to a specific segmentation choice (i.e. sentence initial adverbial). Furthermore, in the PDTB, Entity relations are annotated apart from the rhetorical ones, which led to some confusion for the automatic systems that chose either to exclude (as non rhetorical) or to include them in the class Expansion (but with no solution at level 2), maybe because some of them are annotated as Elaboration in the RST DT. Finally, the merging made for the RST DT erases the differences on temporal ordering, and the level 2 relations of the PDTB only keep the distinction between the synchrony and the asynchrony, losing the distinction between Precedence and Succession corresponding to Flashback and Narration in Annodis. From now on, it may thus be interesting to give results on shareable mappings (Benamara & Taboada 2015, Prasad & Bunt 2015), in order to compare systems built on different corpora. Furthermore, a common relation set opens ways to cross-corpus studies if, of course, we are able to deal with the other differences, including the structure and the segmentation. In particular, the massive amount of data annotated in the PDTB should be useful to improve on tasks, such as discourse parsing, or languages with fewer resources.

Concerning the relations, another choice raising interesting issues from an NLP perspective is the possibility, theoretically justified, of annotating multiple senses between two spans of text in Annodis and the PDTB. The first systems dedicated to the identification of implicit relations in the PDTB kept all the annotated labels using an evaluation metric counting a correct match if one of the possible label is identified (Lin et al. 2009). A more natural way would be to use multi-label classification algorithms (Tsoumakas & Katakis 2006), thus trying to identify all the labels associated with a pair of segments, similar to what is done for document classification. However, very few examples have multiple labels (4.3% of all examples in the PDTB), making it impossible the use of such algorithm. Recent studies simply chose to select the first annotated sense (Wang et al. 2012) leading to a cleaner evaluation but lacking a proper justification. It is therefore important to determine if multiple annotations have been under annotated (Miltsakaki et al. 2005) and if we can expect more data in future releases, or if we need to make *a priori* decisions on the relations we want to identify in the way the RST DT manual defines preferences on the annotated relations (Carlson & Marcu 2001b). Finally, the underspecified labels allowed in

the PDTB, making the annotation easier, are also hard to take into account into classification systems: it has been proposed either to simply discard them (e.g. removing examples annotated at level 1 when working at level 2) – it is probably the most widespread strategy –, to duplicate the examples according to their subclasses (Versley 2011), leading to instances annotated with different labels – an issue for a classification system –, or to keep all labels as they are, thus enlarging the label set with even more skewness (Xue et al. 2015). None of these solutions seems completely satisfactory and more work is probably needed to understand this phenomena.

Another important difference between the corpora is that each document in the RST DT and Annodis is paired with a structure covering the entire text (resp. a tree or a graph), whereas the annotation in the PDTB follows a theory-neutral approach not requiring a full coverage. Concerning automatic systems, this distinction led to the development of two concurrent distinct tasks: discourse parsing for building a tree or graph over a document and discourse chunking for retrieving PDTB-like annotations. Discourse parsing has clearly attracted more attention with incremental improvements over the years. The creation of a new shared task (Xue et al. 2015) has put, however, a new emphasis on end-to-end discourse chunking. Surprisingly, discourse parsing seems easier than chunking, with performance around 36% (Annodis, see Muller et al. 2012) and 61%[7] (RST DT) for the former and 33% for the latter (Lin et al. 2010). Whereas retrieving a full embedding of the spans of text could seem harder, and despite the much higher number of documents annotated in the PDTB, the task of matching the PDTB annotations could be made harder by some segmentation choices. The low scores could also be due to the fact that no joint solution has been proposed yet: whereas, on the RST DT, algorithms and methods inspired by syntactic parsing have led to improvements, end-to-end discourse chunkers rely on a pipeline of modules dedicated to specific tasks (i.e. identifying the connectives, their arguments, their senses, and the senses of non explicit relations). Studies identifying the kind of structures annotated in the PDTB (Lee et al. 2006) could lead to a better understanding and thus to the development of more informed solutions. Studies trying to cast existing structures from RST, SDRT and D-LTAG into a single formalism (Venant et al. 2013) are also interesting, and the use of similar dependency parsing algorithms for building discourse parsers on the RST DT (Li et al. 2014) and Annodis (Muller

[7] Ji & Eisenstein 2014. Note that one can expect about 7% loss considering an automatic segmentation.

et al. 2012) creates a way for cross corpora and language systems. Now, it could also be interesting to see if the RST DT and PDTB annotations could be used together, for instance by making discourse chunking from the RST DT data. This raises the question of the usefulness of discourse chunking for downstream applications. It has been shown that RST trees are useful for many applications (Taboada & Mann 2006a). But discourse chunking has also been proved useful for summarization (Louis et al. 2010, Lin et al. 2012), automatic translation (Meyer & Popescu-Belis 2012, Meyer & Webber 2013) or sentence compression (Sporleder & Lapata 2005). This tends to show that a covering structure may not be required, and more studies need to be done to evaluate this assumption for various tasks.

Other differences between the corpora rely on the kind of encoded information. For example, one of the great feature of the PDTB is that it makes a distinction between different types of relations, especially between explicit and implicit ones, leading to studies that showed that the discourse markers are not very ambiguous (Pitler & Nenkova 2009), whereas implicit relations are really hard to identify, with low scores even when taking into account a large range of indices (Lin 2009, Rutherford & Xue 2015). The release of this corpus finally led to a large number of studies dedicated to implicit relations that interestingly allowed improvements for parsers built on the RST DT (Feng & Hirst 2014).

Finally, other differences concern the segmentation, a crucial and complex problem (Carlson 2001a; Taboada & Mann 2006b) The nature of the elementary discourse units (EDU) naturally defines the task performed by a segmenter but also influences the other systems, the EDU being the basic unit forming their inputs. One issue is the treatment of the embedded units in the RST DT that led to the use of a specific relation (Same-Unit) to link the parts of an EDU made discontinuous by an embedded unit. If this choice allows to keep the adjacency constraint, it also implies that parsers in fact try to learn a segmentation mixing complete and partial discourse units and a labeling mixing rhetorical with textual organization links. Annodis and its graph structure does not need such a pseudo relation. Another issue is the known difficulty of the segmentation in the PDTB that raises the question of the usefulness of an exact match of the arguments for the task, and, besides, the CoNLL shared task this year[8] will allow for partial matching.

[8] http://www.cs.brandeis.edu/~clp/conll16st/index.html

We have discussed some specific features of the three main annotation schemes for discourse showing that choices made during annotation led to decisions in automatic systems that could seem somewhat arbitrary or questionable (for instance, why include Same-Unit as a rhetorical relation? What should be done with multiple labels or specific Entity relations annotated in the PDTB?), that make them difficult to compare, and that end in a split of the domain into two groups (parsing *vs.* chunking) seemingly hard to reconcile. We think that NLP researchers in discourse need now to think more about the interoperability of their systems across frameworks and languages, thus beginning with the use of the existing shareable sets of relations when evaluating their systems, and then possibly the definition of a common task, or at least a shareable evaluation method, taking into account recent works on structure comparison.

References

- Afantenos, St., Asher, N., & Benamara, F. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In: *Proceedings of LREC*.
- Asher, N. & Lascarides, A. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.
- Benamara, F. & Taboada, M. 2015. Mapping different rhetorical relation annotations: A proposal. In: *Proceedings of Starsem*.
- Carlson, L., Marcu, D. & Okurowski, M. E. 2001a. Building a discourse-tagged corpus in the framework of rhetorical structure theory, *Proceedings of SIGdial*.
- Carlson, L. & Marcu, D. 2001b. *Discourse Tagging Reference Manual.* Rapp. tech. University of Southern California Information Sciences Institute.
- Feng, V. F. & Hirst, G. 2014. A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing. In: *Proceedings of ACL*.
- Ji, Y. & Eisenstein, J. 2014. Representation Learning for Text-level Discourse Parsing. In: *Proceedings of ACL*.
- Joty, S. R., Carenini, G. & Ng, R. T. 2012. A Novel Discriminative Framework for Sentence-Level Discourse Analysis. In: *Proceedings of EMNLP*.
- Lee, A., Prasad, R., Joshi, A., Dinesh, N. & Webber, B. 2006. Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax. In: *Proceedings of the Workshop on Treebanks and Linguistic Theories*.

- Li, S., Wang, L., Cao, Z. & Li, W. 2014. Text-level Discourse Dependency Parsing. In: *Proceedings of ACL*.
- Lin, Z., Kan, M-Y. & Ng, H. T. 2009. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In: *Proceedings of EMNLP*.
- Lin, Z., Ng, H. T. & Kan, M-Y. 2010. *A PDTB-styled end-to-end discourse parser.* University of Singapore.
- Lin, Z., Liu, Ch., Ng, H. T. & Kan, M-Y. 2012. Combining coherence models and machine translation evaluation metrics for summarization evaluation. In: *Proceedings of ACL*.
- Louis, A., Joshi, A. & Nenkova, A. 2010. Discourse indicators for content selection in summarization. In: *Proceedings of SIGDIAL*.
- Mann, W. C. & Thompson, S. A. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8: 243–281.
- Meyer, Th. & Popescu-Belis, A. 2012. Using sense-labeled discourse connectives for statistical machine translation. In: *Proceedings of the Workshop on Hybrid Approaches to Machine Translation*.
- Meyer, Th. & Webber, B. 2013. Implicitation of discourse connectives in (machine) translation. In: *Proceedings of the Workshop on Discourse in Machine Translation*.
- Miltsakaki, E., Dinesh, N., Prasad, R., Joshi, A. & Webber, B. 2005. Experiments on sense annotation and sense disambiguation of discourse connectives. In: *Proceedings of the FourthWorkshop on Treebanks and Linguistic Theories*.
- Muller, Ph., Afantenos, St., Denis, P. & Asher, N. 2012. Constrained decoding for text-level discourse parsing. In: *Proceedings of COLING*.
- Pitler, Emily & Nenkova, A. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In: *Proceedings of the ACL-IJCNLP*.
- Prasad, R., Dinesh, N., Lee, A. 2008. The Penn Discourse Treebank 2.0. In: *Proceedings of LREC*.
- Prasad, R., & Bunt, H. 2015. Semantic relations in discourse: The current state of ISO 24617-8. In: *Proceedings of the ACL-ISO Workshop on Interoperable Semantic Annotation*.
- Rutherford, A. & Xue, N. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In: *Proceedings of NAACL-HLT*.
- Sporleder, C. & Lapata, M. 2005. Discourse chunking and its application to sentence compression. In: *Proceedings of HLT/EMNLP*.

- Taboada, M. & Mann, W. C. 2006a. Applications of Rhetorical Structure Theory. *Discourse Studies* 8: 567–588.
- Taboada, M. & Mann, W. C. 2006b. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies* 8: 423–459.
- Tsoumakas, G. & Katakis, I. 2006. *Multi-label classification: An overview.* Aristotle University of Thessaloniki.
- Venant, A., Asher, N., Muller, Ph., Denis, P. & Afantenos, St. 2013. Expressivity and comparison of models of discourse structure. In: *Proceedings of SIGDIAL*.
- Versley, Y. 2011. Towards finer-grained tagging of discourse connectives. In: *Proceedings of the Workshop Beyound Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena*.
- Wang, X., Li, S., Li, J. & Li, W. 2012. Implicit Discourse Relation Recognition by Selecting Typical Training Examples. In: *Proceedings of COLING*.
- Webber, B., Egg, M. & Kordoni, V. 2011. Discourse Structure and Language Technology. *Natural Language Engineering* 18 (4): 437–490.
- Xue, N., Ng, H. T. & Pradhan, S. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In: *Proceedings of CoNLL*.

Ludivine Crible, Liesbeth Degand, Anne Catherine Simon

Interdependence of annotation levels in a functional taxonomy for discourse markers in spoken corpora

Université catholique de Louvain
ludivine.crible@uclouvain.be; liesbeth.degand@uclouvain.be;
anne-catherine.simon@uclouvain.be

Despite the proliferation of corpus-based studies focusing on the complex category of discourse markers in the recent years, consensus remains to be found regarding the most reliable yet informative model to describe their behavior in authentic data. Major frameworks (e.g. the Penn Discourse TreeBank, Prasad et al. 2008; Rhetorical Structure Theory, Mann & Thompson 1988) disagree both on the top levels (number and type of generic annotation levels, if any) and the specific relations included in them. It is precisely the relation between top levels and corresponding sublevels that is addressed in the present study, starting from a recent proposal of functional taxonomy (Crible in press) applied to the French-English spoken corpus DisFrEn and its revision in the framework of the LOCAS-F corpus (Martin et al. 2014).

This paper compares the structure and content of the original and revised coding schemes. By weighing their benefits and drawbacks, the comparison paves the way for several theoretical and methodological considerations regarding the operationality and cognitive validity of annotation models for discourse-structuring devices.

Domains and functions in DisFrEn and LOCAS-F

In DisFrEn, four top-level functions or "domains" are distinguished, from the revision of existing proposals for both speech and writing (Cuenca 2013, Gonzalez 2005, Halliday & Hasan 1976, Zufferey & Degand in press): ideational (ob-

jective relations), rhetorical (subjective, metadiscursive functions), sequential (structuring functions) and interpersonal (speaker-hearer relationship). In the original model, these four domains comprise a total of thirty functions, each of them belonging to one – and only one – domain. For instance, the function labeled "CAUSE" is always ideational, while "MOTIVATION" is always rhetorical, although in other frameworks they would be considered as two subtypes of the same causal relation. This interdependent system is intended to maximize the informativity of the annotation labels (one label for one function, vs. combinations of labels) while giving the opportunity to filter the distribution from thirty to four conceptually coherent categories, hence more efficient for quantitative purposes.

The revised version currently being applied to LOCAS-F aims at reducing the number of options and enhancing the reliability and cognitive validity of the model. It was elaborated through a back-and-forth work between *i)* inductive, conceptual grouping of different relations by their semantic proximity and *ii)* corpus testing on authentic examples. The main difference between this revision and the original is that domains and functions are no longer interdependent. On the contrary, it is assumed that many functions can be assigned more than one domain. For instance, a [contrast] can be ideational *1),* rhetorical *2),* or sequential *3).*

1) I wasn't looking forward to doing it **but** I am now (DisFrEn EN-phon-01)
2) a rebate is when they send the money back // yes **but** how do you define it in economic terms (DisFrEn EN-clas-02)
3) (after a digression on the industries in Bristol area) **but** Bristol itself is a large metropolis (DisFrEn EN-intf-05)

The pilot study on 200 DMs in LOCAS-F leads to the following results. The set of 30 relational functions could be reduced to 14. The merge between the original ideational and rhetorical domains was nearly total, while a number of specific sequential relations (topic-shift, opening, closing) were retained, but none of the specific interpersonal functions. Results furthermore show that the sequential domain is the most frequent one, followed by the rhetorical and ideational ones, with only a marginal use of the interpersonal domain. The most multi-purpose functions (Geertzen & Bunt 2006) are [addition], [contrast], and [consequence], all of which are used in the ideational, sequential and rhetorical domains, although in different proportions. The functions [opening] and [topic-shift] are exclusive to the sequential domain. Finally, the most polyfunctional

markers are: *et* 'and', *enfin* 'eventually' and mais 'but' fulfilling six different functions, followed by alors 'then/so' and en fait 'in fact'. These are also among the most frequent markers in the data set.

From interdependence to multi-domain functions: implications

The proposal of multi-domain functions addresses a number of theoretical and methodological shortcomings of the original system. In particular, it allows to annotate at two independent levels of precision from the most generic to the most specific, domain first and function second, following previous research (Zufferey & Degand in press) showing that higher annotation levels trigger less disagreements. This in turn calls for an investigation of the relation between offline annotation of discourse relations (by experts) and online interpretation (by speakers).

The revised taxonomy also involves three practical biases: a relational bias (reconstitution of a two-part relation in case of hesitation with a non-relational function); a semantic bias (reconstitution of the lexical meaning if possible, e.g. contrast in but); and a single-label bias that specifies how to choose one of two domains in case of simultaneous double meanings. These decisions acknowledge the artificial nature of annotation compared to online interpretation, and aim at better coverage and reliability of the model. In general, we advocate for a better documentation of all annotation decisions and biases as a methodological principle of transparency and replicability.

This new approach also offers to show empirical evidence for the integration of topic relations into taxonomies of discourse relations, a proposal which is currently not consensual especially in studies on written language. We argue that sequential functions are similar to their ideational and rhetorical equivalents and merely constitute a different, more global level of discourse coherence (Lenk 1998). In this respect, qualitative analysis of double-tagged occurrences annotated with the original taxonomy can be used to identify borderline cases. Further analyses would need to uncover whether some connectives specialize in the marking of sequential relations (potentially qualifying candidates include *bon/ben* 'good/well', *quoi* 'punctuator'), and/or whether the different domains are associated with contextual and/or formal characteristics (e.g. co-occurrence patterns, register variation).

A final point of debate concerns cases of DMs which activate a domain but do not "fit" into any of the functions (e.g. French *alors* which is very rarely exclusively temporal, conditional or causal). Should we allow for under-specified

functions? It seems that ambiguity and under-specification are part and parcel of spoken language use and should be accounted for in theoretical models.

References

- Crible, L. in press. Towards an operational category of discourse markers: A definition and its model. In: Fedriani, C. & Sanso, A. (eds) *Discourse Markers, Pragmatic Markers and Modal Particles: New Perspectives.* Amsterdam: John Benjamins.
- Crible, L. & Degand, L. 2015. Functions and syntax of discourse markers across languages and genres. In: *Towards a multilingual annotation scheme, 14th IPrA*.
- Cuenca, M. J. 2013. The fuzzy boundaries between discourse marking and modal marking. In: Degand, L., Cornillie, B. & Pietrandrea, P. (eds) *Discourse markers and modal particles. Categorizations and description.* Amsterdam: John Benjamins. 191–216.
- Geertzen, J. & Bunt, H. 2006. Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In: *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, 2006,* 126–133.
- Gonzalez, M. 2005. Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse Studies* 7 (1): 53–86.
- Halliday, M. A. K. & Hasan, R. 1976. *Cohesion in English.* London: Longman.
- Lenk, U. 1998. Discourse markers and global coherence in conversation. *Journal of Pragmatics* 30: 245–257.
- Mann, W. C. & Thompson, S. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization, Text 8 (3): 243–281.
- Martin, L., Degand, L., Simon, A. C. 2014. Forme et fonction de la périphérie gauche dans un corpus oral multigenres annoté. *Corpus* 13: 243–265.
- Rashmi, P., 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),* 2008.
- Spooren, W. & Sanders, T. 2008. The acquisition order of coherence relations: On cognitive complexity in discourse. *Journal of Pragmatics* 40: 2003–2026.
- Sweetser, E. E. 2000. *From etymology to pragmatics. Metaphorical and cultural aspects of semantic structure.* Cambridge: Cambridge University Press.
- Zufferey, S. & Degand, L. in press. Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory 10.*

Iria da Cunha

Towards Discourse Parsing in Spanish

Universidad Nacional de Educación a Distancia (UNED), Spain
iria.dacunha@upf.edu

Introduction

Texts can be analysed from different perspectives. One of the most difficult phenomena to process is discourse structure (Hovy 2010). In recent years, one of the main challenges in the field of Natural Language Processing (NLP) has been discourse parsing. Research on this topic has been done for several languages, such as Japanese (Sumita et al. 1992), English (Marcu 2000) and Portuguese (Pardo 2008), among others. Also, for English, the CoNLL-2015 Shared Task focused on Shallow Discourse Parsing.[9] Discourse annotated corpora have been created too, for example for English (Carlson et al. 2002), German (Stede 2004), Portuguese (Pardo 2008) and French (Afantenos 2012). Discourse parsing tools and resources are used to develop NLP applications; for example, automatic summarization, information extraction, text generation, machine translation and sentiment analysis (Taboada & Mann 2004).

The aim of this paper is to present the advances in discourse parsing for Spanish. Specifically, after explaining our theoretical framework, we will detail the tools we have developed for the automatic annotation of discourse information in texts in Spanish and the discourse annotated resources we have created.

Theoretical Framework

Most discourse NLP tools are based on Rhetorical Structure Theory (RST, Mann & Thompson 1988). This is a language independent theory based on the idea that a text can be segmented into Elementary Discourse Units (EDUs)

[9] http://www.aclweb.org/anthology/K/K15/

linked by means of nucleus-satellite or multinuclear discourse relations. In the first case, the satellite gives additional information about the other one, the nucleus, on which it depends (e.g. Result or Concession). In the second case, several elements, all nuclei, are connected at the same level, that is, there are no elements dependent on others and they all have the same importance with regard to the intentions of the text author (e.g. Contrast or Sequence). RST discourse parsing includes three stages: *a)* segmentation, *b)* relations detection and *c)* building of hierarchical rhetorical trees.

Discourse Tools and Resources for Spanish

In this section we explain the discourse tools and resources we have developed for Spanish, in the framework of RST. First, we have developed the discourse segmenter DiSeg (da Cunha et al. 2012), which can be used online.[10] It is based on shallow parsing and a set of linguistic rules that insert segment boundaries into sentences, following specific criteria.[11] DiSeg performance was evaluated using a corpus of manually annotated texts (a gold standard).[12] The system obtained an F-score between 80% and 96% in experiments with a corpus containing medical texts, and an F-Score of 91% with a corpus of texts about terminology.

Second, we have developed a discourse corpus containing texts manually annotated, the RST Spanish Treebank (da Cunha et al. 2011), which can be consulted and downloaded online.[13] The texts have been annotated with the RST-Tool (O'Donnell 2000). The corpus includes 267 specialised texts (from several domains and genres), 52,746 words, 2,256 sentences and 3,349 discourse segments. It is divided into a learning corpus (183 texts) and a test corpus (84 texts).

Third, we have developed a sentence-level discourse parser, DiSeg2 (da Cunha et al. 2012a), which can also be consulted online.[14] To do this, we have analysed the learning corpus of the RST Spanish Treebank in order to manually detect all the markers that show discourse relations. We divided the markers into 3 categories: *1)* traditional discourse markers, *2)* markers including lexical units

[10] http://dev.termwatch.es/esj/DiSeg/WebDiSeg/
[11] Similar to the ones used in: da Cunha & Iruskieta 2010.
[12] http://dev.termwatch.es/esj/DiSeg/index.html
[13] http://corpus.iingen.unam.mx/rst/
[14] http://diseg2.termwatch.es/

(nouns and verbs), and *3)* markers including verbal structures. We obtained 778 markers. Taking these markers into account, we have designed an algorithm to automatically detect intra-sentence RST relations and nuclearity. It is based on linguistic rules including discourse patterns and the aforementioned discourse segmenter. We have evaluated the system with the test corpus, obtaining an accuracy of 81.75 regarding EDUs, SPANs (that is, sets of EDUs) and nuclearity, and 81.75 with regard to relations.

Fourth, we have created DiZer 2.0 (Maziero et al. 2011), an adaptable online platform designed to develop discourse parsers in any language, which integrates a language-independent algorithm to build discourse trees (Marcu 2000). In order to automatically obtain hierarchical rhetorical trees from full texts in Spanish, we have included our discourse segmenter and patterns in this platform. Currently, we are evaluating the performance of this discourse parser for Spanish.

Conclusions and Future Work

The aim of this paper has been to show the main automatic tools and resources related to discourse parsing for Spanish: the discourse segmenter, the RST Spanish Treebank, the sentence-level discourse parser, and the platform to build rhetorical trees. As future work, we plan to evaluate the complete discourse parser and to develop several NLP applications. Also, we plan to research about the cross-linguistic applicability of these tools.[15]

References

- Afantenos, S., Asher, N. & Benamara, F. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In: *Proceedings of the 8th Conference of LREC*, 2012.
- Carlson, L., Marcu, D. & Okurowski, M. E. 2002. *RST Discourse Treebank.* Pennsylvania: Linguistic Data Consortium.

[15] This work has been partially supported by a Ramón y Cajal research contract (RYC-2014-16935) and the research project APLE 2 (FFI2009-12188-C05-01) of the Institute for Applied Linguistics (IULA).

- da Cunha, I. & Iruskieta, M. 2010. Comparing rhetorical structures of different languages: The influence of translation strategies. *Discourse Studies* 12 (5): 563–598.
- da Cunha, I., Torres-Moreno, J-M., Sierra, G. 2011. On the Development of the RST Spanish Treebank. In: *Proceedings of the 5th Linguistic Annotation Workshop (ACL 2011)*. 1–10.
- da Cunha, I., SanJuan, E., Torres-Moreno, J-M., Cabré, M. T. & Sierra, G. 2012a. A Symbolic Approach for Automatic Detection of Nuclearity and Rhetorical Relations among Intra-sentence Discourse Segments in Spanish. *LNCS* 7181: 462–474.
- da Cunha, I., SanJuan, E., Torres-Moreno, J-M., Lloberes, M. & Castellón, I. 2012b. DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Systems with Applications (ESWA)* 39 (2): 1671–1678.
- Feng, V. W. & hirst, G. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In: *Proceedings of the 52nd Annual Meeting of the ACL (2014)*. 511–521.
- Hovy, E. 2010. Annotation. A Tutorial. *Presented at the 48th Annual Meeting of the ACL*.
- Joty, Sh., Carenini, G. NG, R. 2015. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics* 41 (3): 385–435.
- Mann, W. C. & Thompson, S. A. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8 (3): 243–281.
- Marcu, D. 2000. The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach, *Computational Linguistics* 26 (3): 395–448.
- Maziero, E. G. & Pardo, Th. A. S., da Cunha, I., Torres-Moreno, J-M. & San-Juan, E. 2011. DiZer 2.0 – An Adaptable On-line Discourse Parser. In: *Proceedings of the III RST Meeting (STIL 2011)*.
- O'Donnell, M. 2000. RSTTOOL 2.4 – A markup tool for rhetorical structure theory. In: *Proceedings of the International Natural Language Generation Conference (2000)*. 253–256.
- Pardo, Th. A.S. & Nunes, M. G. V. 2008. On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Journal of Theoretical and Applied Computing* 15(2): 43–64.
- Soricut, R. & Marcu, D. 2003. Sentence Level Discourse Parsing using Syntactic and Lexical Information. In: *Proceedings of 2003 HLT and North American ACL Conference (2003)* 149–156.
- Stede, M. 2004. The Potsdam commentary corpus. In: *Proceedings of the Workshop on Discourse Annotation (ACL 2004)*.

- Subba, R. & Di Eugenio, B. 2009. An effective discourse parser that uses rich linguistic information. In: *Proceedings of 2009 HLT and North American ACL Conference (2009).* 566–574.
- Sumita, K., Ono, K., Chino, T., Ukita, T. & Amano, Sh. 1992. A discourse structure analyzer for Japanese text. In: *Proceedings of the International Conference on Fifth Generation Computer Systems 2.* 1133–1140.
- Taboada, M. & Mann, W. C. 2006. Applications of rhetorical structure theory. *Discourse Studies* 8 (4): 567–588.

Laurence Danlos, Pierre Magistry

Discourse Treebanks in a Graph Database

ALPAGE, UMR INRIA – Univ. Paris Diderot
laurence.danlos@inria.fr; pierre.magistry@inria.fr

In this work, we introduce a Discourse TreeBank modeled as a Graph to be stored and queried in a graph oriented database. This work has been conducted on the FDTB (French Discourse Treebank, Danlos et al. 2015), which is in the PDTB style and uses the tools designated to annotate the PDTB (Penn Discourse TreeBank, Prasad et al. 2008). This means that any Discourse TreeBank annotated in the PDTB style could benefit from our work.

Modeling data as a large graph is a new trend in the world of databases (Robinson et al. 2013). It allows users to elegantly and efficiently describe, store and query highly relational data. We use the software Neo4J, which is freely downloadable at http://neo4j.com/. This software comes with a convenient and powerful query language, called Cypher, which allows users to search and to edit the database through graph-pattern matching.

Cypher language enables us to build queries that cannot be formulated with the PDTB browser and so require writing complex programs. For example, with Cypher it is easy to write a query that matches the occurrences of "multiple connectives" sharing the same argument (Arg2) and then to test if they also share the other argument (Arg1) while taking into account the discourse relation they lexicalize.

In the graph database, it is also possible and easy to include, on top of discourse annotation, other linguistic resources such as a connective lexicon, for example LexConn (Roze et al. 2012) for French, and any syntactic parse trees of the raw corpus. We can then design Cypher queries to assess the consistency of the discourse annotation with external information.

This database also allows us to replicate machine learning experiments from Pitler (Pitler & Nenkova 2009) and Johannsen (Johanssen & Søgaard 2013). We wrote a small program that trains and tests a MaxEnt classifier using a set of features extracted from the data thanks to a Cypher query. Under this experi-

mental design, it is straightforward to change the set of features used by the classifier, changing only the query.

In conclusion, the tools we designed, which are released as Open Source software,[16] are quite useful both to explore linguistic issues and to conduct machine learning experiments. They have been tested only on French so far, but we foresee that their adaptation would be easy for any other language for which PDTB style annotation is available. A demonstration of these tools will be performed during the conference.

References

- Danlos, L., Colinet, M. & Steilin, J. 2015. FDTB1, première étape du projet "French Discourse TreeBank": repérage des connecteurs de discours en corpus. *Discours*, Varia (17). [http://discours.revues.org/9065]
- Johannsen, A. & Søgaard, A. 2013. Disambiguating explicit discourse connectives without oracles. In: *Proceedings of the Sixth International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing.*
- Pitler, E. & Nenkova, A. 2009. Using syntax to disambiguate explicit discourse connectives in text. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Association for Computational Linguistics.*
- 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 2008),* Marrackech, Maroc.
- Robinson, I., Webber, J. & Eifrem, E. 2013. *Graph Databases*. Sebastopol, CA: O'Reilly.
- Roze, Ch., Danlos, L. & Muller, Ph. 2012. LexConn: a French Lexicon of Discourse connectives. Discours 10. [http://discours.revues.org/8645]

[16] http://fdtb-v1.gforge.inria.fr/

Jacqueline Evers-Vermeul,[a] Jet Hoek,[a]
Merel Scholman[b]

On Temporality in Discourse Annotation

[a]*Utrecht University, The Netherlands;* [b]*Saarland University, Germany*
J.Evers@uu.nl; j.hoek@uu.nl; m.c.j.scholman@coli.uni-saarland.de

One of the features that determine the coherence of a discourse is the temporal ordering of the segments. Language acquisition studies show that temporal relations are among the first types of coherence relations that are explicitly marked by a connective: children start using *and*, followed by *(and) then* and *because* (Bloom et al. 1980, Evers-Vermeul & Sanders 2009). In contrast to this prominence of temporality in acquisition studies, processing studies reveal that a temporal marking of a coherence relation can easily be overruled by other features of the relation. For instance, Mak and Sanders (2013) have shown that immediate effects of causal relatedness on referential processing occur even with a connective that is not explicitly causal *(when)*. Given these insights, the research question of this paper is: What is the status of temporality in language use in general and in discourse annotation in particular?

We aim to answer this question by bringing together and comparing different kinds of data. First, we will discuss several outcomes of language acquisition and processing studies. Among other things, these studies underline the importance of the phenomenon of underspecification (Spooren 1997): linguistic markers of coherence relations need not exactly match the type of relation intended by the writer/speaker or perceived by the reader/listener. For instance, *and* and *and then* can both be used in cause-consequence relations).

Second, we compare four approaches towards the annotation of temporality in discourse: the ones taken by the Cognitive approach to Coherence Relations (CCR, Sanders et al. 1992), the Penn Discourse Tree Bank (Prasad et al. 2008), the Rhetorical Structure Theory Treebank (Carlson et al. 2003), and Segmented Discourse Representation Theory (Reese et al. 2007). We will focus on questions such as: What kind of temporal relations are listed? And is the temporal ordering of discourse segments considered a relational feature or a segment-specific

feature (e.g. determined by the tense of one or both of the segments)? For example, CCR treats temporal relations as a subtype of additive relations and claims that temporality is not a relational but a segment-specific feature. In contrast, the PDTB presents Temporals as one out of four major classes of discourse relations, distinguishing between three subtypes (Synchronous, Precedence and Succession).

In our discussion of these data and research outcomes, we will address how temporality interacts with other properties of discourse relations, such as order, causality, and subjectivity. This will enable us to draw conclusions on whether temporality should be seen as an independent relational feature or whether it is a segment-specific phenomenon and/or a phenomenon that is a by-product of (certain combinations of) other characteristics of coherence relations.

References

- Bloom, L., Lahey, M. Hood, L. Lifter, K. & Fliess, K. 1980. Complex Sentences: Acquisition of Syntactic Connectives and the Semantic Relations They Encode. *Journal of Child Language* 7 (2): 235–261.
- Carlson, L., Marcu, D. & Okurowski, M. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In: van Kuppevelt, J. C. J. & Smith, R. W. (eds.) *Current Directions in Discourse and Dialogue.* Dordrecht: Kluwer Academic Publishers. 85–112.
- Evers-Vermeul, J., Sanders, T. J. M. 2009. The Emergence of Dutch Connectives: How Cumulative Cognitive Complexity Explains the Order of Acquisition. *Journal of Child Language* 36 (4): 829–854.
- Mak, W. M. & Sanders, T. J. M. 2013. The Role of Causality in Discourse Processing: Effects of Expectation and Coherence Relations. *Language and Cognitive Processes* 28 (9): 1414–1437.
- 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 of Language Resources and Evaluation (LREC 2008),* 2008, Marrakech.
- Reese, B., Hunter, J., Asher, N., Denis, P. & Baldridge, J. 2007. *Reference Manual for the Analysis and Annotation of Rhetorical Structure (version 1.0). Technical report.* Austin: University of Texas, Departments of Linguistics and Philosophy. [http://timeml. org/jamesp/annotation_manual. pdf]

- Sanders, T. J. M., Spooren, W. P. M. S. & Noordman, L. G. M 1992. Toward a Taxonomy of Coherence Relations. *Discourse Processes* 15: 1–35.
- Spooren, W. P. M. S. 1997. The Processing of Underspecified Coherence Relations. *Discourse Processes* 24: 149–168.

Sílvia **Gabarró-López, Laurence Meurant**

Studying the position of Discourse Relational Devices in signed languages: adapting the Basic Discourse Units Model to the signed modality?

University of Namur
silvia.gabarro@unamur.be; laurence.meurant@unamur.be

Introduction

Discourse segmentation is at the basis of the study on how discourse in the oral setting is structured regardless of the modality, i.e. spoken or signed. However, the reality of each modality is very different from the other: while scholars working on spoken languages (SpLs) have developed some models to segment spoken discourse, sign languages (SLs) are far beyond having a model or models, although its necessity and importance has been widely acknowledged in the literature (Börstell et al. 2014, Crasborn 2007, Ormel & Crasborn 2012).

To the best of our knowledge, six different models are used for the segmentation of spoken discourse: the Geneva Model (Roulet et al. 1985), the Val.Es.Co. Model (Briz Gómez & Val.Es.Co. Grupo 2003), the Fribourg Model (Groupe de Fribourg 2012), the Co-Enunciation Model (Morel & Danon-Boileau 1998), the Prominence Demarcation Model (Lombardi Vallauri 2009) and the Basic Discourse Units Model (Degand & Simon 2009a, b). The main difference between these models is that they approach the issue of segmentation from different points of view, namely pragmatic (GM, VAM and FM), prosodic (CEM and PDM) or in a combination of syntax and prosody (BDU).

The question of discourse segmentation in the signed modality is a tricky one because of the delay in the study of SLs as natural languages. Consequently, we are in very preliminary stage of knowledge on how SLs are structured in different linguistic domains (syntax, discourse, etc.) In addition, we cannot entirely

rely on a SpL model because of the specificities of SLs: the two hands are the main articulators and they produce simultaneous constructions, and nonmanuals also participate in the construction of meaning. Despite these difficulties, the need to have a segmentation model that allows working on how discourse is structured from different points of views is modality-independent. Bearing in mind this convergence point, this paper aims to propose a model for the segmentation of signed discourse whose ultimate goal is to allow the study of the position of discourse markers (DMs) through the discourse, i.e. large sets of utterances.

After reviewing the different segmentation models for SpLs, the most suitable model for the segmentation of signed discourse appears to be the BDU Model. Its main advantage is that it is not only applicable to conversation, which was the main drawback for other potentially interesting models that have already been used for the study of DMs such as the VAM or the CEM, but it can also be used for monologic data. Due to the delay in SL research, our model needs to be as versatile as possible (i.e. applicable to as many discourse situations as possible) allowing the use of the "same measures" to segment both monologues and dialogues, and therefore get comparable units in both settings.

Methodology

In order to adapt the BDU Model to the signed modality getting the most interoperable model possible, our corpus contains data from the two SLs that are available to us: French Belgian Sign Language (LSFB) and Catalan Sign Language (LSC). We took 12 signers, 6 from the LSFB referential corpus and 6 from the LSC referential corpus. This sample is balanced in terms of age (2 signers from each SL belonging to one of the following age groups: 18–29, 30–49 and 50–80) and gender (3 men and 3 women per SL). For both video corpora (annotated with ELAN), signers came in couples (both belonging to the same age group) and their conversations (including argumentations, descriptions, explanations and narrations) were guided by a moderator.

Results

As in the BDU Model, our segmentation procedure consists of three different steps: (i) delimiting syntactic units, (ii) delimiting prosodic units, and (iii) finding the convergence point between syntactic and prosodic units in order

to establish BDUs. The first two steps are independent, which means that once the syntactic segmentation is finished, this tier is hidden in order to carry out the prosodic segmentation independently. For the first step, the Dependency Grammar for spoken French is used to delimit clauses (Blanche-Benveniste et al. 1984, Blanche-Benveniste et al. 1990). Verbs are the nuclei of clauses and their actants and circumstants are identified. In some cases, other categories (e.g. nouns or adjectives) are the nucleus of the unit. Adjuncts (such as DMs) are out of the dependency structure of the verb and therefore constitute a syntactic unit by themselves.

For prosodic segmentation, the cues taken to delimit major boundaries in the BDU Model are silent pauses, a lengthening of the syllable (three times longer than the syllables in context) or a sharp rise of f0 (intra-syllabic f0 superior to ten semi-tones). As for the two first cues, the equivalent in signed discourse is quite straightforward: pauses (i.e. periods of no signing) on the one hand; and a sign hold or a lengthened sign with respect to the context on the other. As for the third cue, we looked at eye blinks layered with another prosodic cue such as a head movement or a change in gaze as the SL literature claims that eye blinks have a prosodic function (Sandler 2012).

The final step consists of displaying both the syntactic and prosodic tiers to see where the boundaries coincide and establish BDUs. The tests carried out so far prove that although it is a time consuming process, the adaptation of the BDU Model provides us with a practical protocol for the segmentation of discourse allowing a more fine-grained study of the position of DMs in the signed modality. As a matter of fact, we have seen that the position of the DM AUSSI (here translated as ALSO) in LSFB correlates with its function in a particular context. If we take two common functions of ALSO, i.e. addition (adding information to the same topic) and specification (introducing an example), we can see that each function displays a particular position with respect to the clause and the BDU. Addition is found at the left periphery of the clause and at the syntactic left periphery of the BDU as in Example 1 (http://www.corpus-lsfb.be, session 21, task 04, 2:37–2:42). There are four clauses (delimited with square brackets), two within each BDUs (delimited by slashes). *ALSO* is out of the dependency structure of the verb GO (i.e. clausal left periphery), but it is prosodically integrated at the beginning of the BDU (i.e. syntactic left periphery).

(1) /[hearing i go bicycle learn] [bibycle there go] / *also* [go horse] [i go horse]/
"The Hearing taught me how to cycle. And I went to ride horses…"

Specification is found at the left periphery of the clause but in the medial position of the BDU as in Example 2 (http://www.corpus-lsfb.be, session 27, task 04, 2:29–2:33). In this case, there are three clauses within the same BDU. ALSO is out of the dependency structure of the verb REMEMBER (i.e. clausal left periphery), but it is prosodically integrated in the middle of the BDU (i.e. BDU medial position).

(2) /[YES] *ALSO* [REMEMBER BEFORE LITTLE ALWAYS I] [TODAY SECOND MEMORY CHILD]/
"Yes, for instance I remember when I was young… Well, this is my second child memory today."

This coupling of position and function of ALSO is regular across different examples of our corpus, which includes different signers and different genres. Therefore, the position can be used as a criterion to identify the function of a polysemous DM such as ALSO. Our data suggest that the adaptation of the BDU Model for the segmentation of SLs provides us with enlightening results for the study of, at least, DMs.

References

- Blanche-Benveniste, C., Deulofeu, J., Stefanini, J. & van den Eynde, K. 1984. *Pronom et syntaxe. L'approche pronominale et son application au français.* Paris: SELAF.
- Blanche-Benveniste, C., Bilger, M., Rouget, Ch. & van den Eynde, K. 1990. *Le français parlé: études grammaticales*. Paris: Éditions du CNRS.
- Börstell, C., Mesch, J. & Wallin, L. 2014. Segmenting the Swedish Sign Language corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In: Crasborn, O., Efthimiou, E., Fotinea, S. E., Hanke, T., Hochgesang, J., Kristoffersen, J. H. & Mesch, J. (eds.) *Beyond the manual channel. 6th Workshop on the Representation and Processing of Sign Languages*, Reykjavik, ELRA, 2014, 7–10.
- Briz Gómez, A. & Grupo Val.Es.Co. 2003. Un sistema de unidades para el estudio del lenguaje colloquial. *Oralia* 6: 7–61.
- Crasborn, O. 2007. How to recognise a sentence when you see one. In: Crasborn, O. (ed.) *Identifying sentences in signed languages,* Vol. 10–2. Amsterdam/Philadelphia: John Benjamins. 103–111.

- Degand, L. & Simon, A. C. 2009a. Mapping Prosody and Syntax as Discourse Strategies: How Basic Discourse Units vary across Genres. In: Barth-Weingarten, D., Dehé, N. & Wichmann, A. (eds.) *Where Prosody meets Pragmatics*, Vol. 8. Bingley: Emerald.
- Degand, L. & Simon, A. C. 2009b. Minimal Discourse Units in Spoken French: On the Role of Syntactic and Prosodic Units in Discourse Segmentation. *Discours* 4.
- Groupe de Fribourg 2012. *Grammaire de la période.* Berne: Peter Lang.
- Lombardi Vallauri, E. 2009. *La struttura informativa. Forma e funzione negli enunciati linguistici*. Roma: Carocci.
- Morel, M.-A. & Danon-Boileau, L. 1998. *Grammaire de l'intonation. L'exemple du français.* Paris: Ophrys.
- Ormel, E. & Crasborn, O. 2012. Prosodic correlates of sentences in signed languages: A Literature Review and Suggestions for New Types of Studies. *Sign Language Studies* 12 (2): 109–145.
- Roulet, E., Auchlin, A., Moeschler, J. & Schelling, M. 1985. *L'articulation du discours en français contemporain.* Berne: Peter Lang.
- Sandler, W. 2012. Visual Prosody. In: Pfau, R., Steinbach, M. & Woll, B. (eds.) *Sign Language. An International Handbook*. Berlin/Boston: De Gruyter Mouton. 55–76.

Yulia Grishina and Manfred Stede

Referring expressions as cohesive devices in multiple languages

Applied Computational Linguistics, University of Potsdam
grishina@uni-potsdam.de; stede@uni-potsdam.de

The goal of our study is to investigate how linguistic annotations of discourse coherence can be used for annotation projection and for the development of linguistically annotated resources in multiple languages. In particular, we are dealing with annotations of coreference, which is necessary to establish coherence in discourse (Halliday & Hasan 1976). Anaphoric and cataphoric expressions operate as a type of discourse relational device (DRDs) that form a layer of discourse structure, whose inter-relations with coherence relations are an interesting object of study. It was shown that successful resolution of referring expressions can be exploited for automatic derivation of a text's discourse structure (Schauer & Hahn 2001).

Coreference denotes identification of all the repeated mentions to an object or state of affairs in the natural language discourse. Text coherence is maintained only when all the references and their discourse entities can be easily identified by the reader. Gaining knowledge about the distribution of coreference devices in different languages will support the standardization of a language-neutral annotation schema and mapping from one language to another.

Extended coreference relations include relations other than identity – namely, near-identity and bridging. Bridging relations are indirect relations that can only be inferred based on the knowledge shared by the speaker and the listener (Clark 1975). They encompass a wide range of relations between anaphor and antecedent, such as part-whole, or set membership. For example:

(1) Daisy walked into the office and saw a bunch of flowers on the desk.

From this example, we infer that "the desk" is definite because it is related to "the office" as one of its parts.

Additional complexity arises when two expressions refer to "almost" the same thing, but are neither identical nor non-identical. In this case, we speak of near-identity, which can be seen as a "middle ground" between identity and non-identity coreference (Recasens et al. 2010) and holds between two NPs whose referents are almost identical, but differ in one crucial dimension.

However, coreference resolution requires relatively expensive resources, usually in terms of manual annotation. To alleviate this problem for low-resourced languages, techniques of annotation projection can be applied. A projection approach is used to automatically transfer different types of linguistic annotation from one language to another. It has been successfully applied to different annotation tasks, including PoS tagging and syntactic parsing, semantic role labelling, sentiment analysis, mention detection, or named-entity recognition.

In our project, we applied a knowledge-lean projection algorithm to transfer coreference chains for two relatively similar languages (English-German) and for less similar languages (English-Russian). Furthermore, we were interested in differences incurred by the text genre and therefore used three different genres: argumentative newspaper articles, narratives, and medicine instruction leaflets. For the annotation of the corpus, we created common annotation guidelines that make few assumptions on the structural features of the target languages. We were able to achieve 0.7 MUC score (Vilain et al. 1995) for the inter-annotator agreement which is generally considered good for this task.

To align the corpus, we used a well-known, standard word alignment tool trained on a corpus of moderate size[17]. We aligned the data in both directions and took the intersection of the alignments in order to achieve maximal precision. Our approach was evaluated in three settings: *a)* the quality of the identification of mentions, *b)* projection of full coreference chains and *c)* of coreference chains with minimal spans. The experiment has shown that in a task of coreference projection, in English-German texts we obtained precision of 63.8% for full mentions (setting b) and 85.4% for mentions with minimal spans (setting c), only using a knowledge-lean approach. This indicates the promise of our algorithm for the task of coreference resolution.

We compared our results quantitatively to the most closely related work (Postolache et al. 2006) and argue that they are competitive, in particular because our task setting is more target-language-neutral (we use three languages rather than two) and we work on three different genres of text. These results are presented in detail in our recent paper (Grishina & Stede 2015).

[17] http://www.statmt.org/moses/giza/GIZA++.html

Moreover, we experimented with using limited PoS tagging and syntactic information in order to improve the performance for English and German. We used state-of-the-art parsers[18] to automatically annotate the corpus, and we were able to obtain an average improvement of 15.1% for the F1-score for 6 out of 7 newswire texts. For short stories and one newswire text, we did not get any improvement at all, which we attribute to the structural differences in NPs for different text genres.

This study is now continued by extending the corpus with the annotations of coreference relations other than identity. We developed an annotation schema for bridging and near-identity coreference after a couple of annotation experiments with English and German data. Based on the related work and our pilot annotation rounds, we developed annotation guidelines for extended coreference relations. We introduce our own categories for bridging:

- *Part – Whole*: One NP represents a physical part of the whole expressed by the other NP.
- *Set – Membership*: One can refer to a certain subset or to a single definite element of the set and bridge from this subset or element to the whole collection.
- *Entity – Attribute/Function*: An entity is a person or an object that has certain attributes characterizing it and certain functions it fulfills with respect to some other entity.
- *Event – Attribute*: Events can have necessary attributes (e.g. place, time) and optional attributes (e.g. duration, frequency).
- *Location – Attribute*: As locations we consider geographical entities that have permanent locations in the world which have certain attributes characterizing them.

Our primary goal is to introduce a domain-independent typology of bridging relations, which can be applicable across languages. We apply our annotation scheme to the German side of our coreference corpus of three genres, with a subsequent manual annotation transfer to the English and Russian sides, and for near-identity relations we use the already provided definitions (Recasens et al. 2010). Taking German annotations as source, we annotated the English and Russian sides of our parallel corpus. We present the distribution of bridging and near-identity relations across three different languages and genres and the

[18] https://code.google.com/p/mate-tools/

analysis of the resulting annotations which has shown that our guidelines are in general applicable to the three languages in our corpus; even though there are some differences across languages and genres that we will investigate in more detail.

Our scheme achieves reliable inter-annotator agreement scores on the source language side for anaphor and antecedent selection (64% and 79% F-1 score respectively) and on the assignment of bridging relations (Cohen's kappa = 0.98) which we consider as overall reliable for bridging when compared to the most closely related work on extended coreference (Nedoluzhko 2009). However, the infrequency of near-identity relations in our corpus leaves this part as a step for the future work.

Finally, we conducted a detailed analysis of the nature of bridging relations in the corpus, focusing on the distance between anaphor and antecedent. We found that the average bridging distance (anaphora + cataphora) is 20.55 tokens for all texts, with the average sentence length being 24.87 tokens. Moreover, we examined the relationship between bridging and identity coreference by computing the correlation between the length of identity chain and the number of bridging markables that are linked to this chain. Using Spearman's rank correlation coefficient, we found that there is a strong correlation between the chain length and the number of its bridges: 0.6595, with p-value of 1.35E-008.

Our next steps include *a)* implementing a multi-sourced annotation transfer and its application to a wider range of languages and *b)* refining our typology of extended coreference relation by introducing a set of possible sub-relations, conducting a more detailed comparative analysis of bridging relations across languages using annotation transfer, and exploring in detail the category of near-identity on a larger amount of texts. Still, we aim at keeping our approaches applicable to multilingual data and to different genres of text.

References

- Clark, H. H. 1975. Bridging. In: *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, 169–174.
- Grishina, Y. & Stede, M. 2015. Knowledge-lean projection of coreference chains across languages. In: *Proceedings of the 8th Workshop on Building and Using Parallel Corpora (BUCC).*
- Halliday, M. A. K & Hasan, R. 1976. *Cohesion in English*. London: Longman.

- Nedoluzhko, A., Mírovsky, J., Ocelák, R. & Pergler, J. 2009. Extended coreferential relations and bridging anaphora in the Prague dependency treebank. In: *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009),* 1–16.
- Postolache, O., Cristea, D. & Orasan, C. 2006. Transferring coreference chains through word alignment. In: *Proceedings of LREC-2006*.
- Recasens, M., Hovy, E. H. & Marti, M. A. 2010. A Typology of Near-Identity Relations for Coreference (NIDENT). In: *Proceedings of LREC-2010*.
- Schauer, H. & Hahn, U. 2001. Anaphoric cues for coherence relations. In: *Proceedings of RANLP-2001*.
- Vilain, M. Burger, J., Aberdeen, J., Connolly, D. & Hirschman, L. 1995. A model-theoretic coreference scoring scheme. In: *Proceedings of the 6th Message Understanding Conference (MUC-6)*, 45–52.

Eva Hajičová, Barbora Hladká, Pavlína Jínová,
Jiří Mírovský, Šárka Zikánová

Putting things together: Correlating discourse relations with other types of linguistic data

Charles University in Prague
hajicova@ufal.mff.cuni.cz; hladka@ufal.mff.cuni.cz; jinova@ufal.mff.cuni.cz;
mirovsky@ufal.mff.cuni.cz; zikanova@ufal.mff.cuni.cz

The shift from the traditional emphasis on sentence syntax and semantics towards research focusing on text and discourse[19] that started during the last decades of the last century raised a number of research questions one of them being the study of the factors that participate in making a discourse a "neatly woven texture" (Halliday & Hasan 1976). In the present contribution we concentrate on three factors we believe to be crucial and to play an integrating role, being aware that the list of aspects we focus our attention on is far from exhaustive:

(i) **information structure of the sentence** (its topic-focus articulation, TFA), i.e. the relations expressed within a sentence (or a clause);
(ii) relations that exist **between** elementary parts of the discourse, i.e. relations that combine these elements into larger wholes;
(iii) the connective threads carried out via **coreferential links** and other **associative** relations.

All these aspects are taken care of in the annotation scenario of the Prague Dependency Treebank 3.0 (Bejček et al. 2013); a unique feature of this resource is the fact that all phenomena (including also the morphosyntactic and deep syntactic information) are annotated on the same texts which makes it possible

[19] This work has been supported in part by the LINDAT/CLARIN project No. LM2015071 of the MEYS CR.

to follow the interplay between the phenomena as well as to study (and retrieve) them separately (cf. Zikánová et al. 2015).

For the purpose of our presentation we have chosen one piece of continuous text (not taken from PDT but annotated in the PDT-style) and we demonstrate a detailed analysis of the correlations between the deep syntactic relations, the information structure, inter-sentential relations and coreference relations (both anaphoric and bridging) illustrated on that text.[20] We also apply the information structure analysis together with the analysis of coreferential links in order to follow the development of discourse in terms of the **salience** of the elements of the stock of knowledge assumed by the speaker to be shared by him and the hearer; this aspect of the dynamics of discourse is visualized.

An example of the interplay of the above mentioned aspects as reflected and applied in our annotation scheme is given in Figure 1.

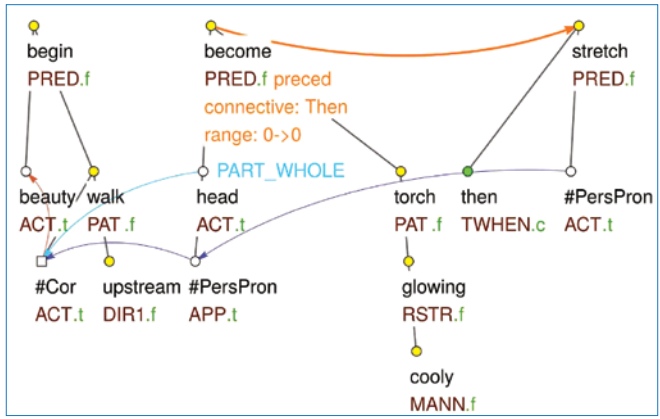


Figure 1. Illustration of the annotation of *a)* the deep syntactic structure (dependency and functors, e.g. ACT, PAT etc.), *b)* topic-focus articulation (based on the values t, f, c), *c)* coreference relations (all arrows in the lower half of the figure), and (d) discourse relations (the arrow in the top of the figure) for the text

[20] The text is a considerably shortened and modified extract of one chapter of Josef Škvorecký's book *Dvorak in Love. A light-hearted dream* (translated from the Czech original Scherzo capriccioso by Paul Wilson, published by Lester & Orpen Dennys Limited, Toronto in 1986). The fact that we have both the original Czech as well the translated English version at our disposal, makes it possible to arrive at some cross-linguistic observations.

segment "The beauty began to walk upstream. Her head became a cooly glowing torch. Then she stretched."

Our contribution will be accompanied by the presentation of the monograph Zikánová et al. (2015) written by members of the Prague TextLink team.

References

- Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J. & Zikánová, S. 2013. *Prague Dependency Treebank 3.0. Data/software* Univerzita Karlova v Praze, MFF, ÚFAL, Prague [http://ufal.mff.cuni.cz/pdt3.0/].
- Halliday, M. A. K. & Hasan, R. 1976. *Cohesion in English*. London: Longman.
- Zikánová, S., Hajičová, E., Hladká, B., Jínová, P., Mírovský, J., Nedoluzhko, A., Poláková, L., Rysová, K., Rysová, M. & Václ, J. 2015. *Discourse and Coherence. From the Sentence Structure to Relations in Text.* Praha: ÚFAL

Jet Hoek, Jacqueline Evers-Vermeul, Ted J. M. Sanders

Discourse Segmentation and Ambiguity in Discourse Structure

Utrecht University, The Netherlands
j.hoek@uu.nl; J.Evers@uu.nl; T.J.M.Sanders@uu.nl

Coherence relations hold between two or more text segments. The process of discourse annotation not only involves determining what type of relation holds between segments, but also indicating the segments themselves. Often, segmentation and annotation are treated as individual steps, and separate guidelines are formulated for each (Carlson & Marcu 2001, Mann & Thompson 1988, Reese et al. 2007, Sanders & van Wijk 1996). Ideally, segmentation results in text segments that correspond to the units of thought related to each other. Although segmenting a text can be fairly simple, there are also fragments in which determining which parts of the discourse are related to each other is more complicated.

When identifying the idea units that are related to each other in a text is not straightforward, this can affect annotation. Fragments containing embedded clauses, for example complement constructions or relative clauses, seem especially prone to ambiguity, since they offer multiple segment candidates. In *1)* a fragment taken from the Europarl corpus (Koehn 2005), for instance, the sentence following *because*, 'it was bringing hard currency into Romania', presents a plausible reason for the BBC to allege that the Romanian authorities knew and approved of the child export. However, it presents an equally plausible reason for Romania to approve of the child export in the first place.

(1) The BBC recently produced evidence that ‚wombs', as they described it, were for sale in Romania – that women were being paid to have children for export to Member States of the European Union. Furthermore, the BBC alleged that this was being done with the tacit approval of the Romanian authorities because it was bringing hard currency into Romania. {ep 00-03-15}

In this presentation we argue that accurate segmentation is in part dependent on taking into account the propositional content of text fragments, and that completely separating segmentation and annotation (i.e. treating it as a two-step process) does not always yield text segments that correspond to the text units between which a conceptual relationship (potentially signaled by a connective) holds (see also Verhagen 2001). We will address ambiguity in discourse segmentation and explore the interaction between segmentation and annotation. In particular, we will focus on the role of connectives in text ambiguity. We propose that connective features that can either allow or resolve ambiguity are for instance the subordinating or coordinating nature of the connective, or the encoding of specific relation characteristics, such as subjectivity or volitionality. Extending our knowledge about variation in discourse structure can help formulate strategies in dealing with constructions or discourse elements for which multiple segmentation options should be considered.

References

- Carlson, L., Marcu, D. 2001. *Discourse Tagging Reference Manual*. [http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf]
- Koehn, Ph. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In: *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand, 79–86.
- Mann, W. C. & Thompson, S. A. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8: 243–281.
- Reese, B., Hunter, J., Asher, N., Denis, P. & Baldridge, J. 2007. *Reference Manual for the Analysis of Rhetorical Structure*. Unpublished manuscript. Austin: University of Texas at Austin.
- Sanders, T. J. M. & van Wijk, C. 1996. PISA – A Procedure for Analyzing the Structure of Explanatory Texts. *Text* 16: 91–132.
- Verhagen, A. 2001. Subordination and Discourse Segmentation Revisited, or: Why Matrix Clauses May Be More Dependent than Complements. In: Sanders, T. J. M., Schilperoord, J. & Spooren, W. P. M. S. (eds.) *Text Representation: Linguistic and Psycholinguistic Aspects,* Amsterdam/Philadelphia: John Benjamins. 337–357.

Murathan Kurfali,[a] Deniz Zeyrek,[a]
Teresa Gonçalves[b]

Automatic prediction of implicit discourse relations in Turkish

[a]*Middle East Technical University (METU), Ankara, Turkey*
[b]*Universidade de Évora, Évora, Portugal*
murathankurfali@gmail.com; dezeyrek@gmail.com; tcg@uevora.pt

Discourse relations can be marked overtly by a connecting device (but, because, so) or left implicit, in which case the adjacency of the clauses gives a hint as to the type of the discourse relation (henceforth, DR). DRs not made explicit with a connecting device are called implicit DRs. They are easily interpreted by readers during semantic interpretation but inducing accurate models that predict their sense is a highly challenging task for discourse processing. Automatic recognition of English explicit discourse connectives (henceforth, DCs) have been performed with an accuracy as high as 93% (Prasad et al. 2014), while prediction of implicit DRs yielded lower accuracies (Pitler et al. 2008, 2009, Louis et al. 2010). The challenge is even higher for Turkish, for which there has not been any work on discourse parsing. Here we focus on predicting two Level 1 implicit DR types in the PDTB sense hierarchy, namely Contingency and Expansion, as well as their frequent types and subtypes in the data. We use several features, including linguistically informed ones. Predicting the sense of the targeted implicit DRs with these features significantly outperforms the baseline of using mere words.

Data, methodology and experiments

The recent extensions on a 40,000-word subcorpus of the METU Turkish Discourse Bank are used. The data contains 747 explicit and 375 inter-sentential implicit DRs annotated for their two arguments and senses in the PDTB style. We chose implicit DRs with a minimum of 50 instances of each DR type (Table 1).

We used Mallet's Maximum Entropy[21] and WEKA's SVM classifier (Hall et al. 2009). Due to limited space, only Maximum Entropy classification results are given, which are overall higher than the results of the SVM classification. We experimented the following five features used in previous studies (Pitler et al. 2009, Hen-Hsen & Hsin-His 2011, Park & Cardie 2012): *Words:* bag of Words in both arguments*; Part-of-Speech (POS):* bag of POS tags in both arguments*; Common Word (CW):* whether or not there is at least one common word in the arguments*; Number of Common Words (NCW):* number of common words in the arguments*; Polarity (POL):* whether or not the main verb of the argument is negated (Turkish being a SOV language, the verb at the end of the argument is taken to be the main verb).

Relation type	Sense	Training set #	Training set with removed explicit DC annotations #	Test set #
Level 1 Class	Contingency	102	290	25
	Expansion	140	336*	34
	TOTAL	242	726	49
Level 2 Type	Contingency: Cause	96	145*	23
	Expansion: Restatement	103	111	25
	TOTAL	199	306	48
Level 3 Subtype	Contingency: Cause: Reason	44	106	10
	Contingency: Cause: Result	51	88	12
	Expansion: Restatement: Specification	62*	90	20
	TOTAL	177	284	52

Table 1. Distribution of Level 1, 2, and 3 relation types of implicit DRs used as training instances.
*The stars show where certain instances were excluded to keep the training data more balanced.

To increase the implicit DRs in the corpus, we removed explicit DCs from the DRs that have the same senses as the implicit ones and added the clause pairs to the training data as in a previous study (Marcu & Echihabi 2002). We assumed these are further instances of the implicit DRs we are interested in. We then experimented with 11 different combinations of the five features to assess their effectiveness at each level separately. 20% of the implicit DRs is allocated as the test set. Our baseline is the F1 scores obtained by using mere words. Our features have led to a considerable increase in the F1 scores over the baseline on all three levels (Table 2). Below, we briefly discuss the results per relation type at their respective levels.

[21] http://mallet.cs.umass.edu/

Contingency, Expansion: Regarding Contingency, although the gain is not so high, the combination of Words and POS is useful for automatic prediction. Artificially increasing the training instances leads to improvement as well. Concerning Expansion, the combination of three features (Words, CW, POS) is highly predictive, which is one of our best F1 results.

Cause, Restatement: For detecting the Level 2 senses, the combination of four features (Words, CW, POL, POS) gives the best F1 results. The POL feature is especially helpful for detecting Restatement, where only 4% of the DRs possess a negated verb. This ratio is 30% in Cause DRs.

Specification, Reason, Result: For all the Level 3 senses, the combination of Words, CW and POS gives the best results, with Specification exhibiting our best performance. The POL feature decreases the F1 score in all cases. In particular, it decreases performance in predicting Reason and Result because overall, approximately 70% the Reason and Result DRs have positive verbs on both arguments, with the remaining ones having a negated verb only in one argument.

DR type		F1 score			Effect of exp. DRs
		Baseline F1	Best F1	Best Feature Set	Difference (F1)
Level 1	Contingency	0.36 (0.28)	0.52 (0.56)	Words + POS	+0.13
Class	Expansion	0.68 (0.79)	0.76 (0.88)	Words + CW +POS	-0.02
Level 2	Cause	0.52 (0.56)	0.68 (0.65)	Words + CW	-0.01
Type	Restatement	0.48 (0.44)	0.73 (0.76)	+ POL +POS	-0.17
				Words + CW	
				+ POL +POS	
Level 3	Specification	0.65 (0.60)	0.76 (0.85)	Words + CW +POS	-0.28
Subtype	Reason	0.28 (0.30)	0.48 (0.30)	Words + CW +POS	0
	Result	0.28 (0.25)	0.46 (0.50)	Words + CW +POS	-0.11

Table 2. F1 (accuracy) results. The last column indicates the effect of adding explicit DRs (stripped of overt DCs) to the training data in terms of the obtained best F1 score.

Overall, our experiments show that except for Contingency, POS information and CW improve the performance for predicting the senses we looked at. The gains for predicting Expansion, its type and subtype are high. But because Expansion is a frequent DR in the data, the features we use may well prove to be useful for detecting the sense of other DR types. We aim to examine this in the future. Artificially increasing the data seemed helpful only for Contingency detection but as various studies have pointed out, the method may not provide a good model for automatic sense prediction (Sporleder & Lascarides 2008, Webber 2009). In further work, we will look into the role of more linguistically

informed features, such as lexical classes, modality, and Levin word classes, as well as possible genre specific features as discussed in Webber (2009: 681).[22]

References

- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. 2009. The WEKA data mining software: an update. *SIGKDD Explorations* 11: 10–18.
- Hen-Hsen, H. & Hsin-Hsi, Ch. 2011. Chinese Discourse Relation Recognition. In: *Proceedings of the 5th International Joint Conference on Natural Language Processing,* 2011. 1442–1446.
- Louis, A., Joshi, A. K., Prasad, R. & Nenkova, A. 2010. Using entity features to classify implicit discourse relations. In: *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 59–62.
- Marcu, D. & Echihabi, A. 2002. An unsupervised approach to recognizing discourse relations. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,* 368–375.
- Park, J. & Cardie, C. 2012. Improving implicit discourse relation recognition through feature set optimization. In: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 108–112.
- Pitler, E., Louis, A. & Nenkova, A. 2009. Automatic sense prediction for implicit discourse relations in text. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 688–690.
- Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A. & Joshi, A. K. 2008. Easily identifiable discourse relations. In: *Companion volume: Posters and Demonstrations, Proceedings of COLING 2008*, Manchester, UK, 85–88.
- Prasad, R., Webber, B. & Joshi, A. K. 2014. Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics* 40: 921–950.
- Sporleder, C. & Lascarides, A. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering* 14: 369–416.

[22] We thank Cem Bozşahin for his useful comments. The second author acknowledges TÜBİTAK (The Scientific and Technological Research Council of Turkey) for supporting her visit to University of Lisbon (scheme no. 2219, project no. 1059B191500896).

- Webber, B. 2009. Genre distinctions for Discourse in the Penn TreeBank. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 674–682.

Veronika Laippala[a],
Aki-Juhani Kyrolainen[a], Johanna Komppa[b],
Maria Vilkuna[b],
Jyrki Kalliokoski[b], Filip Ginter[a]

Sentence-initial Discourse Relational Devices in the Finnish Internet

[a]*University of Turku;* [b]*University of Helsinki*
mavela@utu.fi; akkyro@gmail.com; johanna.komppa@helsinki.fi;
mvilkuna@mappi.helsinki.fi; jyrki.kalliokoski@helsinki.fi; figint@utu.fi

Discourse relations, or coherence relations, are crucial in the understanding of a discourse and its organization (e. g. Mann & Thompson 1988). They carry information about the relationships between states of affairs represented in a text and provide knowledge about its organization. They may but not need to be signaled by discourse relational devices (DRDs), which are often polysemous, context-sensitive and vary across genres (van der Vliet & Redeker 2014).

In Finnish, DRDs predominantly cover conjunctions, utterance particles and some connectives (Hakulinen et al. 2014). The lack of corpora and available tools for Finnish has, however, until recently prevented their large-scale exploration. In this study, we investigate the use of sentence-initial DRDs in the entire Finnish Internet. What markers are the most frequent? How are they typically used, i.e. what kinds of linguistic structures do they typically co-occur with? And, finally, which DRDs are used similarly? From a methodological perspective, we introduce dependency profiles for DRDs to investigate their usage patterns in large corpora.

The DRDs analyzed consist of both a closed class of sentence coordinating conjunctions (as defined in standard grammars of Finnish), utterance initial particles and a selection of connecting elements chosen from a more open class of (potential) sentence initial DRDs. The findings give valuable large-scale information on their use in Finnish, as well as provide a basis for future contrastive approaches. Further, the study enables the comparison of DRDs representing different semantic categories, such as consequence and addition, and syntactic classes, such as strictly defined conjunctions and other markers.

Data and methods

The data come from Finnish Internet Parsebank[23], a web-crawled, nearly 4-million word corpus of Finnish with automatic syntactic analyses. The size of the data offers a unique large-scale insight to the use and variation of DRDs. In addition, the syntactic analyses allow for a deep level investigation, going beyond individual words.

To operationalize the linguistic elements DRDs typically co-occur with, we transformed each text segment with a DRD to unlexicalized syntactic biarcs, three-token subtrees of dependency syntax analyses with the lexical information removed (Kanerva et al. 2014). The removal of the lexical items allows for a generalization of the analysis beyond individual discussion topics to describe structural and more general topical characteristics of texts (Laippala et al. 2015). The analyzed segments include the sentence with the sentence-initial DRD and the preceding sentence. As Prasad et al. (2008) note that 76% of the sentence-initial connectives refer to the previous sentence, this should cover the majority of the text spans the analyzed DRDs refer to. The DRD and its syntactic information are deleted to prevent their unnecessary effect on the clustering result.

The co-occurrence of the DRDs with the unlexicalized syntactic ngrams forms a dependency profile. To analyze the differences between the DRDs, we carried out a cluster analysis, a data-driven method to investigate structuring in the data, on the dependency profiles in R using the package flexclust. This will allow us, in the future, to estimate the most typical syntactic ngrams for each cluster, based on their TF-IDF weights (Spärck 1972), thus giving the opportunity to study the typical syntactic structures co-occurring with the DRDs.

Results

To analyze the DRDs used and to define the ones chosen for further analysis, we first automatically extracted all the sentence-initial words tagged as coordinating conjunctions or adverbs in the Parsebank, and counted the frequencies of these tokens. Then, out of the 100 most frequent possible DRDs, we manually selected 24 DRDs for further analysis. The selection aimed at including different DRDs representing varied part-of-speech classes and semantic categories, and at the same time excluding the most polysemous markers.

[23] bionlp.utu.fi

Table 1 below presents the DRDs included in the analysis, their frequencies as well as their division to clusters. The clustering resulted in a solution of five clusters offering the best fit to the data. Apart from standard coordinating conjunctions, the DRDs mainly represent words mentioned as connectives in Hakulinen et al. (2014), but many of them are simultaneously characterized as modal (e.g., *kyllä* 'sure', *toki* 'certainly') or focus particles (*esimerkiksi* 'for example', *erityisesti* 'in particular') or temporal adverbs (*samalla* 'at the same time', *ensin* 'first').

Discourse Marker	Frequency	Cluster
ensin ‚first, firstly'	156,777	1
ja ‚and'	3,297,640	2
mutta ‚but'	2,339,647	2
eli ‚in other words, or'	644,530	2
niin ‚so, thus'	632,734	2
tai ‚or'	564,269	2
kyllä ‚sure, to be sure'	510,543	2
toisaalta ‚on the other hand'	405,939	2
siksi ‚therefore'	347,605	2
kuitenkin ‚however'	303,967	2
siis ‚then, so, thus'	291,779	2
joten ‚so, therefore'	201,635	2
silti ‚still, yet'	165,455	2
tosin ‚although, but'	362,588	3
näin ‚so, thus, therefore'	800,859	4
sitten ‚then'	702,458	4
toki ‚certainly, of course'	245,563	4
niinpä ‚so, accordingly'	198,747	4
samoin ‚similarly'	174,653	4
lisäksi ‚in addition'	1,183,613	5
myös ‚also'	1,004,886	5
esimerkiksi ‚for example'	466,602	5
samalla ‚at the same time'	296,246	5
erityisesti ‚in particular'	154,143	5

Table 1. The frequency distribution of the DRDs along with their estimated cluster membership.

Ensin 'first' and *tosin* 'although, but, of course' were grouped to separate clusters consisting of only these DRDs, suggesting that their uses stand apart from the others. The cluster 2 includes the coordinating conjunctions and many DRDs expressing manner, consequence and opposition. Interestingly, the focus particle *kyllä* 'sure' is also in the cluster 2 while the other focus particles are in the cluster 5 together with the additives *myös* 'also' and *lisäksi* 'in addition'. The cluster 4, finally, gathers mainly DRDs expressing temporality and consequence, as well as some markers reflecting concession and manner.

To conclude, the division of the DRDs to the clusters seems to reflect partially the semantic relation described by the DRD and its part-of-speech class, although these aspects do not explain the division entirely. The analysis of the typical syntactic ngrams co-occurring with the DRDs of each cluster will shed more light on their typical usage patterns as well as on the features applied in their grouping into clusters. This will be presented in the poster during the TL2016 conference.

References

- Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. R. & Alho, I. 2014. *Iso suomen kielioppi* [The Comprehensive Grammar of Finnish]. Helsinki: Suomalaisen Kirjallisuuden Seura.
- Kanerva, J., Luotolahti, J., Laippala, V. & Ginter, F. 2014. Syntactic N-gram Collection from a Large-Scale Corpus of Internet Finnish. In: *Proceedings of the Sixth International Conference Baltic HLT*.
- Laippala, V., Kanerva, J. & Ginter, F. 2015. Syntactic Ngrams as Keystructures Reflecting Typical Syntactic Patterns of Corpora in Finnish. *Procedia – Social and Behavioral Sciences. Current Work in Corpus Linguistics* 198: 233–241.
- Mann, W. C. & Thompson, S. 1988. Rhetorical Structure Theory: A 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 LREC,* Marrakech, Morocco.
- Spärck, J. K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28: 11–21.
- van der Vliet, N. & Redeker, G. 2014. Explicit and implicit coherence relations in Dutch texts. In H. Gruber & G. Redeker (ed.), *The Pragmatics of Discourse Coherence*. Amsterdam: John Benjamins, 23–52.

Ekaterina Lapshinova-Koltunski[a],
Kerstin Anna Kunz[b] and Anna Nedoluzhko[c]

From monolingual annotations towards cross-lingual resources: An interoperable approach to the analysis of discourse

[a]*Saarland University;* [b]*Universität Heidelberg;* [c]*Charles University of Prague*
e.lapshinova@mx.uni-saarland.de; kerstin.kunz@iued.uni-heidelberg.de;
nedoluzko@ufal.mff.cuni.cz

Introduction

In this paper, we perform a cross-lingual analysis of discourse phenomena in German and Czech, using two corpus resources annotated monolingually within two different frameworks: Functional Generative Description (Sgall et al. 1986) for Czech, and textual cohesion (Halliday & Hasan 1976) for German. We take advantage of the existing resources reflecting systemic peculiarities and realisational options of the languages under analysis. In our previous work (Lapshinova et al. 2015), we have shown that the annotations of the involved resources are comparable if abstract categories are used and only the phenomena with a direct match in German and Czech are taken into consideration. Our analysis is a first step towards unifying separate analyses of discourse relations in Germanic and Slavic languages. At the same time, it demonstrates that the application of 'theoretically' different resources is possible in one contrastive analysis. This is especially valuable for NLP, which uses annotated resources to train language models for various tools.

Related work

Slavic languages have a richer, more fusional morphology than Germanic languages. Even though German has conserved more of the inflectional mor-

phology of Proto-Indo-European than other Germanic languages, it has a more isolating character than Czech. The morphological reduction in German partially results in a less flexible constituent word order as compared to Czech, although some positional options are possible. We expect these contrasts to have an effect on the creation of discourse properties. There is a vast number of theoretical studies comparing Germanic and Slavic languages on a general level (Šticha 2003) and on anaphoric relations (Komárek 1994), whereas quantitative comparisons are rare.

Methodology

For our analysis, several texts of written discourse with comparable topics on economic and political issues were selected. For the German data, nine texts were excerpted from CroCo (Hansen-Schirra et al. 2012) comprising 14930 tokens and 736 sentences in total. The corpus is annotated on several levels including morphological, syntactic and textual information. The information on the latter was annotated with the help of semi-automatic procedures (Lapshinova-Koltunski & Kunz 2014). The Czech texts were taken from the Prague Dependency Treebank (PDT 3.0, Bejček et al. 2013). They are annotated with morphological, analytical and tectogrammatical information. The latter also contains annotation of information structure attributes and inter-sentential relations (Zikánová et al. 2015). Since texts are shorter in PDT than in the German data, 17 texts were excerpted to arrive at a similar number of tokens and sentences (11769 and 763 respectively). In our previous work we attempted to unify the Czech and the German-English frameworks for the annotation of discourse properties creating an interoperable scheme. We use this scheme to test whether this can be applied for contrastive analyses of Czech and German, which can be extended to more general comparisons of Germanic and Slavic languages in the future. The main categories are labelled as IDENTITY, NON-IDENTITY, ELLIPSIS and DISCOURSE RELATIONS. These categories include also subcategories, which we clarify in Table 1 in the next section. We analyse these categories in both languages with respect to their overall distribution, the degree of explicitness, as well as the type of textual categories preferred. We also examine variation in the degree of dependence of these phenomena on lexico-grammatical constraints.

Analyses

Overall, German (GO) and Czech (CZ) texts do not differ significantly in their overall degree of cohesiveness if all four categories are taken together. The differences get pronounced if we compare the distributions for each category (Figure 1). Taking a closer look into the subcategories (Table 2 based on the overall frequencies per category per total number of words in texts), we find that the higher frequencies of IDENTITY in Czech exclusively stem from id3. Qualitative analyses show that more coreference relations are underspecified in Czech than German in terms of explicit accessibility markers, since the definite article does not exist in Czech and accessibility of referents is indicated by information structure more often than in German. By contrast, the frequencies for DISCOURSE RELATIONS are higher in German than in Czech, as German seems to prefer signaling logico-semantic relations by an explicit discourse marker, especially for temporal and expansion relations. Finally, the higher number of NOMINAL ELLIPSIS in Czech than German points to a higher preference for expressing comparison by fragments. This tendency towards implicitness may, however, stem from the greater syntactic flexibility of Czech.

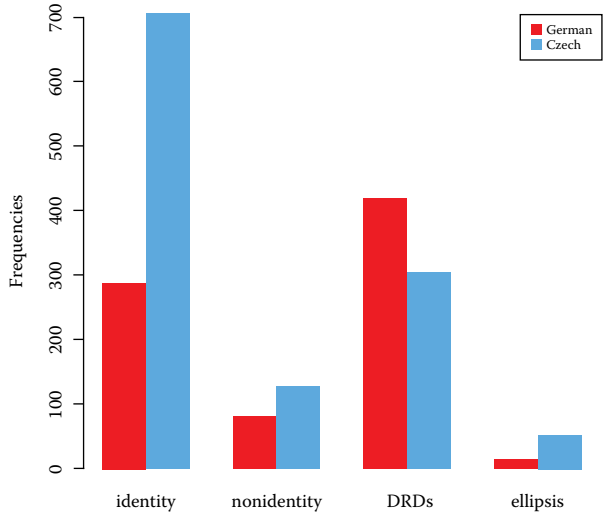


Figure 1. Discourse phenomena in German and Czech

	framework for Czech	framework for German	featID	German	Czech
IDENTITY	coreference with pronouns	coreference with heads (no extended reference)	id1	88.41	97.71
	pronouns with arrows to segments and events	extended reference	id2	38.18	64.58
	NP coreference	coreference with modifiers or def. articles	id3	144.68	597.33
	coreference with the word *same*	general comp.reference	id4	3.35	0.00
	coreference with local and temporal adverbs	coreference with local and temporal adverbs	id5	12.06	10.20
NON-IDENTITY	relations of MERONYMY	relations of MERONYMY	nonid1	52.91	88.37
	bridging CONTRAST	particular comparative reference and antonyms	nonid2	28.80	37.39
DISCOURSE RELATIONS	temporal	temporal	temp	106.50	14.44
	contingency	causal	cont	52.24	66.28
	comparison (contrast)	adversative	comp	79.04	86.67
	expansion	additive	expan	181.51	136.80
ELLIPSIS	textual ellipsis	cohesive ellipsis	ellipsis	14.07	50.13

Table 1. Frequencies of discourse categories

Conclusion

Our preliminary results show that our interoperable scheme permits a multilingual analysis of discourse-annotated corpora originating from different approaches. On the one hand, we are able to validate the interoperable scheme in an application. On the other hand, this indicates possible interoperability in the existing resources, which saves time and effort as no compilation of additional resources is required. Furthermore, the results yield first insights into differences between German and Czech in terms of the annotated phenomena.[24]

[24] We acknowledge support from the Grant Agency of the Czech Republic (grant 16-05394S). This work has been using language resources developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071). The project GECCo has been supported through a grant from the Deutsche Forschungsgemeinschaft (German Research Society).

References

- Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J. & Zikánová, S. 2013. *Prague Dependency Treebank 3.0*.
- Halliday, M. A. K. & Hasan, R. 1976. *Cohesion in English*. London/New York: Longman.
- Hansen-Schirra, S., Neumann, S. & Steiner, E. 2012. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German.* Berlin/New York: Mouton de Gruyter.
- Komárek, M. 1994. On Relative Pronouns in Czech and German (jenž - který, der - welcher). In: Čmejrková, S. & Štícha, F. (eds) *The Syntax and of Sentence and Text: A Festschrift for Frantisek Danes.* Amsterdam: John Benjamins, 359–364.
- Lapshinova-Koltunski, E. & Kunz, K. 2014. Annotating cohesion for multilingual analysis. In: *Proceedings of the 10th Joint ACL – ISO Workshop on Interoperable Semantic Annotation*, Reykjavik, Iceland.
- Lapshinova, E., Nedoluzhko, A. & Kunz, K. 2015. Across Languages and Genres: Creating a Universal Annotation Scheme for Textual Relations. In: Rehbein, I. & Zinsmeister, H. (eds) *Proceedings of the Workshop on Linguistic Annotations, NAACL-2015*, Denver, USA.
- Sgall, P., Hajicova, E. & Panevova, J. 1986. *The meaning of the sentence in its semantic and pragmatic aspects.* Dordrecht: Reidel.
- Šticha, F. 2003. Česko-německá srovnávací gramatika. Prague: Argo.
- Zikánová, S., Hajičová, E., Hladká, B., Jínová, P. & Mírovský, J., Nedoluzhko, A., Poláková, L. Rysová, K., Rysová, M. & Václ, J. 2015. *Discourse and Coherence. From the Sentence Structure to Relations in Text.* Prague: ÚFAL.

Julia Lavid and Lara Moratón

Annotating metadiscourse markers in the English-Spanish MULTINOT corpus: preliminary steps

Universidad Complutense de Madrid
julavid@filol.ucm.es; laramoraton@gmail.com

The work reported in this paper is part of a larger research effort within the MULTINOT project (Lavid et al. 2015), focused on the multidimensional annotation of a register-diversified bilingual corpus of comparable and parallel English and Spanish texts with lexicogrammatical, semantic and discourse features with the aim of developing a multifunctional resource which can be used by a variety of potential users and in a number of theoretical and applied contexts. While previous work by members of the research team has focused on the annotation of features such as modality (Zamorano et al. 2014), global discourse structures, rhetorical relations and thematic patterns (Arús, Moratón & Lavid 2013, Moratón & Lavid 2013), in this paper we report on the recent extension of our annotation tasks to metadiscourse markers (Hyland 2004), as potential realisational devices and markers of some of the previously annotated features. For this task we use a subpart of the MULTINOT corpus, namely, sixty-two newspaper texts, consisting of sixteen news reports, sixteen editorials and twenty letters to the editor, evenly divided into English and Spanish, all of them collected from British and Spanish high-circulation newspapers between 2009 and 2013 and preprocessed with the GATE platform (Cunningham et al. 2002). We found Hyland's distinction into interactive and interactional markers particularly useful as the basis for the design of the annotation scheme, although we decided to use Halliday's terminology and distinguish between 'textual' (interactive) and 'interpersonal (interactional) markers, given its wider acceptance. The former are concerned with ways of organising discourse to anticipate readers' knowledge and include transitions, frame markers, endophoric markers, evidentials and code glosses. The latter focus on the participants of the interaction and "seek to display the writer's persona and a tenor consistent with the norms of

the disciplinary community" (Hyland 2004: 139). These include hedges, boosters, attitude markers, engagement markers and self-mention markers. In the paper we present an annotation scheme for these metadiscourse markers in English and Spanish, report on experiments to validate it and the problems encountered during the annotation phase. We also report on the genre and language-specific variation found in the distribution of these metadiscourse markers in the annotated corpus.

References

- Cunningham, H., Maynard, D. & Bontcheva, K. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).* Philadelphia.
- Hyland, K. 2004. Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing* 13: 133–151.
- Lavid, J., Arús, J., DeClerck, B. & Hoste, V. 2015. Creation of a high quality, register-diversified parallel corpus for linguistic and computational investigations. In: *Current Work in Corpus Linguistics: Working with Traditionally–conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015). Procedia – Social and Behavioral Sciences*, Volume 198, 24 July 2015, 249–256.
- Moratón, L. & Lavid, J. 2013. Thematic Progression Patterns in English and Spanish Newspaper Genres. *Paper presented at COLDOC Conference, 13–14 November 2013.* Published online Actes du IX Colloque de Linguistique de Doctorants et Jeunes Chercheurs. 109–118. [https://coldoc2013fr.files.wordpress.com/2014/11/livret_actescoldoc_version2.pdf]
- Zamorano, J. R., Carretero, M. & Lavid, J. 2014. The annotation of modality and evidentiality in English-Spanish comparable and parallel texts. *Paper presented at EMEL'14 CONGRESS, Evidentiality and Modality in English (6–8 October 2014), Universidad Complutense de Madrid.*

Pierre Lejeune, Amália Mendes, Nuno Martins

Some considerations on the use of main verbs to express rhetorical relations

Centre of Linguistics, Faculty of Arts, University of Lisbon
lejeunepierre@hotmail.com; amalia.mendes@clul.ul.pt;
nunoferreiramartins@gmail.com

Rhetorical relations are typically expressed by discourse structuring devices that ensure textual cohesion and coherence (Halliday&Hasan 1976). Resources such as the PDTB (Prasad et al. 2007; 2008; 2014) target specifically the annotation of these devices, while describing alternative lexicalizations of such relations (AltLex).

Our preparatory work to develop a discourse treebank for Portuguese in the PDTB framework has provided ground for some considerations regarding the status, in intra-sentential coherence, of main verbs that internally carry a causative meaning. We have first focused on the annotation of the rhetorical senses Reason, Result, Pragmatic justification as expressed explicitly by discourse structuring devices (conjunctions, adverbs, phrases and prepositions), taken as elements that express a two-place semantic relation filled by propositional arguments. However, these relations are also frequently marked by other devices (AltLex).

The introduction of the annotation manual mentions that the PDTB "has annotated the argument structure, senses and attribution of discourse connectives" which "are treated as discourse-level predicates that take two abstract objects such as events, states, and propositions" (Prasad et al. 2007). Recently, PDTB's authors have come to the conclusion that "DRMs [Discourse Relational Markers] are a lexically open-ended class of elements which may or may not belong to well-defined syntactic classes" (Prasad et al. 2010: 1024). They specify that one condition for instances of AltLex to be annotated is that "A discourse relation can be inferred between adjacent sentences", which means that a condition – being inter-sentential – is imposed on AltLex that is not imposed

generally on connectives (which include subordinating and coordinating conjunctions with Arg1 and Arg2 in the same sentence). Under a "related work" heading, the authors mention a few articles (Danlos 2006, Kibble 1999, Power 2007) that analyze the verbalization of discourse relations at the intra-clausal level, but it is not clear whether they envisage at all the possibility of annotating instances of the verbs involved as AltLex. One could argue that verbs that mark discourse relations ("discourse verbs", Danlos 2006: 6) should be included, provided they link "events, states and propositions", whatever the grammatical realization of the arguments (nominalization, non-finite clause, etc.) is.

We will focus on causal discourse verbs such as *provocar* 'to cause', *obrigar* 'to force' and *reduzir* 'to reduce'. We will discuss to what extent these verbs have a cohesive function in texts, taking into account their semantic content and the nature of their arguments, based on contexts extracted from the corpus CINTIL (Barreto et al. 2006), a 1M word corpus annotated for part-of-speech and manually revised.

We illustrate the question at hand with examples of the verbs *provocar* and *reduzir*. When considering proposals that decompose lexical meaning into semantic primitives expressed by a conceptual structure (Jackendoff 1983 and 1990) or a lexical conceptual structure (Rappaport&Levin 1988, Pustejovsky 1988), the verb *provocar* 'to provoke' may be expressed as an internally complex event formed by a causative and an existential meaning [CAUSE [TO BE]]. The two arguments of the verb *provocar* are frequently nominalizations, ("a by-product of explicit realization of the relations as verbs and propositions"; "Typically, nominalized forms denote a property, an event or process, or the state resulting from an event" [Kibble 1999: 49]). In (1), where the two arguments are underlined, the verb establishes a causal coherence relation between the event *the refusal of France and Germany (…)* and the event *the recent collapse (…):* [the refusal [CAUSE [the recent collapse TO BE]]. Sentence (1a) could be paraphrased by two clauses linked by a connective, as illustrated in (1b) in English.

(1) a. A força do euro é tal que nem pestanejou com o recente colapso do pacto de estabilidade e crescimento *(PEC)* PROVOCADO pela recusa da França e Alemanha em se submeterem às suas regras de disciplina orçamental. (Público, 2.12.2004) 'The strength of the euro is such that it didn't even flinch with the recent collapse of the Pact for Stability and Growth caused by the refusal of France and Germany to submit themselves to the rules of budget discipline.'

b. The Pact for Stability and Growth recently collapsed because France and Germany refused to submit to its rules.

A different verb type is illustrated by *reduzir* 'to reduce', whose conceptual structure may be expressed as [CAUSE [TO DECREASE]]. The meaning of the verb expresses both a causative value and the variation of a variable's attribute, and consequently the verb is marked both as an alternative lexicalization and part of the second argument[25] in (2).

(2) De acordo com especialistas, uma subida de 10 por cento do dólar REDUZ o crescimento da eurolândia em um ponto. (Público, 2.12.2004) 'According to specialists, a 10% increase of the dollar reduces the growth of euroland in one point.'

In order to annotate verbs like *reduzir* in the PTDB framework, we could rely on Framenet. In Frame Semantics, a frame is constituted by a lexical unit called the target and by frame elements that combine with it. Typically, but not necessarily, the target is a verb and the core frame elements are complements or adjuncts. Some frames semantically encapsulate discourse relations and, quite naturally, their expression through verbs (e. g. for contingency relations: [causation], [cause to X], [concessive], [conditional occurrence], [creating], [evidence]). A system of annotation layers allows lexical units to be annotated at the same time as targets and frame elements. A similar technique might be used for annotating Altlex verbs twice: as DRMs (discourse relation markers) and as part of the Arg 2 (in the case of *reduzir*, this method would account for both the semantic elements *to cause* and the predicate *to decrease* applicable to Arg 2).

Contexts such as those illustrated in (1) and (2) are at the crossroad between syntax and discourse and consequently challenge the limits of the annotation performed in the framework of the PDTB, namely the concepts that we explore in this paper of Alternative Lexicalization (taking into account the fact that they "convey more than just the meaning of the relation", Prasad et al. 2010: 1027)[26] and nominalization.

[25] Danlos (2006) makes a similar distinction between two categories of causal verbs: "Besides cause, there exists a number of causal verbs. On the one hand, there exist other verbs such as provoke, launch, trigger, etc., which are quite similar to cause. On the other hand, there exist causative verbs which lexically encode the effect". She gives *to irritate, to break* and *to give a headache* as examples of the second category.

[26] One important element conveyed by causal verbs, if we compare them to the corresponding connectives, is modality (epistemic, axiological and deontic). For a discussion on modality with notions such as *letting, hindering, helping,* cf. Wolff (2002).

References

- Barreto, F., Branco, A., Ferreira, E., Mendes, A., Bacelar do Nascimento, M. F., Nunes, F., Silva, J. 2006. Open Resources and Tools for the Shallow Processing of Portuguese: the TagShare project. *Proceedings of the V International Conference on Language Resources and Evaluation LREC 2006*, Genova, 2006, 1438–1443.
- Danlos, L. 2006. "Discourse Verbs" and Discourse Periphrastic Links. In C. Sidner, J. Harpur, A. Benz, P. Kühnlein (eds) *Proceedings of the Second Workshop on Constraints in Discourse*. Maynooth, Ireland, 59–65.
- Halliday, M.A.K. & Hasan, R. 1976. *Cohesion in English*. London: Longman.
- Jackendoff, R. 1983. *Semantics and Cognition*. Cambridge: The MIT Press.
- Jackendoff, R. 1990. *Semantics structures*, Cambridge: The MIT Press.
- Kibble, R. 1999. Nominalisation and rhetorical structure. In G-J. M. Kruijff & R. T. Oehrle (eds.) *Proceedings of ESSLLI Formal Grammar conference*, Utrecht, 1999, 49–60.
- Power, R. 2007. Abstract verbs, in *ENLG '07*. In: *Proceedings of the Eleventh European Workshop on Natural Language Generation*, Morristown, USA, Association for Computational Linguistics, 93–96.
- Prasad, R., Miltsakaki, E., Dinesh, A. L., Joshi, A., Robaldo, L., Webber, B. 2007. *The Penn Discourse Treebank 2.0 Annotation Manual*. The PDTB Research Group.
- Prasad, R., Dinesh, A. L., 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*, Marrakech, 2961–2968.
- Prasad, R., Joshi, A., Webber, B. 2010. Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Beijing, 2010, 1023–1031, http://www.aclweb.org/anthology/C10-2118, accessed 12 March 2016.
- Prasad, R., Webber, B., Joshi, A. 2014. Reflections on the Penn Discourse TreeBank, Comparable Corpora and Complementary Annotation. *Computational Linguistics* 40 (4): 921–950.
- Pustejovsky, J. 1988. The geometry of events. In: C. Tenny (ed.) *Studies in Generative Approaches to Aspect*. Cambridge: The MIT Press, 19–39.
- Rappaport, M., Levin, B. 1988. What to do with θ-- roles. In: W. Wilkins (ed.) *Syntax and Semantics 21: Thematic relations*. New York: Academic Press, 7–36.

- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Scheffczyk, J. 2006. *FrameNet II: Extended Theory and Practice*. Berkeley: International Computer Science Institute.
- Wolff, P. 2002. Models of causation and causal verbs. In: M. Andronis, C. Ball, H. Elston and S. Neuval (eds.) *Papers from the 37th Meeting of the Chicago Linguistics Society, Main Session*, Vol. 1., Chicago: Chicago Linguistics Society, 607–622.

Barbara Lewandowska-Tomaszczyk[a],
Paul A. Wilson[b]

Categories and Annotation of Negative Emotionality Discourse Markers in Spoken Language

[a]*University of Lodz; State University of Applied Sciences in Konin, Poland*
[b]*University of Lodz, Poland*
blt@uni.lodz.pl; p.wilson@psychology.bbk.ac.uk

The present paper is a follow up of two of our contributions relating to a contrastive analysis of negative emotion pragmatic markers generated from spoken parts of the British National Corpus and the National Corpus of Polish. The materials are accompanied by relevant recordings and in Polish additionally by automatically generated pitch profiles (Fig. 1). With respect to English, a simplified pitch change annotation (Demenko et al. 2006) is planned to be provided. To analyze the contrasts in more detail, translational (parallel) English-to-Polish and Polish-to-English corpora (Cartoni et al. 2013) are consulted, available at http://pelcra.clarin-pl.eu/. The focal research questions refer to the identification of the negativity elements in the linking devices, particularly those which convey higher *emotional arousal* in the sense of Dziwirek & Lewandowska-Tomaszczyk (2010) and Lewandowska-Tomaszczyk & Wilson (2011). The main goal of this work is to present a corpus-based functional cross-linguistic scale of negative emotion markers in English and Polish.

As proposed before (Lewandowska-Tomaszczyk, 1996, 2004), elements which possess either overt or opaque properties of negativity are both a cognitively more conspicuous, more salient as well as more powerful rhetorical device in discourse than less marked corresponding positive forms. Negative emotions are also less controllable and potentially more revealing with regard to the mental state and stance expression than positive emotions. Therefore, the tracing of the elements of negativity in discourse markers can contribute to the emotion and negation research as well as uncover new vistas in the analysis of discourse markers, also contrastively. Classes of emotional negativity senses

evolve through *frame-switching* (Fillmore 1982) of the content represented in the presupposed part of the utterance, typically preceding a structure involving the negative marker.

Relevant categories and uses are identified in the corpora (http://spokes.clarin-pl.eu/) and their criterial properties and cross-linguistic typology proposed in order to provide some more explicit Event Linking Device *annotation clues*. The primary prosodic clues are considered (as in Fig. 1), particularly pitch and stress patterns generated from the recording of spoken utterances (Pęzik 2014), and in the categories of *contextual* types, a larger context is consulted. The basic category types involve the clines between the *basic-social* and *(un)pleasantness* emotion criteria, mapped onto the *direct - indirect* negativity modes, *dialogic* versus *(internally) monologic* as well as *primary* and *contextual* types of the emotional expressive utterances.

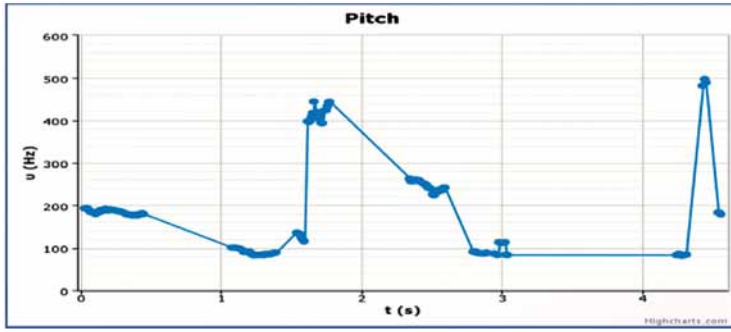


Figure 1. Pitch-stress pattern Pol. […] ***COŚ** ty, **NIE** chcę* lit. '**WHAT** do you [lexicalized negative emotionality (disagreement) marker], I do **NOT** want'

The basic category emotion pattern types proposed involve i.e., Direct Negative Dialogic Frames (with a possible confrontational element and an increasing emotional arousal scale (emphasis) *not at all, not really,* e.g., Pol. **no nie powiesz** żebym ja cię namówiła Eng. **well, you can't say** I made you do it', Pol. strong *(no) co(ś) ty!, (no) wiesz!),* Indirect Negative Dialogic Frames (a weaker arousal scale e.g., Pol. tak było codziennie i było atmosfera była taka że **no nie wiem** żyć nie umierać) with a class of Ironic (counterfactual) Frames (*no, great, really!* Pol. no..no dobrze 'the atmosphere was such that. **well, I don't know.** great, well..all right'), Internal Monologic Expressive Negative Frames *(oh, no, not that again!),* and a range of Contextual Expressive Negative Frames, engaging vari-

ants of the Adversative *but*, and the Temporal/Confrontative *less, while* scales. The categories and scales are not evenly distributed in Polish and English and sets of unique discourse markers in each of the languages (e.g., *zanim* in Polish Eng. strong 'before'/'unless') can be identified, which uniquely link particular structures in one language but not necessarily in the other.[27]

References

- Cartoni, B., Zufferey, D. & Meyer, Th. 2013. Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation Spotting Technique. In: Dipper, S., Zinsmeister, H. & Webber, B. (eds) *Discourse and Dialogue. Beyond Semantics: the Challenges of Annotating Pragmatic and Discourse Phenomena*. Vol. 4, No 2. 65–86.
- Demenko, G., Grocholewski, A., Wagner, A. & Szymanski, M. 2006. Prosody Annotation for Corpus-based Speech Synthesis. In: Warren, P. & Watson. C. L. (eds) *Proceedings of the 11th Australian International Conference on Speech Science and Technology. Australian Speech Science and Technology Association*, 460–466.
- Dziwirek, K. & Lewandowska-Tomaszczyk, B. 2010. *Complex Emotions and Grammatical Mismatches*, Berlin: Mouton de Gruyter.
- Fillmore, Ch. 1982. Towards a Descriptive Framework for Spatial Deixis. In: Jarvella, R. J. & Klein, W. (eds) *Speech, Place and Action*. London: John Wiley. 31–59.
- Lewandowska-Tomaszczyk, B. & Wilson, P. A. 2011. Culture-Based Conceptions of Emotion in Polish and English. In: Goźdź-Roszkowski, St. (ed.) *Explorations across Languages and Corpora. PELCRA 2009.* Frankfurt am Main: Peter Lang. 229–239.
- Lewandowska-Tomaszczyk, B. 1996. *Depth of Negation. A Cognitive Semantic Study.* Lodz: Lodz University Press.
- Lewandowska-Tomaszczyk, B. 2004. Conceptual Blending and Discourse Functions. *Research in Language* 2: 33–47.
- Pęzik, P. 2014. Graph-based Analysis of Collocational Profiles. In: Jesenšek, V. & Grzybek, P. (eds) *Phraseologie im Wörterbuch und Korpus (Phraseology in Dictionaries and Corpora), ZORA 97*, Maribor, BielskoBiała, Budapest, Kansas, Praha, Filozofska fakuteta.

[27] The present study is carried out within the COST Action IS1312 *Structuring Discourse in Multilingual Europe* (TextLink).

Amália Mendes and Pierre Lejeune

LDM-PT – A Portuguese Lexicon of Discourse Markers

Centre of Linguistics / Faculty of Arts, University of Lisbon
amalia.mendes@clul.ul.pt; lejeunepierre@hotmail.com

The Lexicon of Discourse Markers (LDM-PT) provides a set of lexical items in Portuguese that have the function of structuring discourse and ensuring textual cohesion and coherence at intra-sentential and inter-sentential levels (Halliday & Hasan 1976). Each connective is associated to the set of its rhetorical senses, following the PDTB typology (Prasad et al. 2008).

We take discourse markers as a broad category that includes cohesive devices and also pragmatic markers with interactional and modal meanings (Cuenca & Marín 2009: 903) but we focus for now on discourse connectives. Our immediate goal is to provide data for the annotation of discourse relations in a Portuguese Discourse Treebank, although a listing of discourse connectives will certainly prove to be useful for applications dealing with tasks such as parsing, text processing and summarization of Portuguese. Lexical resources available for Portuguese deal essentially with content words and even those focusing on multi word expressions favour content expressions. However, the DPDE online[28] does provide a Portuguese equivalent to the set of Spanish discourse particles, and an experiment in the fully automatic identification of multilingual lexica, including Portuguese has been reported (Lopes et al. 2015).

We consider that discourse connectives do not vary regarding inflection, they express a two-place semantic relation, have propositional arguments and are not integrated in the predicative structure. This includes conjunctions, adverbs and phrases, but also prepositions, which we consider in our list of connectives, an option that is common to the German lexicon DiMLex (Stede 2002) and the French lexicon LEXCONN (Roze et al. 2012).

[28] Antonio Briz–Salvador Pons Bordería–José Portolés (dirs.) *Diccionario de partículas discursivas del español*. online since 2003, www.dpde.es.

The identification of discourse connectives was first performed through a constrastive approach to English, based on the parallel Europarl corpus and on the list of connectives labelled in the PDTB. We locate discourse markers in the English corpus and inspect the Portuguese sentences to identify the corresponding connectives. We apply a manual approach with several goals in mind: to procure fully accurate data, to identify potential new senses of the Portuguese connectives, to spot semantic and pragmatic differences between discourse connectives denoting the same sense. The approach is close to the Translation Spotting Technique (Cartoni et al. 2013), although our motivation is not to capture the different meanings of a given connective in the source language but to acquire a diversified set of connectives in Portuguese. The manual identification of connectives based on a contrastive language analysis brings our attention to the limits of the category (for instance, the case of referential NPs that perform a cohesive function) and to other strategies that express coherence relations between text spans, such as paraphrases. This approach is now complemented using our preparatory work to develop a discourse treebank for Portuguese in the PDTB framework by annotating texts of the corpus CINTIL (Barreto et al. 2006), a 1M word corpus annotated for part-of-speech and manually revised. Our annotation focuses on discourse connectives (conjunctions, adverbs, phrases and prepositions), taken as elements that express a two-place semantic relation filled by propositional arguments.

The lexicon is structured as pairs of discourse connectives/rhetorical senses, so as to cover polysemous connectives. The lexicon includes at the moment 210 pairs of discourse connectives/rhetorical senses and is, for now, implemented in excel format (an illustration of the discourse connectives is provided in Table 1).

Portuguese DM	Rhetorical Relation	Other rhetorical relations	English DM
com efeito, é que, na medida em que, pois, porque, visto (que)	reason	Pragmatic Cause: Justification	for
daí (que), de onde, por conseguinte, por consequência, portanto	result		hence
apesar de, embora, contudo	contrast		though

Table 1: Discourse Markers in the LDM-PT

Additional information on the category of the connective is provided in a required field *Category* (conjunction, preposition, adverb), and other optional information is encoded, such as the equivalent English connective, a corpus

example and restrictions on the context (e.g., the presence of a negative particle, mood selection (indicative or subjunctive), modifiers). The latter might be especially important to deal with connectives that share a common rhetorical sense although they don't occur in the same contexts since "connectives are not always interchangeable and therefore cannot be treated as equivalents" (Cartoni et al. 2013: 79).

We also include in the lexicon Alternative Lexicalizations (AltLex), i.e, alternative expressions that denote a cohesive relation, making it redundant to supply an implicit connective in the context (Prasad et al. 2010: 1025). For instance, the cohesive relation *contrast* is frequently denoted through the following AltLex: *acontece que* 'it happens that', *diga-se que* 'let it be said that', *dito isso* / *posto isso* 'this being said', *não deixa de ser verdade que* 'it is nevertheless true that'. We have also encountered borderline cases of intra-sentential discourse relations marked by a main causative verb (Danlos 2006), such as *provocar* 'to provoke', obrigar 'to force', *reduzir* 'ro reduce', which typically establish a causal coherence relation between two nominalizations (Lejeune et al. this volume).

The lexicon is viewed as an open list that integrates both the results of the contrastive analysis between English and Portuguese discourse connectives and of our corpus annotation following the PDTB model.

References

- Barreto, F., Branco, A., Ferreira, E., Mendes, A., Bacelar do Nascimento, M. F., Nunes, F., Silva, J. 2006. Open Resources and Tools for the Shallow Processing of Portuguese: the TagShare project. *Proceedings of the V International Conference on Language Resources and Evaluation LREC 2006*, Genova, 2006, 1438–1443.
- Briz, A., Pons Bordería, S., Portolés, J. (dirs.) *Diccionario de partículas discursivas del español*, online since 2003, www.dpde.es.
- Cartoni, B., Zufferey, S., Meyer, T. 2013. Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation Spotting Technique. *Dialogue and Discourse* 4 (2): 68–86.
- Cuenca, M. J., Marín, M. J. 2009. Co-occurrence of discourse markers in Catalan and Spanish oral narrative. *Journal of Pragmatics* 41: 899–914.
- Danlos, L. 2006. "Discourse Verbs" and Discourse Periphrastic Links. In C. Sidner, J. Harpur, A. Benz, P. Kühnlein (eds) *Proceedings of the Second Workshop on Constraints in Discourse*. Maynooth, Ireland, 59–65.

- Halliday, M.A.K. & Hasan, R. 1976. *Cohesion in English*. London: Longman.
- Lejeune, P., Mendes, A., Martins, N. 2016. Some considerations on the use of main verbs to express rhetorical relations, this volume.
- Lopes, A., Matos, D., Cabarrao, V., Ribeiro, R., Moniz, H., Trancoso, I., Mata, A. I. 2016. Towards Using Machine Translation Techniques to Induce Multilingual Lexica of Discourse Markers. March 2015, http://arxiv.org/abs/1503.0914, accessed 15 January 2016.
- Prasad, R., Dinesh, A. L., 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*, Marrakech, 2961–2968.
- Prasad, R., Joshi, A., Webber, B. 2010. Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Beijing, 2010, 1023–1031, http://www.aclweb.org/anthology/C10-2118, accessed 12 March 2016.
- Roze, C., Danlos, L., Muller, P. 2012. Lexconn: a French lexicon of discourse connectives, *Revue Discours*, http://discours.revues.org/8645.
- Stede, M. 2002. DiMLex: A Lexical Approach to Discourse Markers. In: A. Lenci & V. Di Tomaso (ed.) *Exploring the Lexicon - Theory and Computation*, Alessandria (Italy): Edizioni dell'Orso.

Philippe Muller, Juliette Conrath, Stergos Afantenos, Nicholas Asher

Data-driven discourse markers representation and classification

IRIT, Toulouse University, Toulouse, France
muller@irit.fr; j.conrath@gmail.com; stergos.afantenos@irit.fr; asher@irit.fr

The literature on discourse markers and rhetorical relations contains many different classifications of discourse connectives, drawing upon a wide range of evidence including textual cohesion (Halliday & Hasan 1976), hypotactic conjunctions (Martin 1992), cognitive plausibility (Sanders et al. 1992), substitutability (Knott et al. 2001), and psycholinguistic experiments (Louwerse 2001). Due to the different theoretical aspects at the basis of each classification and their motivations, there is a general lack of consensus on the characterization of discourse connectives. Additionally, the usual limitations of manually constructed resources apply in this domain as well: the process is very labor-intensive and the resulting resources are often incomplete.

Because of this, some research has been directed at constructing such classifications of connectives automatically. The idea is to use non-biased evidence from natural instances of connective usage in large corpora to induce empirically-grounded classes. Hutchinson (2004) aimed to automatically acquire the meaning of discourse connectives with regard to three aspects often found in hand-coded taxonomies: polarity, veridicality and type. The latter aspect concerns the type of relation expressed by a connective. In this approach, classes were first manually defined for each aspect. Concerning types, three distinct classes are defined: additive, temporal and causal. Instances of connective usage were then extracted from a large corpus based on string patterns. Features were extracted to describe each instance, in terms of lexical co-occurrences in the clauses linked by the connective, as well as other linguistic information. A portion of these instances was then manually annotated to produce training data.

Finally, classification models were trained on this data to obtain a classification which was compared to a golden standard compiled from previous manual classifications (Knott et al. 2001, Louwerse 2001).

Although this approach obtained highly accurate results with respect to this gold standard, it doesn't seem to appropriately solve the problems we mentioned. Indeed, the classes to which connectives are assigned have been decided manually, and are thus necessarily biased. Additionally, since the evaluation of the results is based on previous manual classifications, they can only prove that they manage to recreate a similar classification, or, perhaps, to validate existing ones. Finally, the reliance on manually annotated instances to learn the models implies the necessity of important manual work prior to the applicability of this method, as well as a bias relative to the annotators. A more suitable approach to alleviate these problems is that of Alonso et al. (2002), who propose to use a clustering method to automatically group instances of connective usage extracted from a large corpus. It should be noted however that the aim is not to classify connectives directly, but to classify instances of their usage in context.

The main goal of clustering is to identify partitions in an unstructured set of objects described by certain features. This identification relies only on these features, and no annotated data is required. Instead of manually identifying classes to which the instances need to be assigned, only the number of clusters needs to be defined. Groups of instances are then created based on their similarity with respect to the features. Alonso et al. (2002) rely on two sets of features. The first set is derived from a hand-coded lexicon of connectives with syntactic, discourse segmental, and rhetorical information, including "rhetorical content", which consists in relations such as reinforcement, concession, consequence or enablement. The second set of features is based on shallow text processing of the instances, and contains features relative to the position of the connective in the segment, the words surrounding the connective, the presence of a negation, etc. The analysis of the results demonstrates that the clusters contain mostly instances with similar syntactic behavior of the connective. Various rhetorical contents can be found across clusters, and there is no clear-cut distinction between subordinating and coordinating connectives, contrary to what is found in manual and supervised classifications.

We propose a different approach, with a slightly different goal. We aim to automatically derive empirically-grounded clusters of connectives based on the significance of association between connectives and pairs of verbal predicates in context. In order to arrive at such clusters, we use co-occurrence data we collected on the English Gigaword corpus, yielding triplets consisting of two

predicates and a discourse marker, and their occurrence count. The association measure of each triplet is then computed with a variant of pointwise mutual information (Conrath 2015). We used the PDTB list of discourse markers as a basis, with no *a priori* grouping. We thus produce a matrix of dimensions {number of verb pairs} x {104 connectives}. When a pair never appears with a certain connective, the score is set to zero in the corresponding position of the matrix. In effect, each verb pair is thus represented by a set of 104 feature values. Finally, we apply a dimensionality reduction procedure in the feature space, that is the space of the connectives. We use NMF, non-negative matrix factorization, a technique that produces a lower-dimensionality representation while conserving positive values in the new dimensions, so that the results are more easily interpretable. The number of dimensions k to which the feature space is factorized needs to be predefined. In order to mirror the granularity of the usual descriptions of human analysts, we set this number to six dimensions, for easier manual analysis.

In Figure 1, we show the strength of association of each connective in each group.

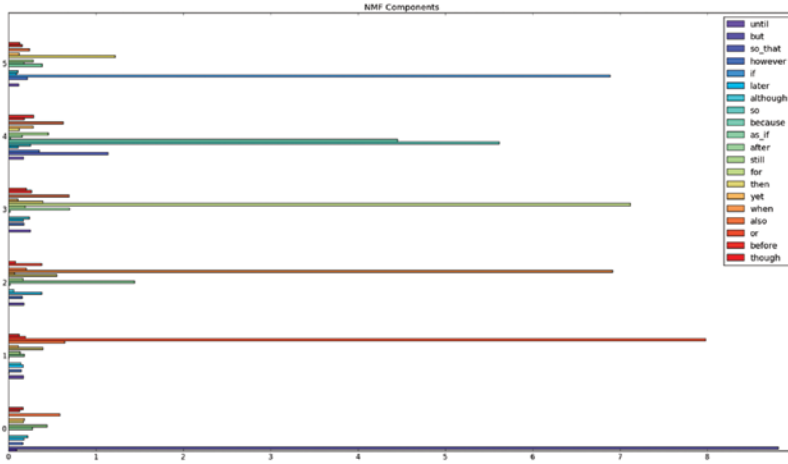


Figure 1. Weight of each connective for each group

In Table 1 we show the most prominent markers for each new dimension. Many markers appear with much lower weight in the new space, mainly because their usage is too infrequent.

Dimension 1	but, (also, still)
Dimension 2	or, (also)
Dimension 3	when, after, (then)
Dimension 4	for, (also)
Dimension 5	because, so, so that, (still, also)
Dimension 6	if, then

Table 1. Most prominent markers for each dimension after NMF (sometimes in several dimensions)

Some of these dimensions isolate broad classes of semantic relations: for instance dimension 1 represents a contrastive or opposition type relation, dimension 2 an alternation type relation, dimension 3 a temporal one, dimension 5 a causal one (regrouping explanation, result and intentional causal relations like goal), dimension 6 a conditional or suppositional relation. Dimension 4 might also be a causal intentional family of relations, though the distinction between dimensions 4 and 5 is not so clear.

Parallelism marked with also seems to pervade several of the semantic classes, but this is not so surprising for theories in which several relations may be at play between two discourse units: parallelism often combines with several other relations. We can see this as multiple markers in clauses are relatively common and natural when one of the markers is an indication of parallelism *also, too, as well because he also, as a result she also, but he … too, although she … as well* (Asher 1993).

References

- Alonso, L., Castellón, I., Gibert, K. & Padró, L. 2002. An empirical approach to discourse markers by clustering. *Topics in Artificial Intelligence* 2504: 173–183.
- Asher, N. 1993. *Reference to abstract objects in discourse*. Amsterdam: Kluwer Academic Publishers.
- Conrath, J. 2015. *Unsupervised extraction of semantic relations using discourse information.* PhD thesis, Toulouse University. [https://www.irit.fr/EVT/PDF/evt-265-fr.pdf]
- Halliday, M. A. K. & Hasan, R. 1976. *Cohesion in English.* London: Longman.

- Hutchinson, B. 2004. Acquiring the meaning of discourse markers. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. 684.
- Knott, A., Oberlander, J., O'Donnell, M. & Mellish, Ch. 2001. Beyond elaboration: The interaction of relations and focus in coherent text. In: Sanders, T., Schilperoord, J. & Spooren, W. (eds) *Text representation: linguistic and psycholinguistic aspects*. Amsterdam/Philadelphia: John Benjamins, 181–196.
- Louwerse, M. 2001. An analytic and cognitive parametrization of coherence relations. *Cognitive Linguistics* 12 (3): 291–316.
- Martin, J. R. 1992. *English text: System and structure*. Amsterdam: John Benjamins.
- Sanders, T. J. M., Spooren, W. P. M. & Noordman, L. G. M. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15 (1): 1–35.

Arne Neumann, Uladzimir Sidarenka, Manfred Stede

A new approach to merging and querying parallel text annotations

Applied Computational Linguistics, University of Potsdam, Germany
neumann@mailbox.org; Uladzimir.Sidarenka@uni-potsdam.de;
stede@uni-potsdam.de

For many purposes, an annotated corpus is particularly useful when *multiple layers* of different annotation types are available for the texts. In our scenario, the 175 German texts in the Potsdam Commentary Corpus (PCC, cf. Stede & Neumann 2014), a selection of short editorials from a local newspaper, have been manually or semi-automatically annotated for sentence syntax, coreference, connectives/arguments, and rhetorical structure. Being able to issue queries that combine different layers now allows for systematically studying correlations between the different realms, and for testing to what extent one annotation layer could contribute to automatically inferring another.

Primary data

One of our initial decisions in the PCC project was to use layer-specific tools for the primary annotation of the data. At the time, there was no "universal" annotation tool available, which would support different types of marking in the same way; and also today, while many new tools have emerged and many of them are quite versatile, we still find that for some of our layers a specific tool is to be preferred in order for annotators to have a user interface that is intuitive to use and best supports the task. This holds in particular for annotation layers where a semi-automatic annotation is to be preferred over a completely manual

one. In our setting, this is the case for syntax (the 'annotate'[29] tool suggests a syntax tree, which the annotator may re-arrange) and for connectives ('Conano'[30] suggests connectives and their argument spans). Coreference is manually annotated with 'MMAX2'[31], which highlights co-referring expressions in the text view. Finally, for RST trees we use 'RSTTool'[32], which allows for an intuitive graphical tree construction.

Merging

The first technical challenge is to "bring together" the different annotation formats produced by the tools. In our new approach this is achieved by the *discoursegraphs* component (Neumann 2015), which is a freely available library[33] of import/export modules for a great variety of annotation tools and formats, and which also takes care of merging multiple layers of annotation of a document into a single directed graph, i.e., to make sure that all parallel annotations are mapped to the correct spans of the base text. Once the layers are merged, the graphs (one for each document) can be queried using the Python programming language or exported for inspection by specialized graph analytics tools.

Querying

While database functionalities are not required to search the corpus across the different layers, they may be worth considering, e.g., if your corpus is large, or if you prefer a concise query syntax. In line with our choice of data format just described, we opted for the neo4j graph database[34], where users' queries apply directly to graphs storing the different portions of information. Thus we implemented a mapping from *discoursegraphs* to the specific format *geoff* used by neo4j.

---

[29] This tool is no longer operational. A successor is synpathy (http://www.mpi.nl/tools/synpathy.html).

[30] http://angcl.ling.uni-potsdam.de/resources/conano.html (see Stede & Heintze 2004)

[31] http://mmax2.sourceforge.net/

[32] http://www.wagsoft.com/RSTTool/

[33] https://github.com/arne-cl/discoursegraphs

[34] http://neo4j.com/

Using this database with the PCC corpus, we can, for example, determine how often a coreference chain starts in a discourse segment that has nuclear or satellite status in RST (it turns out that the ratio is roughly 3:2). Or, we find that syntactically-subordinated clauses are mostly satellite segments in RST, but – against some predictions in the literature – not always (the ratio is roughly 1:5). Similarly, we can now determine how often discourse relations are being signaled by which connectives (or none), and in what cases annotators arrived at different "scope" decision in the RST analysis and the connective annotation.

Outlook

Our approach has been implemented for the PCC, but the idea can easily be transferred to other corpora, also in other languages, provided that annotations have been produced on the same base text; this in particular means that tokenization must not produce different results for different layers. Another prerequisite is that the annotation tools are being supported by the *discoursegraphs* library; these are listed in the paper mentioned above. All the tools that we have mentioned here are freely available for research purposes.

References

- Neumann, A. 2015. Discoursegraphs: a graph-based merging tool and converter for multilayer annotated corpora. In: *Proceeding of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015),* Vilnius/Lithuania. 309–312.
- Stede, M. & Heintze, S. 2004. Machine-assisted rhetorical structure annotation. In: *Proceedings of the International Conference on Computational Linguistics (COLING),* Geneva, 2004
- Stede, M. & Neumann, A. 2014. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In: *Proceedings of the Language Resources and Evaluation Conference (LREC),* Reykjavik, 2014.

Elena Pascual

Annotating discourse units in spontaneous conversations: The challenge of self-repairs

Universitat de València
elena.pascual@uv.es

The analysis of spoken language requires a system of units that is able to describe and analyze discourse phenomena since syntactic units have proven to be too narrow a segmentation tool for spontaneous speech (Narbona 1991). Such a need is evident in the proliferation of various segmentation models (Pons 2014), such as those of Geneva (Roulet 2001), the Sorbonne (Morel & Danon-Boileau 1998), the Val.Es.Co. Research Group (Briz and Grupo Val.Es.Co 2003) and the BDU (Degand & Simon 2009) among others. Despite the fact that these models use different units and employ heterogeneous criteria to define them, the models all face common challenges when accounting for the segmentation and annotation of particular oral phenomena such as speech ruptures, that is fragmentary segments of speech (e.g. hesitations, restarts, repairs, repetitions) that arise in conversations due to the spontaneous construction and planning of discourse.

In most of the mentioned models, truncated segments are considered only partially or not at all: sometimes they are annotated as syntactical units (Degand & Simon 2005); sometimes they are regarded as elements that can appear in some prosodic units and whose study is to be developed (Morel & Danon-Boileau 1998); on other occasions they are simply excluded from the analysis and considered residues (Briz and Grupo Val.Es. 2014). This study departs from the theoretical question of how to deal with truncated segments of speech when segmenting discourse into units, and focuses on the corresponding practical problem of how to label discourse ruptures when annotating discourse units.

In order to find an answer to the previous questions, this paper carries out a case-study of speech ruptures that comprise specific phenomenon named *self-repair*. A self-repair can be defined as the self-correction conducted by speakers

when they try to rectify the troubles that occur when speaking (Schegloff et a. 1977). In the following example, taken from conversation 46 in the *Corpus Val.Es.Co. 2.0.* (Cabedo & Pons 2013) of Spanish colloquial conversations, we observe two truncations marked with hyphens in the transcript that are uttered by speaker "A":[35]

(1) A: entonces se– // se pasó dos dí– y me dijo Viqui *se ha pasado dos o tres días vomitando*[36]

Repair 1

entonces se– // se pasó dos dí–

original utterance repair

Repair 2

se pasó dos dí- y me dijo Viqui *se ha pasado dos o tres días vomitando*

original utterance editing term repair

As illustrated in the figure above, the first repair contains two parts[37]: (1) the original utterance, "entonces se–", enunciated before the interruption "–", and (2) the repair "se pasó dos dí–".[38] The second repair contains three parts: (1) the original utterance, "se pasó dos dí–", (2) an editing term that often consists of a discourse marker or a phrase, "y me dijo Viqui", which is integrated before the repair, and (3) the repair, *"se ha pasado dos o tres días vomitando"*.

Self-repairs offer an insight into speakers' attempts to plan and organize their speech. They often include editing terms that could be considered DRD (Discourse Relational Devices), given that such editing terms help speakers to struc-

[35] This example could be translated into English as follows: A: so she sp- spent two da- and Viqui told me *she spent two or three days throwing up*

[36] The slashes indicate the presence of a pause of 0,5–1 seconds; the italics indicate direct speech.

[37] The description of the parts of a self-repair is taken from Levelt (1983).

[38] Note that, as Schegloff et al. (1977: 636) observe, a repair does not imply either the correction of an error (in the example the speaker does not make a mistake), or the presence of a replacement in the repair (in the example the repair consists of a repetition, not of a replacement).

ture their discourse, as occurs with the discourse marker "bueno" uttered by "B" in the following example:[39]

(2) C: [yo no lo veo][40]
B: [no quieres] tener ese trabajo o yo no querría o me lo pensa- bueno me lo [pensaría]
C: [sí][41]

The aim of this study is to integrate self-repairs and their constituents (truncated segments, editing terms – DRD – and repairs) into discourse units in order to allow a complete segmentation of Spanish colloquial conversations without residues. This paper examines the problem of annotating self-repairs in the context of the Val.Es.Co. model of units, where ruptures of speech are generally considered segments that are non-intentional and external to the informative and syntactic structure of discourse (Pérez 2004: 883). Since truncated segments cannot be complete semantic, syntactic or pragmatic units (Herrero 1995), they are excluded from annotations carried out using the Val.Es.Co model and constitute a residue of approximately 6% of the delimited segments of a conversation (Pascual 2014).

More than 1000 interventions of colloquial conversations from the *Corpus Val.Es.Co. 2.0.* (Cabedo & Pons 2013) are segmented into units. Only self-repairs that include truncated segments and explicit editing terms are taken into account. A linguistic and structural analysis of the prosodic, morphosyntactic, semantic and pragmatic features of the selected self-repairs is carried out. On the basis of these features, a typology of self-repairs[42] is devised, together with a scheme to annotate them as parts or correlates of the minimal monological discourse units of the Val.Es.Co. system: act and subact.[43]

---

[39] Taken from the conversation 44 in Cabedo and Pons, *Corpus Val.Es.Co 2.0.,* n.p.n.

[40] The square brackets mark the presence of simultaneous speech

[41] This example could be translated into English as follows:
C: [I don't see the point]
B: [you don't want] to have that job or I wouldn't want to or I would thi- well I would [think about it]
C: [yes]

[42] Based on Levelt (1983: 44–45).

[43] According to Briz & Val.Es.Co. (2014: 37), the act is the minimal identifiable and intentional unit; the subact (ibid. 53–60) is the minimal informative unit, generally equivalent to an intonation group. Depending on the nature of their semantic content, subacts can be substantive – director (SSD), subordinate (SSS) or topicalized (SSS) – or adjacent – textual

The results of this study comprise a protocol for annotating self-repairs in the segmentation of a conversation into units. In addition, it is shown that, when used in conjunction with a system of discourse units, the linguistic analysis of self-repairs may reveal recurrent patterns of units that tend to constitute these self-repairs. The process of mapping these patterns elucidates the role of self-repairs in some discursive functions, such as planning and mitigating. Finally, a brief description is offered of the annotated types and functions of DRD which constitute the editing terms of self-repairs.

This paper seeks to demonstrate the advantages of a system for annotating discourse units that takes into account residues and that shows their composition at a microanalytical level. This last example displays two annotations into units: (a) does not take into account the sequence of self-repairs, whereas (b) does take self-repairs into account:

(3) A: pero es que eso no es m- es que no- a ver no es malo[44]

(a) A: <A><SAT>pero</SAT> <SAT>es que</SAT> eso no es m- <SAT>es que</SAT> no- <SAT>a ver</SAT> <SSD>no es malo</SSD></A>
Annotation pattern: <A><SAT></SAT> <SAT></SAT> eso no es m- <SAT></SAT> no- <SAT></SAT> <SSD></SSD></A>

(b) A: <A><SAT>pero</SAT> <SAT>es que</SAT> <SSX>eso no es m-</SSX> <SAT>es que</SAT> <SSD>no- <SAT>a ver</SAT> no es malo</SSD></A>
Annotation pattern: <A> <SAT></SAT> <SAT></SAT> <SSX></SSX> <SAT></SAT> <SSD><SAT></SAT></SSD></A>[45]

(SAT), modal (SAM) or interpersonal (SAI) –. For a further description see Briz & Val. Es.Co. (2014: 37–60).

[44] This example could be translated into English as follows:
A: but the thing is that this is not b- the thing is not- let's see is not bad

[45] The annotation is made with TEI tags in XML format, as described in Salvador Pons, Cómo dividir una conversación en actos y subactos, (in preparation):

<A> </A>	Act
<SAT> </SAT>	Textual Adjacent Subact
<SSD> </SSD>	Director Substantive Subact
< SSX> </SSX>	Undetermined Substantive Subact

In conclusion, this study contributes to: (1) implement the annotation and segmentation of colloquial conversations; (2) shed light on the processes of structuring and planning spontaneous discourse; and (3) advocate the importance of analyzing truncated segments of speech, which are currently considered mere syntactical residues.

References

- Briz, A. & Grupo Val.Es.Co. 2003. Un sistema de unidades para el estudio del lenguaje coloquial, *Oralia* 6: 7–61.
- Briz, A. & Grupo Val.Es.Co. 2014. Las unidades del discurso oral. La propuesta Val.Es.Co. de segmentación de la conversación (coloquial), *Estudios de Lingüística del Español* 35 (1): 11–71. [http://infoling.org/elies/35/elies35.1-2.pdf]
- Cabedo, A. & Pons, S. (eds) 2013. *Corpus Val.Es.Co 2.0*. [http://www.valesco.es]
- Degand, L. & Simon, A. C. 2009. Minimal discourse units in spoken French: On the role of syntactic and prosodic units in discourse segmentation. *Discours* 4. [http://discours.revues.org/5852]
- Degand, L. & Simon, A. C. 2005. Minimal Discourse Units: Can we define them, and why should we? In: Aurnague, M. et al. (eds) *Proceedings of SEM-05. Connectors, Discourse Framing and Discourse Structure: From Corpus-based and Experimental Analyses to Discourse Theories*, Biarritz, 14–15 November 2005. 65–74.
- Herrero, G. 1995 Sobre construcciones fragmenradas. *Philologia Hispalensis* 10: 99–113.
- Levelt, W. J. M. 1983. Monitoring and self-repair in speech. *Cognition* 12: 41–104.
- Morel, M-A.& Danon-Boileau, L. 1998. *Grammaire de l'intonation. L'exemple du français oral*, Paris: Ophrys.
- Narbona, A. 1991. Sintaxis coloquial y análisis del discurso, *Revista Española de Lingüística*, 21 (2): 187–204.
- Pascual, E. 2014. *Aproximación a la segmentación del subacto en la conversación coloquial española*. Unpublished Master's thesis. Universitat de València.
- Pons, S. 2014. Discourse segmentation in Romance languages: an overview. In: Pons, S. (ed.) *Discourse Segmentation in Romance Languages*. Amsterdam/Philadelphia: John Benjamins. 1–21.
- Pons, S. in prep. *Cómo dividir una conversación en actos y subactos*.

- Pérez, M. 2004. Sobre algunas estructuras truncadas en la conversación coloquial. *Interlingüística* 14, *Actas del XVIII Encuentro de la Asociación de Jóvenes Lingüistas*. 875–886.
- Roulet, E., Laurent, F. & Grobet, A. 2001. *Un modèle et un instrument d'analyse de l'organisation du discours*. Berne: Peter Lang.
- Schegloff, E. A., Jefferson, G. & Sacks, H. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53 (2): 361–382.

Kateřina Rysová, Eva Hajičová, Magdaléna Rysová, Jiří Mírovský

Several Observations from the Annotation of Discourse Connectives in the Prague Dependency Treebank

Charles University in Prague, Faculty of Mathematics and Physics
katerina.rysova@post.cz; hajicova@ufal.mff.cuni.cz;
magdalena.rysova@post.cz; mirovsky@ufal.mff.cuni.cz

In our presentation, we introduce the annotation of especially multiword discourse connectives in Czech (i.e. expressions like *abychom to shrnuli 'to conclude', z tohoto důvodu 'for this reason', výsledkem bylo 'the result was'* etc.) on the data of the Prague Dependency Treebank (PDT) and on the basis of the annotation results, we focus on the possible borderlines of discourse connectives as a general class.[46]

In the PDT annotation, we distinguish two categories within discourse connectives especially according to their degree of grammaticalization: primary connectives (mostly fully grammaticalized connectives like *therefore, however, while* etc.) and secondary connectives (yet non-grammaticalized phrases like *as a result*, *under these conditions* etc.) (Rysová & Rysová 2015).

In our presentation, we pay attention especially to the comparison of primary and secondary connectives in the PDT annotation mainly in the following points: *a)* frequency (i.e. how often primary and secondary connectives appear in authentic Czech texts in absolute numbers), *b)* semantico-pragmatic rela-

[46] This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic (grant LD15052 *TextLink: Structuring Discourse in Multilingual Europe* and LINDAT/CLARIN project LM2015071), by the international COST Action IS1312 *(TextLink)* and by the Ministry of Culture of the Czech Republic (project DG16P02B016 *Automatic Evaluation of Text Coherence in Czech*).

tions (i.e. how primary and secondary connectives differ in frequency of the individual types of discourse relations), *c)* inter- vs. intra-clausal relations (i.e. whether primary and secondary connectives express discourse relations rather within a sentence or across the sentence boundary).

The results demonstrated that primary and secondary connectives differ mainly in frequency (primary connectives form 94.6% and secondary connectives 5.4% within all discourse connectives annotated in the PDT) and in inter- vs. intra-clausal relations (while primary connectives prefer intra-clausal relations in 70%, secondary connectives inter-clausal relations in 63%).

During the annotation, it has also appeared that the structure of connectives may have a direct influence on syntactic realization of discourse arguments.[47] A group of secondary connectives (consisting of a noun + verb *to be* like *důvodem je 'the reason is,' podmínkou je 'the condition is,' příkladem je 'the example is,' výsledkem je 'the result is'* etc.) prefer nominalization of the second discourse argument in 80% – i.e. in Czech, it is more common to say *důvodem je **napadení** ředitele 'the reason is the **attack** on the director'* than *důvodem je, že napadl ředitele 'the reason is that he attacked the director'*.

In our presentation, we introduce the most important criteria according to which we may characterize discourse connectives in Czech (based on a detailed annotation of the PDT, i.e. almost 50,000 of sentences) and we describe their behaviour in authentic Czech texts.

References

- Asher, N. 1993. Reference to Abstract Objects in Discourse. Dordrecht: Kluwer Academic Publishers.
- Prasad, R., Webber, B. & Joshi, A 2014. Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics* 40 (4): 921–950.
- Rysová, M. & Rysová, K. 2015. Secondary Connectives in the Prague Dependency Treebank. In: Hajičová, E.; Nivre, J. (eds.) *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala, Sweden: Uppsala University. 291–299.

[47] In the PDT, discourse arguments are defined according to Asher's (1993) abstract objects and minimality principle (Prasad et al. 2014).

- Rysová, M. & Rysová, K. 2014. The Centre and Periphery of Discourse Connectives. In: Aroonmanakun, W. et al. (eds.) *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC 28).* Bangkok, Thailand: Department of Linguistics, Faculty of Arts, Chulalongkorn University, 2014. 452–459.

Ted Sanders[a], Vera Demberg[b],
Jacqueline Evers-Vermeul[a], Jet Hoek[a],
Merel Scholman[b], Sandrine Zufferey[c]

How Can We Relate Various Annotation Schemes? Unifying Dimensions in Discourse Relations

[a]*Utrecht University, the Netherlands;* [b]*Saarland University, Germany;*
[c]*University of Bern, Switzerland*
T.J.M.Sanders@uu.nl; vera@coli.uni-saarland.de; J.Evers@uu.nl;
j.hoek@uu.nl; merel_scholman@hotmail.com;
sandrine.zufferey@rom.unibe.ch

Over the last fifteen years, annotating discourse relations has gained increasing interest of the linguistics research community. Indeed, it is a promising and challenging research area, which allows for systematic cross-linguistic comparison at the discourse level. A lot of progress has been achieved thanks to the development of large discourse-annotated corpora. Some leading examples of frameworks used to annotate these corpora are the Penn Discourse Treebank (Prasad et al. 2008), the Rhetorical Structure Theory (RST) Treebank (Carlson et al. 2003) and SDRT (Reese et al. 2007).

However, existing discourse annotation guidelines differ in important aspects, such as the type of relations that are distinguished. Some proposals present sets of approximately 20 relations (Mann & Thompson 1988). The PDTB contains a three-tiered hierarchical classification of 43 sense tags (Prasad et al. 2008), and the annotation scheme used for the RST Treebank distinguishes 78 relations that can be partitioned in 16 classes (Carlson et al. 2003). Hence, it is not clear which and how many categories (for example, *contingency*, *causal,* or *informational*) and end labels (for example, *result, volitional cause*, and *cause-consequence* are all labels for causal relations) are needed to adequately describe and distinguish coherence relations.

One thing that is clear is that annotation has proven to be a difficult task, which is regularly reflected in low inter-annotator agreement scores. Furthermore, it is often hard to compare outcomes of corpus-based studies, because frameworks differ in the precise relations they distinguish. This is unfortunate; it would be much better if all these annotated corpora could be compared.

Our goal is to suggest how the discourse relation annotations used by the different schemes can be mapped onto one another. An important consideration is to be able to represent all of the annotations that the different schemes have considered relevant for discourse relation annotation (also see a proposal for a mapping between SDRT and RST, Benamara & Taboada 2015). More specifically, we will compare PDTB, RST and SDRT in terms of a limited set of dimensions, and show how they map onto each other. For instance, all systems distinguish between positive and negative relations. We will show how this dimension allows for similar clusterings across systems. Some dimensions are similar to a Cognitive approach to Coherence Relations (CCR, Sanders et al. 1992), while additional criteria capture more fine-grained distinctions. We describe discourse relations in terms of this limited set of dimensions and criteria, and show how the various existing proposals can be related to each other. This leads to a unifying proposal, which allows us to 'translate' outcomes from one framework to the terminology of another. This way, we want to contribute to the ultimate goal: to make optimal use of existing corpora and facilitate discussion among researchers working in different paradigmata.

References

- Benamara, F. & Taboada, M. 2015. Mapping different rhetorical relation annotations: A proposal, In: *Proceedings of the 4th Joint Conference on Lexical and Computational Semantics (*SEM),* collocated with the *Conference of the North American Association for Computational Linguistics*. 2015, Denver.
- Carlson, L., Marcu, D. & Okurowski, M. E. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In: van Kuppevelt, J. C. J. & Smith, R. W. (eds.) *Current Directions in Discourse and Dialogue*. Dordrecht: Kluwer Academic Publishers. 85–112.
- Mann, W. C. & Thompson, S. A. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text* 8 (3): 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 Interna-*

tional Conference of Language Resources and Evaluation (LREC 2008), 2008, Marrakech.

- Reese, B., Hunter, J., Asher, N., Denis, P. & Baldridge, J. 2007. *Reference Manual for the Analysis and Annotation of Rhetorical Structure (version 1.0). Technical report.* Austin: University of Texas, Departments of Linguistics and Philosophy. [http://timeml. org/jamesp/annotation_manual. pdf]
- Sanders, T. J. M., Spooren, W. P. M. S. & Noordman, L. G. M 1992. Toward a Taxonomy of Coherence Relations. *Discourse Processes* 15: 1–35.

Tatjana Scheffler, Rike Schlüter, Manfred Stede

Discourse Structuring Devices on Twitter

University of Potsdam, Germany
tatjana.scheffler@uni-potsdam.de; r.schlueter@rejoice.de;
stede@uni-potsdam.de

It is known that the means to establish discourse coherence differ between discourse types. Specifically, spoken and written discourse employ different discourse connectives, rhetorical structures, and other discourse structuring devices. Here we analyze discourse structuring devices in Twitter conversations, a text type that shows properties of informal, spoken-like interaction, features determined by written transmission, as well as additional phenomena innovated in this medium.

Data

Twitter is a social medium composed of microblogs (called 'tweets'). Users frequently reply to other users' tweets, creating branching dialog structures. It has been shown that up to 40% of all tweets are part of such conversations (Scheffler 2014). It is not well-studied which devices users employ in order to establish coherence in these public conversations, which are potentially read (and contributed to) by many unknown participants. We have extracted large numbers of full conversations from German Twitter data1, which we take as the basis of our work in this paper.

Discourse Connectives on Twitter

Our first question concerns the distribution of connectives in Twitter conversation: Do the relative frequencies of usage correlate with those found in other modes? To this end, we identified all connectives (using the base set in

DiMLex, Stede 2002) in 100 conversations, consisting of 451 tweets in total.[48] This yielded 207 connective instances in 165 tweets; i.e., 37% of all tweets contain one or more connective, with an average of 1.25. The five most frequent German connectives on Twitter are (in this order) aber ('but'; 38), und ('and' 37), wenn ('when'/'if'; 35), dann ('then'; 29), and also ('so'; 8) – notice the big gap from rank 4 to rank 5. For comparison, the top five connectives in our German corpus of newspaper editorials (Stede & Neumann 2014) are und, wenn, aber, doch, denn; also is only on rank 16.

The Case of Causal Connectives

In order to answer more specific questions about discourse connectives, we focus on the subclass of causal relations. Even though individual tweets are short, reasons or justifications are frequently given. It is known that register has a large effect on the kind of causal connective employed in German. In spoken German, *weil* is by far the most common causal connective, while *denn* and *da* are almost equally common in some written registers (Wegener 1999). We have compared the frequency of *weil*, *denn*, *da*, and *nämlich* in a corpus of 250,000 tweets and found that similar to the spoken corpora, *weil* by far dominates (more than 5:1)[49]. This suggests an informal style, though Rehbein (2014) showed that this does not extend to the even more informal variant of *weil* with V2 word order, which is frequent in speech but very rare on Twitter.

In a subsequent study, we analyzed 200 instances of *weil* in conversations on Twitter. The goal was to identify which of the semantic levels proposed by Sweetser (1990) the causal connective targets and where its arguments are located. The causal connective *weil* relates two arguments, Cause and Consequence. We found that, in conversations, the two arguments can be located in two different utterances, usually uttered by two different speakers. This phenomenon was found in 47 of the 200 instances.

Sweetser (1990) assumes that the interpretation of connectives not only depends on the form of the sentence but also on the speaker's conversational intentions. She differentiates between three semantic levels: propositional, epistemic and speech-act. In our study, 79.5% of the instances of *weil* were interpreted on the propositional level, 14% on the epistemic level and 6.5% on the speech-act

[48] This analysis was carried out by Nataliia Vorona in her 2015 Uni Potsdam B.Sc. thesis.
[49] 2 Cf. Scheffler (2014).

level. A similar distribution can be found among the cross-tweet-uses. However, in these cases, the epistemic use is a little more frequent with 19%, because in many cases Twitter users were asked to justify their arguments (and then provide a 'because'-clause in a separate tweet). Table 1 compares the distribution of the semantic levels in causal relations with *weil* on Twitter (our work) with Volodina's (2010) corpus study of semantic levels in spoken language, considering all the main connectives (*weil, da, denn*, and *nämlich*).

	Tweets (*weil* only)		Spoken (*weil, da, denn, nämlich*)
	Number	Percent	
Propositional level	159	79.5%	34.7%
Epistemic level	28	14%	34.9%
Speech act level	13	6.5%	28%

Table 1. Distribution of semantic levels for causal relations in Twitter vs. speech.

Other Coherence Phenomena

Finally, we have also studied the use of URLs and hashtags in Twitter wrt. coherence: In particular, we have found that the discourse relation connecting URLs to the rest of the tweet/conversation is usually left implicit (in 80% of cases), and could be causal, elaborating, exemplifying, or evaluative.[50] Explicit linguistic devices such as connectives are rarely used to identify the discourse function of a URL. Interestingly, URLs can be used as satellites (providing additional information) as well as as nuclei (central to the discussion) when analyzed in an RST framework. In almost all cases, the reader must open the link to understand the meaning and function of the URL. This suggests that the specific properties of the medium Twitter lead writers to provide fewer explicit cues for discourse relations of certain utterances, contrary to what has been observed for standard written texts (Taboada 2006).

[50] The analysis of the discourse function of URLs was carried out by Laura Stelter in her 2015 Uni Potsdam B.Sc. thesis.

References

- Rehbein, I. 2014. *Using Twitter for linguistic purposes: Three case studies.* Talk at DGfS. [http://www.sfb632.uni-potsdam.de/~rehbein/slides/dgfs2014-webcorpora-rehbein.pdf]
- Scheffler, T. 2014. A German Twitter snapshot. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* Reykjavik, Iceland.
- Stede, M. 2002. DiMLex: A lexical approach to discourse markers. In: Lenci, A. & Di Tomaso, V. (eds.) *Exploring the Lexicon – Theory and Computation*. Alessandria: Edizioni dell'Orso.
- Stede, M. & Neumann, A. 2014. Potsdam Commentary Corpus 2.0: Annotation for discourse research, In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* Reykjavik, Iceland.
- Sweetser, E. E. 1990. *From Etymology to Pragmatics.* Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.
- Taboada, M. 2006. Discourse Markers as Signals (or Not) of Rhetorical Relations. *Journal of Pragmatics* 38 (4): 567–592.
- Volodina, A. 2010. Sweetsers Drei-Ebenen-Theorie: Theoretische Überlegungen vor dem Hintergrund einer korpuslinguistischen Studie über konditionale und kausale Relationen. [http://annavolodina.de/dokumente/Volodina_2010_ Sweetsers_Drei-Ebenen-Theorie.pdf]
- Wegener, H. 1999. Syntaxwandel und Degrammatikalisierung im heutigen Deutsch? Noch einmal zu weil-Verbzweit. *Deutsche Sprache* 27 (1): 3–26.

Merel C. J. Scholman[a], Ted J. M. Sanders[b], Pim W. Mak[b]

How do expectations based on contextual signals guide the processing of additive and causal relations?

[a]*Saarland University, Germany;* [b]*Utrecht University, the Netherlands*
merel_scholman@hotmail.com; t.j.m.sanders@uu.nl; w.m.mak@uu.nl

Introduction

Several studies have shown that causally related information is processed faster and represented better than non-causally related information (Black & Bern 1981, Kuperberg et al. 2011, Myers et al. 1987, Sanders & Noordman 2000). At the same time, discourse relational devices like connectives have a clear function as processing signals: they guide and speed up the processing of upcoming information. Connectives signaling non-causal relations seem to reverse readers' expectations of upcoming coherence relations, making the anticipation of unexpected relations stronger (Asr & Demberg 2015, Drenhaus et al. 2014, Koornneef & Sanders 2013, Xiang & Kuperberg 2015). In this contribution, we investigate whether such signals can influence the causal processing preference. More specifically, we examine whether cues other than connectives that precede the coherence relation can facilitate the processing of non-causal relations. It is investigated (1) whether a context sentence signaling an upcoming list relation facilitates the processing of a subsequent list relation, and (2) whether a list signal in the context sentence perturbs the processing of a subsequent causal relation. This will provide more insight into the relative strength of and interaction between contextual signals and general processing preferences.

Method

An eye tracking-while-reading experiment was conducted with a 2x2 design: context (list signal vs. no signal) and relation type (list relation vs. causal relation). Participants (n = 44) read 32 stories in Dutch that consisted of a context sentence and a pair of connected sentences, as in Example 1. In conditions A and C, the context sentence refers to multiple instances (e.g., 'a few', 'several'); in conditions B and D, the context did not signal an upcoming list. The coherence relation following the context sentence was either a *list* relation signaled by the marker 'ook' (*also)* (conditions A and B), or a *consequence-cause* relation signaled by the marker 'namelijk' *(namely),* which can be used in Dutch to signal causal relations (conditions C and D).

1 A ***[List signal – list relation]*** De student had vandaag een aantal financiële meevallers. Ze kreeg korting van een winkelier. Ze kreeg ook een gratis kop koffie in een café.
'The student had a few financial breaks today. She got a discount from a shopkeeper. She also got a free cup of coffee in a café.'

B ***[No list signal – list relation]*** De student was vandaag de stad in gegaan. Ze kreeg korting van een winkelier. Ze kreeg ook een gratis kop koffie in een café.
'The student went into the city today. She got a discount from a shopkeeper. She also got a free cup of coffee in a café.'

C ***[List signal – causal relation]*** De student had vandaag een aantal financiële meevallers. Ze kreeg korting van een winkelier. Ze had namelijk een kortingsbon mee uit een dagblad.
'The student had a few financial breaks today. She got a discount from a shopkeeper. She had [namely] brought a coupon from a magazine.'

D ***[No list signal – causal relation]*** De student was vandaag de stad in gegaan. Ze kreeg korting van een winkelier. Ze had namelijk een kortingsbon mee uit een dagblad.
'The student went into the city today. She got a discount from a shopkeeper. She had [namely] brought a coupon from a magazine.'

Results

The results show shorter regression path durations at the end of the final sentence in condition A compared to B (1019 ms vs. 1278 ms, $p < .001$), providing evidence that the list signal facilitated integration of the second argument of a list relation with previous content. Moreover, the overall mean for first gaze duration at 'namelijk' *(namely)* was longer in condition C than in D (236 ms vs. 214 ms, $p < .01$), indicating that the list signal immediately perturbed the processing of the causal marker.

Conclusion

The results suggest that a list signal facilitates the integration of the second argument of a list relation with the previous context, and that it hinders the processing of a causal marker. It is concluded that contextual signals can override general processing preferences, but more research is necessary to further investigate the strength and effect of a list signal on the processing of a subsequent coherence relations.

References

- Asr, F. T. & Demberg, V. 2015. Uniform Information Density at The Level of Discourse Relations: Negation Markers and Discourse Connective Omission. *IWCS 2015:* 118.
- Black, J. B. & Bern, H. 1981. Causal Coherence and Memory for Events in Narratives. *Journal of Verbal Learning and Verbal Behavior* 20: 267–275.
- Drenhaus, H., Demberg, V., Köhne, J. & Delogu, F. 2014. Incremental and Predictive Discourse Markers: ERP Studies on German and English. In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014).*
- Koornneef, A. W. & Sanders, T. J. M. 2013. Establishing Coherence Relations in Discourse: The Influence of Implicit Causality and Connectives on Pronoun Resolution. *Language and Cognitive Processes* 28: 1169–1206.
- Kuperberg, G. R., Paczynski, M. & Ditman, T. 2011. Establishing Causal Coherence Across Sentences: An ERP study. *Journal of Cognitive Neuroscience* 23(5): 1230–1246.

- Myers, J. L. Shinjo, M. & Duffy, S. A. 1987. Degree of Causal Relatedness and Memory. *Journal of Memory and Language* 26: 453–465.
- Sanders, T. J. M. & Noordman, L. G. M. 2000. The Role of Coherence Relations and Their Linguistic Markers in Text Processing. *Discourse Processes* 29: 37–60.
- Xiang, M.& Kuperberg, G. R. 2015. Reversing Expectations During Discourse Comprehension. *Language, Cognition and Neuroscience* 30: 648–672.

Istvan Szekrenyes

Automatic Prosodic Annotation for DRD Analysis

University of Debrecen, Hungary
xepenator@gmail.com

Introduction

The poster aims to demonstrate a new automatic annotation method (called *ProsoTool,* Szekrényes 2015) which can generate annotation labels representing the prosodic structure of a recorded dialogue. The algorithm is implemented as a Praat (Boersma & Weenink 2015) script and it requires only a two-level annotation of speaker change to identify turn-takes, overlapping speeches and the individual vocal ranges of the speakers. Unlike *ToBi* (Rosenberg 2010), the annotation-system is language-independent. The psychoacoustic model of tonal perception (Hart 1976) was used as theoretical background describing only the perceptually relevant features of prosodic events by smoothing and stylizing the original F0 curves. The annotation labels indicate the shape ("rise", "fall" etc.) and the relative location of prosodic movements based on their initial and final values and the individual vocal range of the speaker.

The results can be associated with the annotation of discourse structure including turn-taking, discourse markers and topic boundaries, as we already made them available in the *HuComTech Multimodal Corpus* (Hunyadi et al. 2012a). Prosodic features as possible non-verbal cues of communicative functions can be very important factor for the analysis of any discourse related phenomenon (Hunyadi et al. 2012b). Beyond the technical aspects of prosodic annotation, the poster focuses on the prosodic cues of different dialogue types (formal or informal) and topic structure: how they can contribute to the efficiency of automatic categorization and detection.

Methodology

The first, pre-processing step of the algorithm is the isolation of the selected speaker's voice based on the manual annotation of speaker change. The preliminary analysis of F0 data uses a dynamic extraction method (Frid & Ambrazaitis 2010) and aims to identify the individual vocal range of the speaker defining five levels of F0 values: L_2, L_1, M, H_1, H_2 (where L_2 is the lowest and H_2 is the highest level).

The further operations are performed on every speech segments annotated in the original audio file:

- F0 extraction using speaker-dependent parameters
- F0 smoothing and stylization: segmentation into prosodic events
- Classification and labelling of F0 movements.

In earlier versions, Merten's *Prosogram* (d'Alessandro & Mertens 2004) was used for stylizing F0 contour into syllable-size vectors. In the current version of *ProsoTool*, smoothing and stylization are performed by the built-in functions of the Praat program to exclude perceptually not relevant, momentary excursions of the actual measured F0 data resulting longer stretches of tonal movements. The classification is based the duration and the amplitude of the movements (following the parameters of the Tilt intonation system, Taylor 2000), which are compared to the individual vocal range of the speaker.

Evaluation

The ProsoTool script was tested using the audio recordings of annotated two-party dialogues from the *HuComTech Corpus*. In my previous work (Szekrényes 2014), three manual and an automatic annotation of the same speech segment were compared within the framework of a perceptual experiment. The poster aims at presenting the results of further experiments on the perception and computational analysis of speech prosody in relation to the pragmatic structure of different dialogue types.

References

- Boersma, P., Weenink, D. 2015. *Praat: doing phonetics by computer (Version 6.0.14)* [Computer program] [http://www.praat.org/]
- d'Alessandro, Ch. & Mertens, P. 2004. Prosogram: semiautomatic transcription of prosody based on a tonal perception model. In: *Proceedings of the 2nd International Conference of Speech Prosody,* 2004. 23–26.
- Frid, J. & Ambrazaitis, G. 2010. Automatic estimation of pitch range through distribution fitting. In: Schötcz, S & Ambrazaitis, G (eds.) *Proceedings from Fonetik* (Working Papers 54.). Lund University. 4–46.
- Hart, J.'t 1976. Psychoacoustic backgrounds of pitch contour stylization. In: *IPO – Annual Progress Report 11.* Eindhoven, The Netherlands. 11–19.
- Hunyadi, L., Földesi, A., Szekrényes, I., Kiss, H., Abuczki, Á. & Bódog, A. 2012a. Az ember-gép kommunikáció elméleti-technológiai modellje és nyelvtechnológiai vonatkozásai. In: Prószéky, G. & Váradi, T. (eds.) Általános Nyelvészeti Tanulmányok XXIV.: Nyelvtechnológiai kutatások. Budapest: Akadémiai Kiadó. 265–309.
- Hunyadi, L., Szekrényes, I. Borbély, A. & Kiss, H. 2012b. Annotation of spoken syntax in relation to prosody and multimodal pragmatics. In: *3rd International Conference on Cognitive Infocommunications (CogInfoCom 2012),* Kosice, Szlovakia, 2012. December 2–5, IEEE. 537–541.
- Rosenberg, A. 2010. AuToBI – A Tool for Automatic ToBI annotation. In: *11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010).* Curran Associates. 146–150.
- Szekrényes, I. 2015. ProsoTool, a method for automatic annotation of fundamental frequency. In: *6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom),* 19–21 Oct. 2015, Győr, IEEE. 291–296.
- Szekrényes, I. 2014. Annotation and interpretation of prosodic data in the HuComTech corpus for multimodal user interfaces. *Journal on Multimodal User Interfaces* 8 (2): 143–150.
- Taylor, P. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America* 107 (3): 1697–1714.

Mutsuko Tomokiyo

An annotation by speech act labels for dialogue discourse analysis in Japanese, French and English

LIG-GETALP, BP 53
mutsuko.tomokiyo@imag.fr

Purpose

We have made telephone dialogue annotations using labels based on speech act theory (Austin 1962, Searle 1979) for Japanese, French and English documents in order to analyze their discourse and contribute to semantic representation for Japanese-French-English machine translation (MT).

Rationale

The speech act theory[51] lies in analyzing meanings in a relation among linguistic conventions correlated with words or sentences in utterances. According to the theory, the speaker utters a sentence and performs a speech act to the hearer, and the speech act is associated with speaker's intention in a situation. Various illocutionary acts in terms of an intention can be supposed and be classified (Oishi 2006, Tomokiyo 2000). We suppose the chain of these labels implies dialogue discourse. So, we have experimented a classification of the illocutionary acts by annotating the collected corpora (Tomokiyo 1994a, b).

The used corpora and its annotation

The annotations are manually made by linguistic intuition of experts for following corpora, while referring to some guidelines based on the classifications of meanings by Austin (1962) and Searle (1979).

[51] The speech act theory has been long time discussed since J. L. Austin: B. Russell (1972), J-R. Searle (1975), J. Derrida (1990), E. Oishi (2006), etc.

Corpora	in Japanese[52]	in French[53]	in English[54]
Dialogue topics	enquiry about a conference	hotel reservation	enquiry about a conference
Number of dialogues	66	68	67
Number of turns	2929	3381	1947
Number of words	30333	35444	26334

Classified illocutionary acts

There exist no real commonly agreed-upon standards for establishing a set of speech acts labels. We think illocutionary acts consist of the intentions and communicative manners of interlocutors (Tomokiyo 2000, Weisser 2014). The annotation labels are assigned to surface expressions of each unit. Concerning annotation unit, words, a sentence, a turn, un utterance, and pause unit () are taken account of.

The set of speech acts labels consists of following 28 illocutionary act labels :

"Inform", "Offer", "Offer-follow-up", "Promise", "Yn-question", "Action-request", "Confirmation-question", "Do-you-understand-question", "Permission-request", "Wh-question", "Yes", "No", "Acknowledge", "Thanks", "Thanks-response", "Farewell", "Good-wishes", "Good-wishes-response", "Greet", "Apology", "Apology-response", "Alert", "Instruct", "Confirmation-question-to-self", "Invite", "Vocative", "Topic", "Expressive" (Tomokiyo 1994a).

E.g. (Japanese corpus)

Receptionist: はい (Hello) (Greet) 、京都国際交流センターです。
(Here is Kyoto international exchange center.) (Inform)
Participant: わたくし三浦ともうします。本日開催されますマシンエイドトランスレーションにおけるマルチメヂアの役割についてという講演をききたいんですけれど。
(My name is Miura. (Greet) I'd like to attend a lecture "About the role of multimedia on Machine-aided translation") (Inform)

(French corpus)

Client: Bonjour Madame. (Hello, Madame) (Greet) Je voudrais une chambre pour deux nuits. (I'd like to reserve a room for 2 nights) (Action-request)
Receptionist: Oui, (Acknowledge) pour quelle date? (for what day?)

[52] The corpus has been developed at ATR in Japan in 1996.
[53] The corpus has been developed at CLIPS-GEOD in France in 1999.
[54] The corpus was developed at ATR in Japan in 1996.

(Wh-question) (English corpus)
Receptionist: Good morning (Greet) Conference office (Inform), can I help you? (Offer)
Participant: Yes (Acknowledge), I'm trying to find my way to the International telecommunications and right now I'm at Kyoto station. (Inform) (Tomokiyo 1994a)

Modelization of dialogue discourse

In order to model structures of dialogues, we have also studied utterance connectors in the corpus, and have proposed a semantic representation for two adjacent utterances in each corpus (Tomokiyo 2000):

E.g. "donc" in French: P: Oui, puisque je viens en train. (Yes, because I'm coming by train.)
Q: Donc, c'est simple. (So, it's simple.)

```
[[sem [[reln *donc*]
      [type *conséquent*]                    (logico-semantic information)
      [niveau *plusieurs*]
      [obj [[P [[reln *Inform*]              (pragmatic information)
               [agen *locuteur*]
               [recp *interlocuteur*]
               [type *declaratif-positif*]
               [nature *stimulant*]
               [obj [[reln *venir*]          (syntactic-semantic information)
               [subj ?agen]
               [means [[reln *train*]] ]]] ]] ]
          [[Q [[reln *Inform*]
               [agen !recp]
               [recp !agen]
               [type *déclaratif-positif*]
               [nature *stimulé*]
               [obj [[reln "être simple"]
               [subj [[reln *ce*]]] ]] ]] ] ] ] ] ]
```

In the semantic representation for "donc (therefore)", *Inform* is one of speech act labels, and it represents that the utterance Q, which is labeled as "Inform" is connected with the utterance P, labeled as "Inform" in the relation (in type) of "conséquence".

Reflection and perspectives

Corpus annotation by linguistic experts is a time-consuming job. So, it's better and possible to semi-automate the annotation by preparing bigger corpus, adding some ontology to our labels and by making example-based search with some criteria (Leech & Weisser 2003, Tomokiyo 1994b, Weisser 2014), because our set of labels correspond to surface expressions or words employed in a situation.

The final goal of our work consists in improvement of results of source language analysis towards improvement of the quality of translation in Japanese-French MT. So, it's our subject to be taken up to make agglutination of speech act labels with a syntactic-semantic module.

References

- Austin, J. L. 1962. *How to Do Things With Words.* Oxford: Clarendon Press. (Also 1962 Quand dire, c'est faire. Edition du Seuil. Paris.)
- Leech, G. & Weisser, M. 2003. *Generic speech act annotation for task-oriented dialogues. Working paper.* University of Lancaster and Friedrich-Alexander-Universität Erlangen-Nürnberg. [http://www.lancaster.ac.uk/fass/doc_library/linguistics/leechg/leech_and_weisser_2003.pdf]
- Oishi, E. 2006. Austin's Speech act Theory and the Speech Situation. *Esercizi Filosofici* 1. [http://www2.units.it/eserfilo/art106/oishi106.pdf]
- Searle, J. R. 1979. *Expression and Meaning: Studies in Theory of Speech Act*. Cambridge: Cambridge University Press.
- Tomokiyo M. 1994a. *Natural Utterance Segmentation and Discourse Label Assignment. Proceedings of ICSLP94. Vol. 3.,* Yokohama.
- Tomokiyo M. 1994b. *Communicative Labels and their automatic labeling.* Rap. ATR n°TR-IT-0067.
- Tomokiyo M. 1996. *Pause units and Communicative Act units – towards Japanese Discourse Representation*. In: *Proceedings of NL-11,* Tokyo.
- Tomokiyo M. 2000. *Analyse discursive de dialogues oraux en français, japonais et anglais.* Lille: Septentrion.
- Weisser, Martin. 2014. Speech act annotation. In: Aijmer, K. & Rühlemann, C. (eds.) *Corpus Pragmatics*: a Handbook. Cambridge: CUP.

Ildikó Vaskó

Markers of mirativity in Hungarian

Eötvös Loránd University of Budapest
vasko.ildiko@btk.elte.hu

Mirativity is a grammatical category which tells the hearer that the information transmitted was not only new to the hearer at the time when the speaker herself received that information but also unexpected and therefore surprising, whether or not the speaker has first-hand experience with that which surprised her. It is the marking of *unexpected* information, information that somehow shocks or surprises the speaker. This category has been linked to evidentiality however, in recent years based on evidence from Lhasa Tibetan, the Athapaskan language Hare and some other languages. DeLancey (1997, 2001) argues that mirativity should not be subsumed under the category of evidentiality. His position seems to prevail today among typologists with a special interest in modality and evidentiality. There appears to be a connection between mirativity and speech act. Many (but not all) examples of miratives involve an exclamative speech act. Discourse items, particles which encode information about speakers' source of knowledge or expectations of knowledge, markers that encode speaker's expectations of other's knowledge, can signal the surprise element.

In languages that do not encode an attitude of surprise by means of inflection ('mirative mood'), expressions that are functionally equivalent to mirative affixes are often fully or partially grammaticalized items historically derived from verbs that encode concepts related to imagination or reflection/pondering. This is true of the utterance-initial expressions *képzeld* ('imagine', 2nd p.sg. imperative) in Hungarian. I am going to argue that the imperative *képzeld* has undergone a process of grammaticalization that has led to a lexical split between the imperative and a segmentally identical, desemanticized particle that encodes an attitude of wonder or surprise directed at the expressed truth-conditional content, a so-called mirative marker. Mirative statements starting with *képzeld* display several formal and functional properties that are incompatible with the rules pertaining to the lexical verb *elképzel*, among others a marker of emphasis

or intensification which frequently accompanies markers of mirativity in many languages. The mirative particle *képzeld* conveys information that the speaker believes the hearer to be unfamiliar with. The particle will not be used if the speaker believes the information to be already in the interlocutors' common ground, so that the only new information transmitted by the speech act would relate to the speaker's propositional attitude.

1 – Képzeld, megkaptad az állást!
Guess what, you've got the job!

However, if the information is only new to the speaker, the surprise element is conveyed by means of *nahát,* an interjection. Both *nahát* and *képzeld* are non-truth-conditional items, but *nahát* has other information-structural properties than *képzeld*.

2 – Nahát, megkaptad az állást!
Wow, you've got the job!

It is not unusual that discourse markers appear in pairs, enforcing, completing each other. *Képzeld* is often used together with the multifunctional discourse marker *csak (only):*

1' – Képzeld csak, megkaptad az állást!
Guess what, you've got the job!

In case of ex.2' however, the two particles do not belong together, the information the speaker expresses is unexpected for her, but it is already known for the addressee. *Csak* functions rather as an adversative context marker (Gyuris 2009).

2' – Nahát, csak megkaptad az állást!
Wow, you've got the job!

I am going to argue that while *képzeld* and *nahát* can be considered as mirative markers, *csak* does not assume genuine miratívity.

It often occurs that verbs of cognition, such as the Hungarian *képzel* ('imagine'), *gondol* ('think'), the English *imagine*, *think, mean* or the Norwegian *tenke* ('think') are used in various discourse related functions. I am going to analyze

two of these linguistic items *(képzel/tenke)* as particles with little or no verb-like properties, as items that are formally identical to imperative forms of the verbs, more precisely as non-truth-conditional utterance-initial particles (marginally also utterance-finally).

Tenke is a Norwegian polysemous lexical verb, one of whose meanings corresponds directly to the meaning of the Hungarian verb *képzel*. I am going to focus on how the imperative and the particle differ, in both languages in terms of syntactic form as well as semantic import and pragmatic interpretation.

A propositional attitude of surprise at some fact is not compatible with the issuing of an imperative, and you cannot be surprised at something you imagine, nor can you be surprised at something if you are not convinced of its existence. For these reasons alone it would be impossible to maintain that the Hungarian word *képzeld* is still the imperative of the verb meaning 'imagine' when it conveys surprise, and that the Norwegian word *tenk* is still the imperative of the verb meaning 'think' when it conveys surprise.

References

- Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford: Oxford University Press.
- DeLancey, Scott. 1997. Mirativity: The grammatical marking of unexpected information. *Linguistic Typology* 1: 33–52.
- DeLancey, Scott. 2001. The mirative and evidentiality. *Journal of Pragmatics* 33: 369–382.
- Fretheim, Thorstein. 2011. Mirativity markers in Norwegian. Paper presented at *CHRONOS 10,* Aston University, 18–20 April 2011.
- Fretheim, Thorstein and Vaskó, Ildikó 2011. *A contrastive analysis of mirative markers derived from verbs of propositional attitude in Hungarian and Norwegian*. Presentation at *10th ICSH Lund*.
- Gyuris, Beáta. 2009. A hangsúlyos csak diskurzuspartikula interpretációja. In: Márta Maleczki and Enikő Németh T. (eds.) *A mai magyar nyelv leírásának újabb módszerei 7*. SZTE. Szeged. 157–179.
- Zeevat, Henk. 2008. Only as a mirative particle. In: Joergen Villadsen and Henning Christiansen (eds.) *Constraints in Language Porcessing* (CSLP 2008) Roskilde University, Computer Science Research Report 122, 76–88.

Ben Verhoeven, Walter Daelemans

Discourse features for computational stylometry

CLiPS Research Center, Department of Linguistics
University of Antwerp, Antwerp, Belgium
ben.verhoeven@uantwerpen.be; walter.daelemans@uantwerpen.be

We present ongoing research on the extraction of discourse features for computational stylometry in Dutch. More specifically, we are investigating whether author profiling experiments (i.e. predicting characteristics of the author of a text, e.g. age and gender) might benefit from adding discourse characteristics as features to the document representation. Currently, document representations for author profiling are mostly limited to word-based features, sometimes utilizing syntactic information. The hypothesis for using discourse information is that groups of people with a common sociological or psychological factor (e.g. gender) might organize discourse in a similar way, e.g. by using similar discourse structures, similar connectives and similar ways of structuring text in space and time.

We began this research by investigating low-level approaches to discourse, beginning with the creation of lexicons of Dutch discourse adverbs and discourse connectives. In exploring the creation of such lexicons, we mined the Dutch Wiktionary[55] for words from the categories 'adverbs' and 'conjunctions'. The goal of the word lists is to categorize the words according to their discourse meaning or function. The frequencies with which these lexicon categories are used is then an approximation of the frequency of use of certain discourse structures.

The adverbs are categorized based on the adverb typology described in the ANS Dutch grammar (Haeseryn et al. 1997). We have formalized these types into 11 categories: place/direction, time, frequency, grade/intensity, quantification, manner, modality, negation, conjunction, and preposition.

[55] http://nl.wiktionary.org

The conjunctions are categorized according to the Penn Discourse Treebank tagset (The PDTB Research Group 2007: 26–39). We only used the two top-levels of the hierarchy (class level and type level) because the third level (subtype level) is the most ambiguous and can only be annotated in context. We thus annotate for the following 16 categories: TEMPORAL – Synchronous, Asynchronous; CONTINGENCY – Cause, Condition; COMPARISON – Contrast, Concession; EXPANSION – Conjunction, Instantiation, Restatement, Alternative, Exception, and List.

Words can belong to more than one category to take polysemy and ambiguity into account. The final lexicon contains 1256 unique words belonging to one or more of the 27 categories. The lexicon will be made publicly available. We propose this representation as a first approximation of discourse structure.

An association analysis was performed, which is very similar to a correlation analysis, but relates numerical variables (discourse category frequencies) to a categorical variable (gender). For this analysis, we used two Dutch corpora for which author gender is known: the reviews part of the CLiPS Stylometry Investigation Corpus (1298 texts, Verhoeven & Daelemans 2014) and a large corpus of Dutch blog posts (301,080 texts). A logistic regression model is trained on each corpus with the category frequencies as features for each text and gender as classes. The regression coefficients for each feature are indicative of the strength of association. We compute two-sided confidence intervals for confidence levels of 95% and 99%.

We compare the category frequencies over gender in both corpora and indicate where our analysis shows a significant association between category frequency and gender in Table 1. Only the features that were significant in at least one of the corpora are shown here.

		Reviews			Blogs		
		M	F		M	F	
Conjunctions	Concession	152.04	160.92		61.17	61.98	*
	Alternative	30.55	31.56		14.39	14.31	**
	Exception	0.00	0.00		0.0050	0.0035	*
	Comparison	154.91	162.93		61.41	62.21	*
	Condition	84.73	72.71	*	24.06	23.54	**
	Expansion	360.65	373.41		149.22	149.33	**
	Instantiation	11.46	10.04	*	2.776	2.712	
Adverbs	Place	745.18	764.88	*	296.49	294.60	
	Preposition	802.70	755.41	**	281.24	276.30	**
	Question	48.93	56.84		15.19	14.81	*
	Manner	576.90	567.70		210.53	211.55	**
	Frequency	25.54	32.87		8.75	8.97	*
	Negation	103.35	117.07		31.47	31.05	**

Table 1. Frequencies of categories (per 10,000 words) per gender for both corpora. Significance: * $p < 0.05$ and ** $p < 0.01$

We can conclude that there are indeed significant associations between some of the discourse categories and author gender, although some associations are quite weak. More categories are significant for the blog corpus, which is probably due to the difference in size. Over the different corpora, we observe that women tend to use less Condition and less Prepositional adverbs (or prepositions). The differences in frequencies between the corpora are related to the different genres.

In future research, we would like to expand our word lists with more conjunctions. We currently only have 80 conjunctions in our lexicon, while we believe more conjunctions to be around that are not in the list. Another interesting approach would be to extract adverbs and conjunctions from part-of-speech tagged corpora so that our word lists better reflect real language usage.

We have now paved the way for starting gender classification experiments using the category frequencies as features. We hope to show that these features can also add information to a word-based document representation by further improving the performance of those classifiers.

Bibliography

- Haeseryn, W., Romijn, K., Geerts, G., De Rooij, J. & Van den Toorn, M. 1997. *Algemene Nederlandse Spraakkunst: Band 1. Tweede, geheel herziene druk*. Groningen/Deurne: Martinus Nijhoff uitgevers/Wolters Plantyn.
- The PDTB Research Group 2007. *The Penn Discourse Treebank 2.0 annotation manual.* IRCS Technical Reports Series. Pennsylvania, USA: ScholarlyCommons.
- Verhoeven, B. & Daelemans, W. 2014. CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC),* Reykjavik, Iceland, ELRA, 2014.

Bonnie Webber[a], **Rashmi Prasad**[b], **Alan Lee**[c], **Aravind Joshi**[c]

Discourse Annotation of Conjoined VPs

[a]*University of Edinburgh,* [b]*University of Wisconsin-Milwaukee,*
[c]*University of Pennsylvania*
bonnie.webber@ed.ac.uk; shaloo0105@gmail.com;
alanlee@lexicontree.org; joshi@seas.upenn.edu

As frequently noted, discourse relations can hold within a sentence as well within larger units of text. Interest has recently grown in the former *(intra-sentential)* discourse relations (Joty et al. 2015) – *e.g.* to support Statistical Machine Translation (Guzman et al.2014). We have therefore started to expand the annotation of intra-sentential discourse relations in the Penn Discourse TreeBank (Prasad et al. 2008, 2014).

According to English grammar (Huddleston and Pullum, 2002), conjoined clauses, VPs and verbs can have senses other than simply Conjunction (*and*), Disjunction *(or),* and Contrast *(but)*. For example, *X* and **Y** may convey:

1. **Consequence** (*X* and therefore **Y**), as in "Scopes was *convicted* and **fined $100** …" [wsj_0946]
2. **Temporal Sequence (***X* and then **Y**), as in "Tripoli says Rome *kidnapped 5,000 Libyans* and **deported them as forced labor.**" [wsj_0990]
3. **Condition** (if *X* then **Y**), as in "*Give television a chance to cover live any breaking of the law*, and **no second invitation will be required.**" [wsj-0290]
4. **Concession** (despite *X,* **Y**), as in "Blacks and Hispanics currently *make up 38% of the city's population* and **hold only 25% of the seats on the council.**" [wsj_1137]
5. **Temporal Inclusion** (*X* while **Y**), as in "…the government can *ensure the same flow of resources* and **reduce the current deficit.** [wsj_1131]

Since these constructions are common, we took this to justify expanding the Penn Discourse TreeBank to include discourse relations associated with conjoined VPs. Here we briefly describe (1) the identification of tokens to annotate;

(2) the senses used in annotation; (3) their frequency distribution, along with some examples of common senses; and (4) what still needs to be done.

Identification of tokens for annotation

The conjoined VPs come from the Penn TreeBank (PTB), identified through a search for sister VPs separated (optionally) by a conjunction or punctuation, and an optional adverbial. When found, the right-hand VP was pre-annotated as **Arg2** of a potential discourse relation, and the conjunction (if any) was pre-annotated as its explicit connective. The two annotators could then exclude tokens or adjust their **Arg2** span, as well as annotate their **Arg1** span and sense. Pre-annotated tokens were divided into 25 batches, with review and adjudication of each batch as soon as it was complete. The authors and annotators discussed the *hard cases* from each batch, keeping track of decisions, so as to maintain the same conventions in annotating further batches. Here we briefly describe the exclusions and span adjustments, since it is relevant to inducing an automated sense classifier over the data.

Annotators were told to reject pre-annotated VP conjuncts lacking a verb, such as:

6. The bonds are rated double-A by Moody's and **double-A-minus by S&P.** [wsj_1312]
   There were about 40 such tokens, and this decision could be reversed. Tokens fit for annotation might have their spans (projections of VP nodes) adjusted to exclude material that contributes to the sense of both conjuncts. This means that spans in the corpus may not be the projections of VP nodes in the Penn TreeBank, as in:
7. UAL … reversed course and **plummeted in off-exchange trading after the 5:00 p.m. EDT announcement.** [wsj_1305]
   Here the right-hand conjunct was changed to "**plummeted** in off-exchange trading …" since the span starting "in off-exchange trading" also holds of the left-hand conjunct.
   Finally, we decided to retain only those attributions within VP conjuncts that are involved in the semantics of the relation (as in Ex. 8, where *declaring* something a pesticide is intrinsic to the Purpose of pulling that thing from the marketplace), while excluding *said and added*, which don't seem to contribute to the Contrast relation taken to hold in Ex. 9.

8. Give the EPA more flexibility to *declare a pesticide an imminent hazard* and **pull it from the marketplace** [wsj_0964]
9. The company, based in San Francisco, said *it had to shut down a crude-oil pipeline in the Bay area to check for leaks* but added that **its refinery in nearby Richmond, Calif., was undamaged.** [wsj_1884]

Senses used in annotation

We have extended and simplified the sense hierarchy used in PDTB 2.0 annotation.

It retains the same four Level-1 senses, but Level-3 senses now only encode the direction of asymmetric relations like **Condition,** with its two Level-3 senses, **Arg1-as-cond** and **Arg2-as-cond.** New Level-2 senses include **Negative-condition, Purpose,** and **Manner,** while Level-3 senses such as **Hypothetical-conditional** have been removed, as they were rare and posed difficulties for annotators. For the Level-1 senses, the most common specific senses used in annotating conjoined VPs were:

- **Expansion:** Conjunction, Disjunction, Substitution.Arg2-as-subst, Manner.Arg2-as-manner, Level-of-detail.Arg2-as-detail
- **Contingency:** Purpose.Arg2-as-goal, Cause.Result
- **Temporal:** Asynchronous.Precedence
- **Comparison:** Contrast, Concession.Arg2-as-denier

Distribution of annotated relations

The 4583 tokens annotated to date include 4138 explicitly conjoined VPs: 3325 VPs conjoined with *and*, 458 VPs conjoined with *but*, 244 VPs conjoined with *or,* and 111 conjoined with less frequent connectives such as *either…or* (26 tokens), *rather than* (25 tokens), *instead of* (11), and *yet* (8). There are also 401 tokens of VPs conjoined by punctuation (a comma, a semi-colon or a dash), and 44 tokens of VPs conjoined by an explicit conjunction or punctuation, but whose sense seems associated with some other element, which we annotated as *AltLex* (Prasad et al. 2010).

As in annotating the PDTB 2.0, annotators were allowed to record that they inferred more than one sense as holding concurrently. Of the 4138 explicit tokens, 3136 were taken to convey a single sense, while 1002 were taken to con-

vey two or three senses. Of the 401 punctuation-conjoined VPs, only two were annotated with more than one sense. Though relations with *and* were labelled with senses other than **Conjunction** – for example, **Purpose** (Ex. 10) – tokens of *and* in which multiple senses were inferred involved **Conjunction,** plus another sense: **Result** (Ex. 1), **Precedence** (Ex. 2), **Arg2-as-subst** (Ex. 11), **Arg2-as-detail** (Ex. 12) and **Arg1-as-manner**+**Arg2-as-result** (Ex. 13).

10. These "active suspension systems" *electronically sense road conditions* and **adjust a car's ride** [wsj 0956]
11. "We've got to *get out of the Detroit mentality* and **be part of the world mentality,**" declares Charles M. Jordan, … [wsj_0956]
12. Last Monday, Qintex Australia *announced a restructuring plan* and **said it would sell off assets.** [wsj_0979]}
13. Punching away, we *raised what I still think were all the right issues* and **landed more than one hard blow,** … [wsj_0937]

With *but*, the most common sense pairings were **Contrast+Arg2-as-subst** (Ex. 14) and **Precedence+Concession.Arg2-as-denier** (Ex. 15).

14. The carnage among takeover stocks Friday *doesn't mean the end of mega-mergers* but **simply marks the start of a less ambitious game,** … [wsj_2443]
15. Consider Spendthrift Farm, a prominent Lexington horse farm that *went public in 1983* but **hit hard times and filed for bankruptcy-court protection last year.** [wsj_1174]

No tokens of *or* were annotated with more than one sense, nor were any of the less frequent conjunctions.

Inter-annotator agreement (IAA) on sense annotation (full agreement on one or more senses) was 74%. Partial agreement on at least one sense was 74.3%. IAA on both senses and argument spans was 69.8%. Partial IAA on at least one sense and span was 70.1%. Of the 658 sense disagreements, the most common involved **Contrast** and **Concession.Arg2-as-denier** (127/658 =19.3%). Of the 338 tokens labeled **Contrast** by at least one annotator, the annotators disagreed 37.6% of the time. But most of the disagreements between annotators involved only one of them reporting an additional inferred sense: For tokens annotated by both of them (explicit or inferred) the average disagreement was 13%.

Still to do before corpus release

Although all the conjoined VPs in the Penn TreeBank have been annotated, there are some very similar spans that have been parsed as conjoined S-nodes, with a null subject in the right-hand conjunct co-indexed with the subject of the left-hand conjunct. This is the case in Ex. 16:

16. He *joined the firm in 1963* and **bought it from the owners the next year.** [wsj_0305]
17. Nissan handled the die-hards in a typically Japanese fashion: They *weren't fired* but **instead "were neglected,"** says Kouji Hori, … [wsj_0286]

We decided to include them in the corpus as an automated parser would find it hard to distinguish them from conjoined VPs. Such tokens are currently being annotated and adjudicated, and will then be folded into the corpus of conjoined VPs. Also folded into the corpus will be tokens from the PDTB in which a discourse adverbial such as *instead* has been annotated in the right-hand conjunct, as in Ex. 17.

We are aiming to release the corpus in May 2016, to anyone with access to the Penn TreeBank (https://catalog.ldc.upenn.edu/LDC99T42). We believe that this is the first large corpus of its kind, and thus will be of interest to the community. While some conjoined clauses are annotated in the RST-corpus (Carlson et al 2003) but the sense of those conjoined by *(and)* is simply **List.** While tokens of conjoined clauses, verbs and VPs have been annotated with a wider range of senses by Kim and Di Eugenio (2006) and Subba and Di Eugenio (2009) within a corpus of home repair instruction manuals, the corpus is small compared with the current one, though worth examining further because of its focus on a different genre. Our hope is that this extensive new corpus will stimulate efforts to automatically annotate similar constructions in English and other languages.

References

- Carlson, L., Marcu, D. & Okurowski, M. E. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: van Kuppevelt, J. & Smith, R. (eds) *Current Directions in Discourse and Dialogue*. New York: Kluwer.

- Guzman, F., Joty, Sh., Marquez, L. & Nakov, P. 2014. Using discourse structure improves machine translation evaluation. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, June 2014. 687–698.
- Huddleston, R. & Pullum, G. 2002. *The Cambridge Grammar of the English Language.* Cambridge: Cambridge University Press.
- Joty, Sh., Carenini, G. & Ng, R. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics* 41: 385–435.
- Kim, S. N. & Di Eugenio, B. 2006. Coding scheme manual for instructional corpus: Identifying segments, relations and minimal units. Computer Science Department, University of Illinois at Chicago, 13 p.
- Prasad, R., Joshi, A. & Webber, B. 2010. Realization of discourse relations by other means: Alternative lexicalizations. In: *Proceedings, International Conf. on Computational Linguistics (COLING),* 2010.
- Prasad, R., Webber, B. & Joshi, A. 2014. Reflections on the penn discourse treebank, comparable corpora and complementary annotation. *Computational Linguistics* 40: 921–950.
- Subba, R. & Di Eugenio, B. 2009. An effective discourse parser that uses rich linguistic information. In: *Proceedings, Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado, June 2009. 566–574.

Šárka Zikánová[a], Liesbeth Degand[b], Péter Furkó[c], Sandrine Zufferey[d], Ágnes Abuczki[e]

Semantic weakening of discourse structuring devices

[a]*Charles University in Prague;* [b]*Université catholique de Louvain;*
[c]*Károli Gáspár University of the Reformed Church in Hungary;*
[d]*University of Bern, Switzerland;*
[e]*MTA-DE Research Group* for Theoretical Linguistics
zikanova@ufal.mff.cuni.cz; liesbeth.degand@uclouvain.be;
furko.peter@gmail.com; sandrine.zufferey@rom.unibe.ch;
abuczki.agnes@gmail.com

Semantic weakening is a type of semantic change whereby lexical elements shift from a "stronger" to a "weaker" or "less precise" meaning, also applicable to a wider context of use.[56] Typical lexical examples are the semantic weakening *astonish/astound,* from the intense meaning of 'being struck by thunder' to todays "weaker" meaning of 'being (highly) surprised'. This linguistic notion also seems to apply to discourse structuring devices and can be observed in many languages. Here, the shift seems to be mainly one from a semantically specific (cause, time, contrast, etc.) to a less specific or "underspecified" meaning:

(1) Czech: ***Tak*** *jak se máte?*
*[**So,** how are you?]*
Shift from the causal meaning to an underspecified relation

(2) French: *Le problème est réel.* ***Mais,*** *je voudrais attirer votre attention sur l'argument suivant.*
[The problem is real. But, I would like to draw your attention to the following argument.]

[56] This work has been supported in part by the LINDAT/CLARIN project No. LM2015071 of the MEYS CR.

Shift from (semantic) contrast to topic shift/introduction (sequential use)

(3) Hungarian: ***Akkor*** *hogy is fogjuk ezt csinálni?*
[***So (then),*** *how are we going to do it?*]
Shift from temporal through additive to underspecified meaning

Spooren (1997) talks about underspecified discourse relations in case "the semantics of the connective that is used to indicate the link does not fully match the semantics of the relation that is intended by the speaker or writer." (Spooren 1997: 150), e.g. the use of a temporal connective to express a causal relation, or the use of an additive connective (typically *and*) to mark a contrastive relation.

The phenomenon we would like to study here is a little different in that a semantically specific marker is used to express a less specific meaning. While this process has been studied extensively from a diachronic perspective – often in the theoretical context discussing the processes of grammaticalization and/or pragmaticalization (see e.g. Degand & Evers-Vermeul 2015 for an overview), we would like to focus here on a synchronic and cross-linguistic perspective.

The research questions we would like to address are:

(1) Are there some discourse structuring devices that are more prone than others to be used with both their semantically specified and semantically underspecified meanings? Cross-linguistic candidates seem to be the "equivalents" of *but* and *so,* but their might be others.

(2) How can this shift from specification to underspecification be described? In line with diachronic research, we suggest that the ideational (propositional, content, semantic) meaning shifts to "non-semantic" meanings, such as the interactional and textual domains.

In our contribution, we want to open the discussion on two general types of semantic weakening:

(i) Domain shift (e.g. shift from the ideational domain to the sequential (or discourse-structuring) one (Crible in press, Crible, Degand & Simon 2016, Cuenca 2013).
(ii) Shift of relational functions (e.g. shift from the additive function to opposition, within the same domain; e.g. the examples described by Spooren 1997).

Domain shift is grounded in the proposal that language may be used at different levels, domains or planes of discourse (Halliday 1994, Sweetser 1990, Schiffrin 1987). Given the polyfunctional nature of discourse structuring devices, we expect that these markers can be used in different domains and that this shift from one domain to another impacts on the meanings expressed (and therefore on the way they will be annotated in corpora).

Functional/Relational shift is also closely linked to the polysemy of DRDs, but the shift occurs within one domain, typically the semantic/ideational domain. Thus temporal markers, typically expressing temporal relations, can be used to express causal or contrastive meanings.

During the pilot analysis of the language data in Czech, French and Hungarian several common types of shifts were observed in both groups of shifts (domain shift, shift of relational functions):

Topic shift (domain shift)

(4) Hungarian: És *mikor megyünk el a kiállításra?*
*[**And** when are we going to see the exhibition?]*

(5) French: ***Alors,** qu'est-ce qu'on fait aujourd'hui?*
*[**Then,** what do we do today?]*
Authentic example:
*et c'est un des aspects qu'on discutera dans la conclusion /// **alors** euh le phonostyle c'est un un objet d'étude particulier*
*[and that's one of the aspects that we will discuss in the conclusion /// **so/then** uh the phonostyle is a particular object of study]*

(6) Czech: ***A** co ve škole dneska?*
*[**And** what about your school lessons today?]*

Emphasis (domain shift)

(8) French: *tiens encore Jean d'Ormesson // **mais** on entend /// Jean d'Ormesson // à chaque automne ///*
*[hey again Jean d'Ormesson // **but** we hear /// Jean d'Ormesson // every autumn ///]*

(9) Hungarian: *Engem **pedig** nem érdekel!*
*[I'm **however** not interested in the least!]*

(10) Czech: ***Tak,*** *a teď jsi to rozbil!*
*[**So,** and now you have broken it!]*

Relational shift: from conjunction to opposition

(11) Czech: *Pořád se mluví o malém pohraničním styku **a** nikdo vlastně neví, co to je.*
*[People talk about small contact around the borderline **and** nobody knows what it is.]*

(12) French: *Il avait promis de venir **et** il n'est pas venu.*
*[He had promised that he would come **and** he did not come.]*

(13) Hungarian: *Megígérte, hogy eljön* és *nem jött!*
*[He had promised that he would come **and** he did not come.]*

Relational shift: from opposition to conjunction

(14) Czech: *Podle našich informací najdete takové školy v Praze 2, v ulici Moravské… Bližší informace **však*** získáte z brožury Komplexní informace o pražských učilištích…
*[According to our information you can find this type of schools in Prague 2, the Moravská Street… **But** you can obtain more detailed information in the booklet Complex information about Prague training schools.]*

(15) French: *(…) comme vous allez voir on ne travaille pas // euh sur des valeurs brutes de f zéro /// **mais** on travaille sur une représentation stylisée de la fréquence fondamentale ///*
*[as you will see we do not work on uh raw values of F zero /// **but** we work on a stylized representation of the fundamental frequency ///]*

Conclusion and open questions

According to the pilot data analysis, semantic weakening or underspecification does occur cross-linguistically in typologically diverse languages. Work is ongoing to quantify the phenomenon at hand, also in other languages. Identifying and describing these semantic shifts is important in view of our common endeavor on discourse annotation. To what extent do these phenomena challenge existing annotation schemes? Which markers give rise to most ambiguity? What is the role of the marker vs. the discourse relation and / or discourse domain when describing our data?

References

- Crible, L. in press. Towards an Operational Category of Discourse Markers: A Definition and its Model. In: Fedriani, C. & Sansó, A. (eds.) *Discourse Markers, Pragmatic Markers and Modal Particles: New Perspectives*. Amsterdam: John Benjamins.
- Crible, L., Degand, L. & Simon, A. C. 2016. *Interdependence of annotation levels in a functional taxonomy for discourse markers in spoken corpora*. Oral presentation at the 2nd TextLink Action Conference, 11–13 April 2016, Budapest.
- Cuenca, M. J. 2013. The Fuzzy Boundaries between Discourse Marking and Modal Marking. In: Degand, L., Cornillie, B. & Pietrandrea, P. (eds.) *Discourse Markers and Modal Particles. Categorization and Description.* Pragmatics & Beyond New Series. Amsterdam: John Benjamins. 181–216.
- Degand, L. & Evers-Vermeul, J. 2015. Grammaticalization or pragmaticalization of discourse markers? More than a terminological issue. *Journal of Historical Pragmatics* 16 (1): 59–85.
- Halliday, M. A. K. 1994. *An Introduction to Functional Grammar.* Second ed. London: Edward Arnold.
- Schiffrin, D. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.
- Spooren, W. 1997. "The Processing of Underspecified Coherence Relations." *Discourse Processes* 24 (1): 149–68.
- Sweetser, E. 1990. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure.* Cambridge: Cambridge University Press.

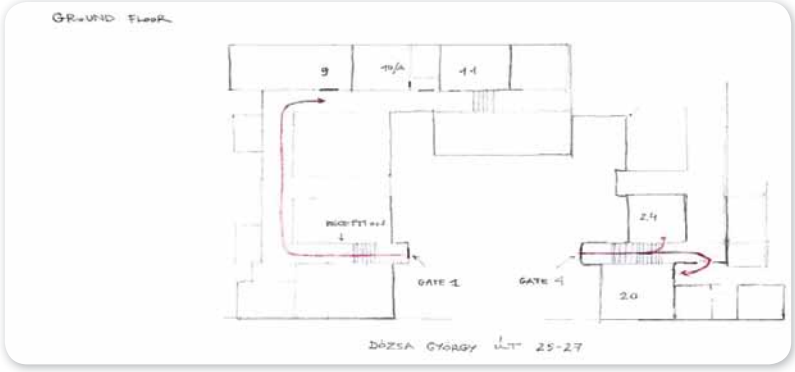
APPENDIX
LAYOUT OF THE MAIN BUILDING, GETTING AROUND BUDAPEST

Venues

The venues are **25 Dózsa György Way, Budapest, H-1146 (Main Building)** and **16 Szabó József Street, Budapest, H-1146 (Synod Building)**

Please note that if you enter the above addresses directly into googlemaps, you might end up with the wrong address, so please use the map below as a point of reference.

Here's a sketch of the layout of the main building. Room 10/A will be the cloak room. Please note that two individual rooms (9 and 11) will be used for lunches and coffee breaks on most of the days, so if one of the rooms is crowded, please try the other one. Room 215 is on the second floor, its location is two floors above room 9. Gate is "KAPU" in Hungarian.



Getting to the Synod from the main building

Sessions marked as SYNOD (both plenaries and some of the oral presentations) will be held in an off-site building of the Synod of the Hungarian Reformed Church (Szabó József utca 16). Turn left as you leave the main building (Dózsa György út 25.) and head south-east towards Abonyi utca. Turn left onto Abonyi utca and walk about 400ms. When you reach Szabó József utca turn right, and you'll find the entrance of the building at number 16. The auditorium is on the first floor.

Accommodation

Textlink is partnering with the **Lion's Garden Hotel and Danubius Hotel Arena** to provide participants with accommodation options (both are 4-star hotels, providing accommodation for 70 / 80€ per night for single / double use, respectively):

- **Lion's Garden Hotel:** http://www.lions-garden-hotel-budapest.com/. Participants wanting to book this hotel should complete the reservation form they receive via email.
- **Danubius Hotel Arena:** http://www.danubiushotels.hu/szallodak-budapest/danubius-hotel-arena Participants wanting to book this hotel should follow the link they receive via email.
- There are other options available on booking.com.

Public transport

Budapest's public transportation systems are operated by the company BKK (website: http://www.bkk.hu/en/main-page/news/). Buses, trams, trolleybuses and metros (4 lines) frequently run from 4.30 a.m. until 11.0 p.m. You can easily get anywhere you like using public transport. It is a good idea to **buy a block of 10 tickets (3000 HUF)** from a machine. You can find these BKK machines at around all terminals and stops (including the bus stop at Ferihegy/Liszt Ferenc airport). A single ticket bought from a machine costs 350 HUF, while a single ticket bought onboard from the driver (only available on buses!) costs 450 HUF.

A Budapest 24-hour travelcard (with unlimited BKK trips) costs 1650 HUF. Be sure to **validate your ticket** using the orange or red ticket-**punching machines** as controllers may ask to see your ticket, and will fine you for having an invalid one. Some ticket-punching machines on buses and streetcars are manual. Be sure to insert your ticket into the top slot and pull the punching mechanism toward you.

Taxis

(REMINDER: if you are being reimbursed through COST for participation in the meeting, the use of taxis is strictly limited to between the hours of 22.00 and 07.00).

Phone numbers of some taxi companies:

- 6x6 Taxi: +36-1-666-6666
- Barát Taxi: +36-70-773-2000
- Buda Taxi: +36-2-333-333
- Budapest Taxi: +36-4-333-333
- City Taxi: +36-2-111-111

The average taxi fare is composed of 3 parts: basic fee (450 HUF), er kilometre charge (280 HUF/km) and waiting fee (70 HUF/min). A taxi from the airport to the centre of Budapest will cost something between 5.000 and 10.000 HUF.

Getting to the hotel (**Lion's Garden Hotel,** Budapest, Cházár András u. 4, 1146: http://www.lions-garden-hotel-budapest.com/) **from the airport** (http://www.bud.hu/english):

- Getting to the hotel using **public transport:** You can buy a bus ticket from a machine at the airport for 350 HUF or from the driver onboard for 450 HUF. However, since you need to transfer from bus to metro several times, it is useful to buy a block of 10 tickets (3000 HUF) from the machine. From 4:00 a.m. to 11:00 p.m. the public airport **bus,** number **200E** commutes between Terminal 2 and Kőbánya-Kispest metro terminal (blue line, M3), about 25 minutes away. Get off the airport bus 200E at **Kőbánya-Kispest** (metro station), and **take metro line 3** (blue line, M3) towards Újpest Központ, and get off at the **Ferenciek tere stop (9 stops, 15 minutes).** From Ferenciek tere **take bus number 5, 7, 110, 112 or 907. Get off the bus at the Cházár András utca stop.** The hotel is located at the beginning of the

street, right across the 100-year-old Cathedral.

- A **taxi** from the airport to the hotel (and in general to the centre of Budapest) will cost something between 8.000 and 10.000 HUF (about 30 EUR). You can use one of the recommended taxi companies (or basically, any taxi company with a logo on it), and you may also contact the hotel for assistance: info@lions-garden.com.

Getting to Danubius Hotel Arena[****] Ifjúság útja 1–3., 1148 Budapest, Hungary **from the airport** (http://www.bud.hu/english):

From 4:00 a.m. to 11:00 p.m. you can take the public airport bus number 200E which commutes leaves from Terminal 2. Get off the airport bus 200E at Felsőcsatári út (8 stops, 16 minutes). From Felsőcsatári út take bus number 95 towards Puskás Ferenc Stadion. Get off at the final destination, Puskás Ferenc Stadion (20 stops, 31 minutes). You will see the hotel near the bus stop at Ifjúság útja 1.

Getting to the TextLink conference venue (Budapest, Dózsa György út 25, 1146) **from the hotel** (Lion's Garden, Budapest, Cházár András u. 4, 1146):

The conference venue is within walking distance from the hotel (500 metres). As you exit Lion's Garden, turn left on Cházár András utca. At the first crossing, turn left onto Abonyi utca. Then at the next crossing, turn right onto Dózsa György út. The conference venue will be on your right (Dózsa György út 25).

From Danubius Hotel Arena[****] **Ifjúság útja 1–3., 1148 Budapest to the conference venue, Károli University: 1146 Budapest, Dózsa György út 25–27:**

As you exit the hotel, turn right on Ifjúság útja and walk 300 meters to the nearest bus stop called Puskás Ferenc Stadion (near a football stadium). Take bus number 75 towards Jászai Mari tér, and get off at Ötvenhatosok tere (5 stops, 9 minutes). From Ötvenhatosok tere you can easily find the university/conference venue in 200 meters on Dózsa György Street (Dózsa György út 25–27) on your left.

Alternatively, you can walk straight down Dózsa György Street, which takes you to the conference venue, it is, approximately, a 20-minute walk.

Directions to the Építészpince Restaurant (venue of the conference dinner)

As you leave the main building turn right and head north-west towards "Ajtósi Dürer sor". Turn left onto "Dembinszky utca" and you will find the bus stop for bus No 74 (it is an electronic red bus, called "troli" [trolley] in Hungarian). Take the bus in the direction of "Astoria M" and go 6 stops. Get off at "Károly körút" (AKA Astoria M) and walk east on "Dohány utca". Turn right onto "Síp utca" then right again onto "Rákóczi út" and find "Puskin utca". Walk south on "Puskin utca" and continue straight on "Pollack Mihály tér", which, in turn, goes straight into "Ötpacsirta utca". The address of the restaurant is 2 Ötpacsirta Street (Ötpacsirta utca 2.).

Directions to the Cruise Ship Dinner

From Lion's Garden Hotel Budapest, Cházár András u. 4, 1146 **to the boat cruise, Jászai Mari tér, hajóállomás (port)**

As you exit the hotel, turn right and walk 450 meters on Cházár András street. From the bus stop Ötvenhatosok tere take bus number 75 towards Jászai Mari tér (14 stops, 21 minutes). Get off the bus at its last stop. Walk to the port (=hajóállomás in Hungarian) (max. 500 meters).

From Danubius Hotel Arena**** Ifjúság útja 1–3., 1148 Budapest **to to the boat cruise, Jászai Mari tér, hajóállomás (port)**

Walk 350 meters to the nearest metro stop, Puskás Ferenc Stadion (near a football stadium). Take metro line M2 towards Déli pályaudvar and get off at Kossuth Lajos tér (5 stops, 10 minutes). You will see the House of Parliament on a square on your left. Pass the Parliament and walk through Kossuth tér which continues in Balassi Bálint utca (800 meters). At the end of Balassi Bálint utca, you will get to Jászai Mari tér and the port (=hajóállomás in Hungarian) will be on your left.

Directions to the venue of the concert

The concert hall is a half-hour walk from the conference venue. One of the organizers, Anna Nagy, will be guiding participants from the main building to the concert venue. If you decide to use public transport to get there from one of our partner hotels, here are the directions:

From Lion's Garden Hotel Budapest, Cházár András u. 4, 1146 **to the concert venue, Liszt Academy of Music,** Concert Centre Budapest, Liszt Ferenc tér 8:

As you exit the hotel, turn right and walk 450 meters on Cházár András street. From the bus stop Ötvenhatosok tere take bus number 74 towards Károly körút (Astoria M). Get off the bus at Wesselényi utca (5 stops, 9 minutes). Turn right and 600 walk meters on Erzsébet körút. After 600 meters, turn left and you will see the concert venue, Liszt Academy of Music, Liszt Ferenc tér 8.

From Danubius Hotel Arena**** Ifjúság útja 1-3., 1148 Budapest **to the concert venue, Liszt Academy of Music,** Concert Centre Budapest, Liszt Ferenc tér 8:

Walk 350 meters to the nearest metro stop, Puskás Ferenc Stadion (near a football stadium). Take metro line M2 towards Déli pályaudvar and get off at Blaha Lujza tér (2 stops, 5 minutes). From Blaha Lujza tér, you can turn left and walk straight ahead for about 900 meters to the concert venue, Liszt Ferenc tér 8. If you don't want to walk, you can take 2 stops on tram 4 or 6 towards Széll Kálmán tér (just 2 minutes). Get off the tram at Király utca. You will find the concert venue after walking 150 meters.

Tourist information about Budapest:
http://www.budapest.com/
http://visitbudapest.travel/budapest-info/

Further useful information:

As for **currency,** although Hungary is part of the European Union, it does not use the Euro as its currency. The Hungarian currency is the **Forint** (Ft, **HUF**). 1 EUR = 310 HUF.

Credit cards are commonly but not universally accepted so it is wise to obtain some Forints (about 10.000 HUF) for your stay even if you expect to use a credit card or debit card for as much as you can. All hotels and most restaurants accept credit cards. However, in some corner shops and kiosks you need to pay in cash for amounts lower than 1000 HUF (approx. 3 EUR).

ATMs are easy to find in Budapest, and there are many options for exchanging cash. Most often, currency exchange kiosks located in tourist areas or shopping malls offer the best exchange rates. Currency exchange is available at the airport at a significant surcharge (10–15%). Currency exchange is also available at banks at a smaller surcharge. (Many international banks have branches in

Hungary.) ATMs dispense Hungarian currency at your bank's daily exchange rate; however, you may be charged a foreign fee on top of the service fees.

Tipping is very much a part of the culture in Hungary, and most people will routinely tip waiters and taxi drivers (10–15%).

As for drinking **water,** tap water is safe to drink in Hungary. If you want to buy **mineral water,** it is useful to know that the meaning of the **colours** of the bottle caps is different in Hungary than in most parts of Europe. Pink cup means mineral water without gas, blue means water with gas, and green means mild sparkling water, just a little bit bubbly.