Cross-Linguistic Discourse Annotation: applications and perspectives. TextLink2018 – Final Action Conference

Lydia-Mai Ho-Dac, Philippe Muller

cost
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

Université de Toulouse

Cross-Linguistic Discourse Annotation Applications & Perspectives

L.-M. Ho-Dac & P. Muller

Final Action Conference
University of Toulouse, March 19-21 2018

CLLE ERSS
IRIT
CNRS · INPT · UPS · UT1 · UTM

Program Committee

Maria Cuenca	Universitat Politècnica de València
Liesbeth Degand	Université catholique de Louvain
Lydia-Mai Ho-Dac	University of Toulouse
Amália Mendes	Centro de Linguística da Universidade de Lisboa
Jiri Mirovsky	Charles University in Prague
Philippe Muller	IRIT, Toulouse University
Hannah Rohde	The University of Edinburgh
Ted Sanders	Utrecht University
Manfred Stede	Univ Potsdam
Jacqueline Visconti	University of Genoa
Bonnie Webber	The University of Edinburgh
Deniz Zeyrek	Middle East Technical University
Sandrine Zufferey	University of Bern

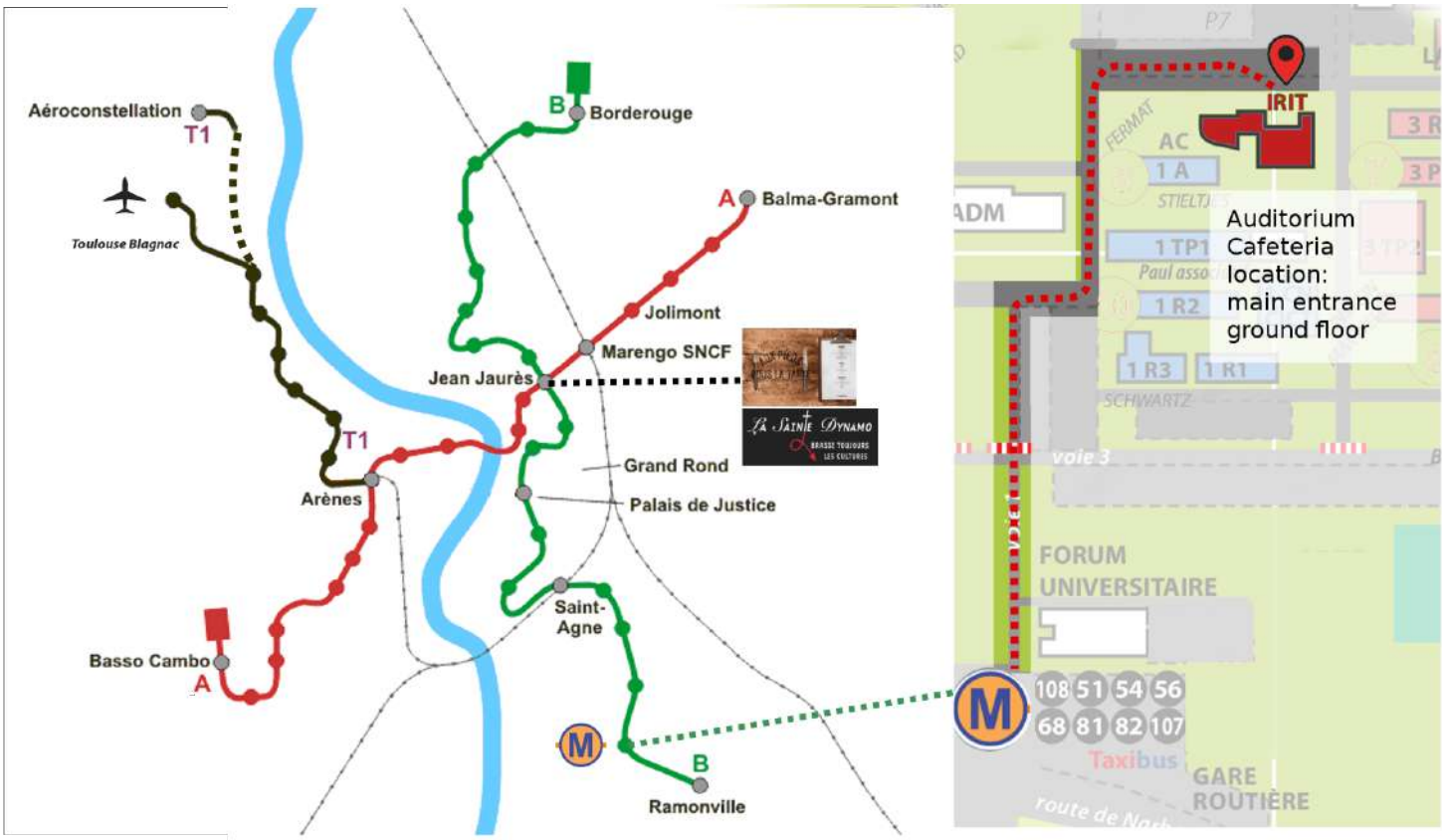
TextLink2018 -- Final Action Conference

Cross-Linguistic Discourse Annotation: applications and perspectives

Conference Programme

Dates	March 19-21, 2018
Venue	Toulouse, France University Paul Sabatier IRIT Rooms : Auditorium and cafeteria (ground floor)
Conference website	http://textlink.ii.metu.edu.tr/final-action-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 third and final Action Conference will be the occasion to take stock of the progress achieved, insisting on evaluation of discourse related resources and bridges between them, as well as opening to applications of the network results.
The meeting is open to everyone, both current members of TextLink and other researchers and practitioners working in the area.



Monday March 19th, IRIT Auditorium

Lunch (IRIT cafeteria)	
2.00 – 2.30	**Welcome and introduction of main results of the TextLink project** (L. Degand)
Auditorium 2.30 -- 4.00	• **Results of the Working Group 1 on ”Resources of DRDs”** (J. Mirovsky and A. Mendes) • **Results of the Working Groups 2 and 3 on ”Interoperable Annotation Guidelines” and “Assessment of Empirical and Cognitive Soundness”** (M. Stede, S. Zufferey, T. Sanders, H. Rohde) • **Results of the Working Group 4 on “Tools for discourse annotation and discourse parsing”** (P. Muller and B. Webber)
coffee break (IRIT cafeteria)	
Cafeteria 4:30 -- 6:00	**Posters** : ✢ Annotating the Meaning of Discourse Marker *so* in TED talks (G. V. Oleškevičienė, N. Burkšaitienė, S. Rackevičienė and L. Mockiene) ✢ Translation of “and” in a parallel TedTalk corpus of English, Czech, Hungarian, Lithuanian and French: functions and omissions (Á. Abuczki, N. Burkšaitienė, L. Crible, P. Furkó, A. Nedoluzhko, G. V. Oleškevičienė, S. Rackevičienė and S. Zikanova) ✢ Adding Senses and New Discourse Relations to Turkish Discourse Bank: Recent Updates (D. Zeyrek, N. Soycan, A. Burcu Güven and M. Kurfalı) ✢ TED Multilingual Discourse Bank: A Parallel Resource Annotated in the PDTB Style (D. Zeyrek, A. Mendes and M. Kurfalı) ✢ A FrameNet lexicon and annotated corpus as DRD resource: Causality in the ASFALDA French FrameNet (L. Vieu) ✢ The path to hearer-old status: Modeling how entities enter common sense knowledge (I. Staliunaite) ✢ French causal connectives at the Grammar-Discourse interface (H. Jivanyan) **Demos** : ✢ TextLink Web Portal (M. Kurfalı, A. Üstün and B. Webber) ✢ A multilingual database of connectives: connective-lex.info (T. Scheffler, M. Stede, P. Bourgonje and F. Dombek) ✢ Describing CzeDLex – a Lexicon of Czech Discourse Connectives (M. Rysová, L. Poláková, J. Mírovský and P. Synková)

8:00 pm
Social Dinner « Aux Pieds Sous La Table »
4-8 Rue Arnaud Bernard, 31000 Toulouse

Tuesday March 20th, IRIT

Auditorium 9:00 – 11:00	• **Aligning connective lexicons for a multilingual database** (Y. Grishina, P. Bourgonje and M. Stede) • **Functions and domains of discourse markers across languages: Testing a two-dimensional annotation scheme** (L. Degand, L. Crible and K. Grzech) • **Annotating Discourse Markers in the MULTINOT corpus: The case of elaborating connectives** (J. Lavid and E Avilés) • **A bottom-up analysis of sentence-initial DRDs in the Finnish Internet** (V. Laippala, A. Kyröläinen, F. Ginter, J. Kanerva, J. Komppa and J. Kalliokoski)
coffee break (IRIT cafeteria)	
Auditorium 11:30 – 12:30	Invited speaker : **Anette Frank** **Resolving Abstract Anaphors in Discourse** **— Uphill Battles with Neural Networks and Automatic Data Generation** *Department of Computational Linguistics, Heidelberg University, Germany*
Lunch (IRIT cafeteria)	
Cafeteria 2:00 -- 4:00	**Posters** : ✦ Identifying DRDs through automatic semantic annotation - Using the UCREL Semantic Analysis System as a pre-annotation tool (P. Furkó) ✦ Correlating DRDs with other types of discourse phenomena: Cross-linguistic analysis of the interplay between DRDs, coreference and bridging (E. Lapshinova-Koltunski and A. Nedoluzhko) ✦ Identification of Thematic Discourse Relations on the Data from an Annotated Corpus of Czech (E. Hajičová and J. Mírovský) ✦ The linguistic marking of coherence relations: The interaction between segment-internal elements and connectives (J. Hoek, S. Zufferey, J. Evers-Vermeul and T. Sanders) ✦ Disambiguating discourse relations with or without a connective: Does “and” really say nothing? (L. Crible and V. Demberg) ✦ Using annotation to identify connective meanings in a multilingual environment (S. Postolea) ✦ Discourse Connectives and Reference (K. Rysová and M. Rysová) ✦ Exploring a corpus annotated in causal discourse relations for the study of causal lexical clues (C. Atallah, M. Bras and L. Vieu) ✦ Naïve annotations of French et and alors: comparison with experts and effect of implicitation (I. Didirkova, G. Christodoulides, L. Crible and A.-C. Simon) **Demos** : ✦ TextLink Web Portal (M. Kurfalı, A. Üstün and B. Webber) ✦ A multilingual database of connectives: connective-lex.info (T. Scheffler, M. Stede, P. Bourgonje and F. Dombek) ✦ Describing CzeDLex – a Lexicon of Czech Discourse Connectives (M. Rysová, L. Poláková, J. Mírovský and P. Synková)
coffee break (IRIT cafeteria)	
Auditorium 4:30 – 6:00	• **Co-occurrence of discourse markers: from juxtaposition to composition** (L. Crible and M. J. Cuenca) • **The automatic analysis of subjectivity and causal coherence in text** (W. Spooren and T. Sanders) • **Testing the interoperability of annotation systems for oral DRDs in Spanish language** (E. Pascual-Aliaga)

ABBAYE DE LA SAINTE DYNAMO
BRASSE TOUJOURS LES CULTURES

7:00 pm
A drink at « La Sainte Dynamo »
6/8 Rue Amélie, 31000 Toulouse

Wednesday March 21th, IRIT Auditorium

9:00 -- 10:30	• **Unifying dimensions in coherence relations: How various annotation frameworks are related** (T. Sanders, V. Demberg, J. Hoek, M. Scholman, S. Zufferey and J. Evers-Vermeul) • **Designing a corpus-based lexicon for spoken DRDs: semantic considerations** (L. Crible and A. Mendes) • **Choosing among alternatives: Conjunction variability comes from both inference and the semantics of discourse adverbials** (H. Rohde, A. Johnson, N. Schneider and B. Webber)
coffee break (IRIT cafeteria)	
11:00 -- 12:30	• **For example, specifically, or because; Individual differences in coherence relation interpretation biases?** (M. Scholman, V. Demberg and T. Sanders) • **Discourse relations with explicit and implicit arguments: The case of European Portuguese "aliás"** (P. Lejeune and A. Mendes)
12:30 – 1:00	• **Closing**
Lunch (IRIT cafeteria)	
4:30 -- Visit of *le Cloître des Jacobins*	

Program Committee:
Maria Josep Cuenca (University of València)
Liesbeth Degand (University of Louvain)
Peter Furkó (Károli Gáspár University of the Reformed Church in Hungary)
Daniel Hardt (Copenhagen Business School)
Lydia-Mai Ho-Dac (University of Toulouse)
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 (Fribourg University)

Organizers:

Lydia-Mai Ho-Dac (University of Toulouse, CLLE-ERSS)

Philippe Muller (University of Toulouse, IRIT)

Contents

Invited Talks

Resolving Abstract Anaphors in Discourse — Uphill Battles with Neural Networks and Automatic Data Generation 2
Anette Frank

Regular Papers

Translation of ”and” in a parallel TED Talk corpus of English, Czech, Hungarian, Lithuanian and French: functions and omissions 4
Ágnes Abuczki, Nijolė Burkšaitienė, Ludivine Crible, Péter Furkó, Anna Nedoluzhko, Giedre V. Oleškevičienė, Sigita Rackevičienė and Sarka Zikanova

Exploring a corpus annotated in causal discourse relations for the study of causal lexical clues . . 12
Caroline Atallah, Myriam Bras and Laure Vieu

Co-occurrence of discourse markers: from juxtaposition to composition 19
Ludivine Crible and Maria Josep Cuenca

Disambiguating discourse relations with or without a connective: Does ”and” really say nothing? . 24
Ludivine Crible and Vera Demberg

Designing a corpus-based lexicon for spoken DRDs: semantic considerations 29
Ludivine Crible and Amalia Mendes

Functions and domains of discourse markers across languages: Testing a two-dimensional annotation scheme 34
Liesbeth Degand, Ludivine Crible and Karolina Grzech

Naïve annotations of French et and alors: comparison with experts and effect of implicitation . . . 38
Ivana Didirkova, George Christodoulides, Ludivine Crible and Anne-Catherine Simon

Identifying DRDs through automatic semantic annotation - Using the UCREL Semantic Analysis System as a pre-annotation tool 45
Péter Furkó

Aligning connective lexicons for a multilingual database 52
Yulia Grishina, Peter Bourgonje and Manfred Stede

Identification of Thematic Discourse Relations on the Data from an Annotated Corpus of Czech . 56
Eva Hajičová and Jiří Mírovský

The linguistic marking of coherence relations: The interaction between segment-internal elements and connectives . . . 64
Jet Hoek, Sandrine Zufferey, Jacqueline Evers-Vermeul and Ted Sanders

TextLink Web Portal . . . 68
Murathan Kurfali, Ahmet Üstün and Bonnie Webber

A bottom-up analysis of sentence-initial DRDs in the Finnish Internet . . . 72
Veronika Laippala, Aki-Juhani Kyröläinen, Filip Ginter, Jenna Kanerva, Johanna Komppa and Jyrki Kalliokoski

Correlating DRDs with other types of discourse phenomena: Cross-linguistic analysis of the interplay between DRDs, coreference and bridging . . . 83
Ekaterina Lapshinova-Koltunski and Anna Nedoluzhko

Annotating Discourse Markers in the MULTINOT corpus: The case of elaborating connectives . . . 89
Julia Lavid and Estefanía Avilés

Discourse relations with explicit and implicit arguments: The case of European Portuguese "aliàs" 91
Pierre Lejeune and Amália Mendes

Testing the interoperability of annotation systems for oral DRDs in Spanish language . . . 96
Elena Pascual-Aliaga

Using annotation to identify connective meanings in a multilingual environment. Romanian and english contrast markers in a parallel corpus . . . 107
Sorina Postolea

Choosing among alternatives: Conjunction variability comes from both inference and the semantics of discourse adverbials . . . 114
Hannah Rohde, Alexander Johnson, Nathan Schneider and Bonnie Webber

Discourse Connectives and Reference . . . 122
Kateřina Rysová and Magdaléna Rysová

Describing CzeDLex – a Lexicon of Czech Discourse Connectives . . . 129
Magdaléna Rysová, Lucie Poláková, Jiří Mírovský and Pavlína Synková

Annotation proposal for Sp. DRDs and their interaction with other units in written texts: an analysis from the Val.Es.Co. discourse segmentation model . . . 136
Shima Salameh Jiménez

Unifying dimensions in coherence relations: How various annotation frameworks are related . . . 139
Ted Sanders, Vera Demberg, Jet Hoek, Merel Scholman, Sandrine Zufferey and Jacqueline Evers-Vermeul

A multilingual database of connectives: connective-lex.info . . . 144
Tatjana Scheffler, Manfred Stede, Peter Bourgonje and Felix Dombek

For example, specifically, or because; Individual differences in coherence relation interpretation biases? . . . 151
Merel Scholman, Vera Demberg and Ted Sanders

The automatic analysis of subjectivity and causal coherence in text . . . 160
Wilbert Spooren and Ted Sanders

Annotating the Meaning of Discourse Marker so in TED talks 165
Giedre Valunaite Oleskeviciene, Nijole Burksaitiene, Sigita Rackeviciene and Liudmila Mockiene

A FrameNet lexicon and annotated corpus as DRD resource: Causality in the ASFALDA French FrameNet 172
Laure Vieu

TED Multilingual Discourse Bank: A Parallel Resource Annotated in the PDTB Style 179
Deniz Zeyrek, Amalia Mendes and Murathan Kurfali

Adding Senses and New Discourse Relations to Turkish Discourse Bank: Recent Updates 185
Deniz Zeyrek, Nihan Soycan, Arzu Burcu Güven and Murathan Kurfali

Invited Talks

Resolving Abstract Anaphors in Discourse — Uphill Battles with Neural Networks and Automatic Data Generation

Anette Frank
University of Heidelberg

Regular Papers

Translation of “and” in a parallel TED Talk corpus of English, Czech, Hungarian, Lithuanian and French: functions and omissions

Abuczki, Ágnes[1], Burkšaitienė, Nijolė[2], Crible, Ludivine[3], Nedoluzhko, Anna[4], Furkó, Péter[5], Valūnaitė Oleškevičienė, Giedre[6], Rackevičienė, Sigita[7] and Zikánová, Šárka[8]

[1] MTA-DE-SzTE Research Group for Theoretical Linguistics, Uni. of Debrecen, Hungary
abuczki.agnes@gmail.com
[2] Mykolas Romeris University, Vilnius, Lithuania
n.burksaitiene@mruni.eu
[3] Université catholique de Louvain, Belgium
ludivine.crible@uclouvain.be
[4] Charles University in Prague, Czech Republic
nedoluzko@ufal.mff.cuni.cz
[5] Károli Gáspár University of the Reformed Church, Hungary
furko.peter@gmail.com
[6] Mykolas Romeris University, Vilnius, Lithuania
gvalunaite@mruni.eu
[7] Mykolas Romeris University, Vilnius, Lithuania
sigita.rackeviciene@mruni.eu
[8] Charles University in Prague, Czech Republic
zikanova@ufal.mff.cuni.cz

Abstract. In this paper we report on the methods and findings of a multilingual corpus study focusing on the functions of *and* in English and its translations into Czech, French, Hungarian and Lithuanian, in a selection of TED Talks. Firstly, we outline the functions of the discourse marker *and*, as it has been previously described in the literature. Secondly, we describe our annotation scheme [1] and our results.

We address the following research questions: (1) What is the functional spectrum of *and* in English TED Talks? (2) How is *and* translated in Czech, French, Hungarian and Lithuanian? Are specific functions of *and* associated with specific translations? (3) Which uses of *and* tend to be omitted in the translations?

Keywords: Discourse Marker, Discourse Connective, Crosslinguistic, Implicitation, Underspecification, Translation Corpora, TED Talks.

1 Introduction

Discourse markers which are defined as joining one sentence with another sentence or one paragraph to another paragraph or even one idea to another. Wrong use of discourse markers may result in hindered communication. Discourse markers are grammatically heterogeneous, multifunctional pragmatic markers [2] that include coordinating conjunctions (*and, but, or*), subordinating conjunctions (*because, although*), adverbs (*well, actually*), verbal phrases (*you know, I mean*), prepositional phrases (*in fact*) and have the function to convey a coherence relation. The difficulty to conduct cross-linguistic comparisons of discourse markers is determined by their polysemy and the ways of expressing coherence relations used in different languages. The problems related to discourse markers become a particular challenge for translators who have to adapt them to a new language and culture, in which textual strategies involving their use are often different from those of the source text [3].

The present study aimed to annotate the English discourse marker *and* cross-linguistically using spoken corpus data from the multilingual corpus TED Talks. The investigation was conducted in two stages, including annotating the domain and functions of the discourse marker *and* in English and its counterparts in Czech, French, Hungarian and Lithuanian followed by the analysis of the translations of this discourse marker into Czech, French, Hungarian and Lithuanian.

2 Theoretical background

Discourse markers have several functions: they connect single text segments into a compound unit, they express a semantic type of relation (contrast, reason, instantiation, etc.) and express various pragmatic functions. Highly frequent among them, the additive discourse marker *and* encodes very little information in its core meaning. Yet, it is used in a variety of contexts where additional meanings can be identified, such as contrast or consequence.

According to Schiffrin, *and* has two basic discourse uses: coordination and continuation. It is a structural device for building text which coordinates ideas and units, however, it has little semantic meaning. Besides, *and* also has contrastive uses, e.g. *We tried to win. And we lost.* It can also preface the outcome of a reason, e.g. *That's one game I remember because we had a driveway and, like we would hide, and they would walk around the driveway? Y'know? And I- I remember it so distinctly* [4]. Moreover, *and* can connect events: [POSITION EVENT] *and* [SUPPORT EVENT] *and* [EVENT], and it can also connect reasons or two pieces of support at a higher level of idea structure: [POSITION SUPPORT 1] *and* [SUPPORT 2]. *And* can connect a general conclusion drawn from a list of specific events which are asyndetically connected, e.g. *I uh I go on trips with 'em, I bring 'em here, we have supper, or dinner here, and I don't see any problem because I'm workin' with college graduates.* [4]. *And* often links structurally similar clauses as well and it doesn't favor tense switching (unlike temporal connectives). *And* also has pragmatic effect: as a marker of speaker-continuation in interaction (which is a consequence of the speaker's situated context-bound use). It can be used to (try to) reopen an interactional unit whose

completion has earlier been interrupted (turn-taking) or to continue/return to a previous question as a request for elaboration (turn-giving): QUESTION 1 ANSWER *and* QUESTION 1. On the other hand, *and* can be used to link pre-arranged questions in a question agenda: QUESTION 1 ANSWER *and* QUESTION 2 [4].

Crible & Degand's taxonomy of domains and functions of discourse markers [1], which is specifically designed for annotating discourse markers used in spoken discourse, consists of four main domains. First, the ideational domain is linked to "states of affairs in the world, semantic relations between real events". Second, the rhetorical domain is linked to "the speaker's meta-discursive work on the ongoing speech". Third, the sequential domain is linked to "the structuring of discourse segments, both at macro- and micro-level." Finally, the interpersonal domain is linked to "the interactive management of the exchange, in other words, to the speaker-hearer relationship" [5].

Table 1. Our annotation scheme: Crible & Degand's revised taxonomy with cross-domain functions [1]

Ideational	**Rhetorical**	**Sequential**	**Interpersonal**
[addition] [alternative] [cause] [closing] [concession] [condition] [consequence] [contrast] [enumeration] [opening] [punctuation] [resuming] [temporal] [topic-shift] [specification]			

In addition to the domains, fifteen functions can be assigned to the discourse markers, including addition, contrast or specification. Domains and functions are independent: any domain can apply to any function and any function can apply to any domain. According to Crible & Degand, annotators "can choose to start at domain-level or function-level, to annotate both levels simultaneously or independently, and could even decide to stop at one level if a particular domain DM token is underspecified for the other level" [1] which enhances the annotation process. Also, the authors believe that this system vouches for reliable annotation (high inter-annotator agreement), because of the reduced number of labels and the independence of the two levels (i.e. domains and functions).

3 Methodology

In this research study, all occurrences of English *and* as a discourse marker along with their translations into Czech, French, Hungarian and Lithuanian have been manually identified in a selection of TED Talks. The originals and their translations were annotated in each language by two experts, following Crible & Degand's functional classification [1].

A pilot study has been carried out and its first results point at regular tendencies regarding implicitation, multiple translation equivalents and functional shifts of *and* across languages. The methodological decisions made in the course of research are related to the choice of the corpus and the annotation method. These choices have

been determined by the aim of the research, that is, to annotate the English spoken discourse marker *and*, to compare its functions with its counterparts in Czech, French, Hungarian and Lithuanian as well as to analyze the translations of *and* into Czech, French, Hungarian and Lithuanian. For these reasons, the multilingual translations of TED Talks were chosen. This choice was made on the grounds that parallel texts are considered to be ideal for optimal comparability between languages as they provide more flexible and accurate ways to compare discourse markers [2]. The choice of the functional approach to be used for this investigation was predetermined by the specific nature of discourse markers, which covers some specific features, e.g. even though most languages possess discourse markers, they have a high degree of contextual variation [1].

The empirical research consists of two stages. Initially, the discourse marker *and* is compared to its Czech, French, Hungarian and Lithuanian counterparts by applying Crible & Degand's [1] taxonomy of domains and functions of discourse markers. Then, the translations of *and* into Czech, French, Hungarian and Lithuanian, found in the annotated samples, are analyzed.

4 Research findings

4.1 Distribution of functions

The DM *and* expresses a very wide functional spectrum, much larger than simply "addition". The functions from the original English are not necessarily the same in the translations. As a result, the functional spectrum of the translation equivalents of "and" in CZ, HU, LIT and FR differ in terms of the types of functions/domains and their proportions. The research reveals that in the annotated sample, the discourse marker *and* with its counterparts in Czech, French, Hungarian and Lithuanian is mainly used in the ideational domain, marking/connecting factual information, and in the sequential domain, representing the structuring of local and global units of discourse and less often in the rhetorical domain which is related to the speaker‘s subjectivity.

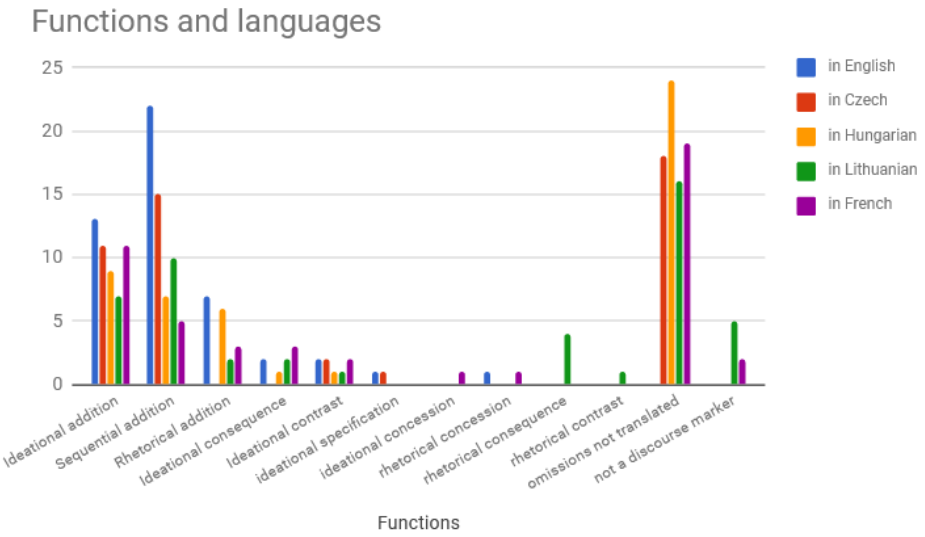


Fig. 1. Distribution of the functions of the translations of AND across languages

It can be seen in Fig.1 that cross-linguistically sequential domain stands out demonstrating the use of the discourse marker *and* for discourse structuring purposes.

4.2 Omissions

Another striking feature in the figures is frequent omissions in cross-linguistic translation of the discourse marker *and*. Cross-linguistically, the uses of *and* in the sequential domain are the most frequently omitted, whereas *and*'s operating in the ideational domain are usually preserved. SEQ-ADD is the basic continuation function and does not bring a lot of information, whereas IDE-ADD really signals a true semantic addition like a "plus" sign, so we lose less information by removing SEQ-ADD than IDE-ADD. Some *and*'s are maintained to avoid juxtaposition and some are removed due to constraints of the translation by subtitles.

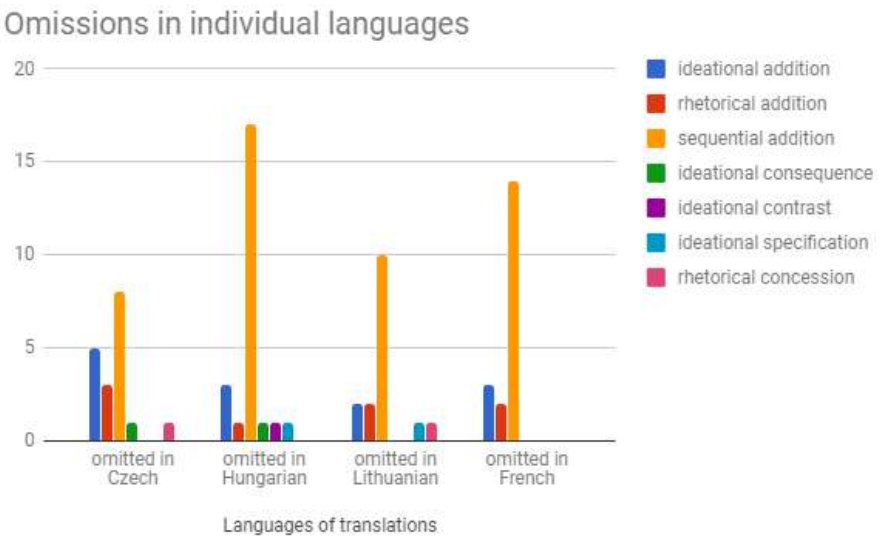


Fig. 2. Distribution of the functions of the omitted English AND's across languages

4.3 Translations

When *and* in sequential or rhetorical addition is translated in French, *et* is annotated by many different labels. In Lithuanian the rhetorical domain also uncovers a number of differences. Here we find rhetorical contrast rendered into the Lithuanian *o* which could represent both addition and contrast. Rhetorical consequence is also observed which is mostly related to the whole argument. The reason could be the nature of the domain since it is related to the speaker's subjectivity.

A striking regularity observed cross-linguistically is the use of *and* for topicalization of the previous focus (in the relation of addition):

- En There'd be a huge spread in her scores. [And] [actually] it's this spread that counts.
- Cz V jejím hodnocení by byl velký rozptyl. [A] [právě] na tomto rozptylu záleží.
- Hu Nagy lesz a szórása a pontoknak. [És] ez az a szórás, ami számít.
- Li Jos balai būtų visiškai pasiskirstę. [Ir] [išties], svarbus būtent tas pasiskirstymas.
- Fr omission of the discourse connective

Similarly, *and* is regularly used in many languages as means of expressing temporal succession of two events:

- En *We give* is mainly used in the ideational domain, marking/connecting factual information, and in the sequential domain,*a little bit of time to play the field, get a feel for the marketplace or*
- *whatever when we're young. [And] then we only start looking seriously at potential marriage candidates once we hit our mid-to-late 20s.*
- Cz *Dopřejeme si čas na hraní a průzkum toho, co je k mání, dokud jsme ještě mladí. [A]*
- *vážný konkurz na kandidáty pod čepec zahájíme až po pětadvacítce.*
- Li *Leidžiam sau šiek tiek išsilakstyti, kol esam jauni, leidžiam suprasti, kas yra rinkoje, ar panašiai. [Ir] tuomet vėlesniame dvidešimtmetyje pradedame į vedybų kandidatus žiūrėti rimtai.*
- Hu, Fr *omission of the discourse connective*

5 Conclusions

In summary, our study of the functions and translations of *and* in a parallel corpus across five languages has shown that the most frequent functions of *and* in the particular register of TED Talks are in the ideational domain, marking/connecting factual information, and in the sequential domain, representing the structuring of local and global units of discourse. The research also reveals systematic tendencies regarding the translation of *and*. In particular, *and* is most frequently omitted in the translations when it expresses sequential addition, that is, a basic continuity function. This study is the first step of a larger research project on the uses of underspecified connectives such as *and*. In future, we will replicate the analysis on a larger sample, and investigate additional connectives such as *now*, *so* or *but*.

References

1. Crible, L.; Degand, L.: Reliability vs. Granularity in discourse annotation: What is the trade-off? In: Corpus Linguistics and Linguistic Theory 14(2), 1–29 (2017).
2. Crible, L.: Discourse markers, (dis)fluency and the non-linear structure of speech: a contrastive usage-based study in English and French. Université Catholique de Louvain, Louvain-la-Neuve (2017).
3. Zufferey, S., Degand, L.: Annotating the meaning of discourse connectives in multilingual corpora. In: Corpus Linguistics and Linguistic Theory 13(2), 1–24 (2017).
4. Schiffrin, D.: Discourse Markers. Cambridge University Press, Cambridge (1987).
5. Crible, L.: Identifying and describing discourse markers in spoken corpora. Université Catholique de Louvain, Louvain-la-Neuve (2014).

Acknowledgements

Work on this paper has been supported by COST Action IS1312, in the framework of the TextLink project, as well as by the Grant Agency of the Czech Republic (pro-

jects GA16-05394S and GA 17-03461S). The research contribution of Ágnes Abuczki to the present study has been supported by the National Research, Development and Innovation Office of Hungary (NKFIH), research project code: PD121009.

Exploring a corpus annotated in causal discourse relations for the study of causal lexical clues

Caroline Atallah[1], Myriam Bras[2], Laure Vieu[3]

[1] LIDILE, Université de Rennes, France
[2] CLLE, Université de Toulouse, CNRS, UT2J, France
[3] IRIT, CNRS, Université de Toulouse, France

1 Introduction

Usually, the study of Discourse Relations (DRs) is based on Lexical Clues (LCs) commonly associated with these DRs, like connectives. For example, a corpus study of causal DRs can be done from the analysis of some connectives commonly associated with causality, like *because*. Such a semasiological approach, that proceeds from a given LC towards DRs, has a significant advantage: it is much easier to locate LCs than DRs in a corpus.

The approach presented here complementarily exploits two types of analysis. We first adopt an onomasiological approach, that proceeds from a given DR towards LCs. In other words, we analyze all the occurrences of this DR in a corpus in order to identify all the LCs that contribute to the DR interpretation. Then, the results of these first analyses are completed by a semasiological analysis: each LC that has been identified is projected on the corpus in order to determine whether it specifically marks the given DR or not.

The onomasiological approach requires working on data that have previously been annotated with DRs. Before the ANNODIS corpus was built (*ANNotation DIScursive de corpus*; Péry-Woodley et al., 2009, 2011; Afantenos et al., 2012), such data did not exist for French and an onomasiological approach, as presented above, was simply impossible for this language. This rather new methodology has already been applied to a few DRs on the ANNODIS corpus (see Vergez-Couret, 2010, for an application to *Elaboration* DR). We propose to focus here on a specific family of DRs: causal DRs, and to base our study on a corpus specifically annotated with causal DRs: the EXPLICADIS corpus (*EXPLication et Argumentation en DIScours*; Atallah, 2014; Atallah, 2015).

2 The EXPLICADIS corpus

In the ANNODIS project, 86 texts were segmented into Elementary Discourse Units (EDUs) and then annotated with a tagset of DRs inspired by SDRT relations (*Segmented Discourse Representation Theory*, Asher and Lascarides, 2003). The EXPLICADIS corpus has been built in the continuity of ANNODIS: the 86 texts were reused and re-annotated with a more complete and accurate new set of causal DRs.

Then 31 more texts were added, segmented and annotated in order to provide a better representation of different text genres: narrative, expositive and argumentative. The whole EXPLICADIS corpus includes 117 texts, 4,580 EDUs and 39,103 tokens.

This new set of causal DRs was adopted in order to remedy the difficulties experienced by ANNODIS annotators with the first set of DRs and to adequately account for the data in a semantically clear set of relations (Atallah, 2014; Atallah et al., 2016). It includes, like the previous one, two types of relations: *Explanation* relations (noted further Rh_Exp) and *Result* relations (noted further Rh_Res)[1]. The new set is original because it distinguishes within both rhetorical types four subtypes of DRs:

- content-level DRs that involve a causal link between the eventualities that are described in the propositional content: *Explanation (α,β)* (1) and *Result (α,β)* (2);
- epistemic DRs that involve a causal link between knowledge items and beliefs: $Explanation_{ep}$ *(α,β)* (3) and $Result_{ep}$ *(α,β)* (4) ;
- inferential DRs that involve a causal link between knowledge items: $Explanation_{inf}$ *(α,β)* (5) and $Result_{inf}$ *(α,β)* (6);
- speech-act (or pragmatic) DRs that involve a causal link between an eventuality that is described in the propositional content and a speech act: $Explanation_{prag}$ *(α,β)* (7) and $Result_{prag}$ *(α,β)* (8).

A total of 319 causal DRs were annotated using this tagset, including 186 Rh_Exp relations and 133 Rh_Res relations. Examples of each type of these DRs are presented below:

(1) [L'armée est déçue,]$_{\alpha}$ [il n'y a aucun viol, aucun pillage, aucun meurtre.]$_{\beta}$
([The army is disappointed,]$_{\alpha}$ [there is no rape, no looting, no murder.]$_{\beta}$)

(2) [le côté gauche de la voiture a mordu l'accotement.]$_{\alpha}$ [L'automobile a perdu sa roue gauche.]$_{\beta}$
([the left side of the car hit the roadside.]$_{\alpha}$ [The car lost its left wheel.]$_{\beta}$)

(3) [Ce phénomène semble se confirmer à Mariana,]$_{\alpha}$ [où on peut observer deux voies parallèles à la sortie sud de la ville.]$_{\beta}$
([This phenomenon seems to be confirmed in Mariana,]$_{\alpha}$ [where two parallel roads can be observed at the south exit of the city.]$_{\beta}$)

(4) [Or la psychomécanique répond à ces deux types d'exigences.]$_{\alpha}$ [Il serait donc intéressant de regarder si les outils théoriques qu'elle a développés permettent de rendre compte de certaines observations faites par la neuropsychologie.]$_{\beta}$
([Yet psychomechanics meets these two types of requirements.]$_{\alpha}$ [It would therefore be interesting to examine whether the theoretical tools it has developed are able to account for certain observations made by neuropsychology.]$_{\beta}$)

[1] « Rh_ » is put for « Rhetorical ». We consider that *Explanation* relations and *Result* relations do not simply differ in the order of presentation, but rather in the rhetorical choice of presentation.

(5) [BITNET était différent d'Internet]$_{\alpha}$ [parce que c'était un réseau point-à-point de type « stocké puis transmis ».]$_{\beta}$
([BITNET was different from Internet]$_{\alpha}$ [because it was a point-to-point network of "stored and transmitted" type.]$_{\beta}$)

(6) [La première exposition avicole de Belfort date de 1922.]$_{\alpha}$ [Cela fait donc plus de trois-quarts de siècle que la digne société du même nom encourage, dans la région, les éleveurs amateurs.]$_{\beta}$
([The first avicultural Belfort exhibition dates back to 1922.]$_{\alpha}$ [Therefore, the honorable society of that name has been supporting farmers for more than seven decades.]$_{\beta}$)

(7) [Mais que ces derniers se rassurent,]$_{\alpha}$ [il y aura encore deux autres tours pour se rattraper.]$_{\beta}$
([These can rest assured[2] that]$_{\alpha}$ [there will be two more rounds to catch up.]$_{\beta}$)

(8) [Suzanne Sequin n'est plus.]$_{\alpha}$ [...] [Nos condoléances.]$_{\beta}$
([Suzanne Sequin is gone.]$_{\alpha}$ [...] [Our condolences.]$_{\beta}$)

After annotation, each of these DRs has been analyzed in order to identify lexical clues (LCs). Within LCs, we draw a line between *clues* and *markers*. We consider that a *clue* is a linguistic unit that plays a potential role in the DR interpretation; while a *marker* has an established function in discourse interpretation, it plays a primordial role in the inference of a DR (Vergez-Couret, 2010; Péry-Woodley, 2000). Thus, for us, a clue is just a potential marker.

To identify causal LCs, we tried to spot every LCs that could have helped to guide our interpretation to a causal DR during the annotation process. It is important to note that those LCs are not necessarily responsible (on their own) for the inference of the causal DR. We consider that actually, in most cases, it is a whole bundle of clues that contributes to the inference of a DR. Thus, by *LCs* we do not mean *discourse markers*, but a simple clue that accompany the DR. To determine the discursive function(s) associated with a LC requires a more in-depth study than the one presented here, a semasiological study of bigger data.

The onomasiological approach we first adopted has its own advantages. For example, it allowed us to study causal DRs associated to LCs but also DRs being annotated without the help of any LC. Those represent 38.87% of the annotated causal DRs. We noticed that the presence of LCs was related to the rhetorical choice, the type of causal DR, but also the text genre. The methodology adopted also helped listing causal LCs, and thereby noticing that LCs associated with Rh_Exp DRs were more diversified (31 LC types for 186 DR occurrences) than LCs associated with Rh_Res (21 LC types for 133 DR occurrences). This observation must however be considered carefully, given the small size of the corpus.

[2] This translation does not keep the imperative form of the verb, impossible in English with a third person. The French construction is similar to a English "But rest assured," in which the imperative is directed to the addressee instead.

3 The LEX-PLICADIS database

We compared our LC list with another existing inventory: LexConn (Roze, 2009; Roze et al., 2012). This resource lists French connectives and associates each of them to one or more DRs[3]. The causal DRs used in LexConn is the classical SDRT set, only including two types of causal DRs: content-level and speech-act relations. Thus, to compare EXPLICADIS LCs with LexConn LCs, we consider that LexConn speech-act causal DRs correspond to one of the three following types of DRs: epistemic, inferential or speech-act DRs.

Among the 52 different LCs we identified, 23 LCs were not recorded at all in LexConn and 4 were listed but not associated with causality. We therefore decided to complete LexConn with EXPLICADIS data in order to create a new database: LEX-PLICADIS.

To fill it, we completed the onomasiological analysis with a semasiological one. We first projected each LC identified on the whole EXPLICADIS corpus, in order to verify whether it was specialized in the expression of causality or not. Results were then compared with LexConn. We also analyzed the 70 LCs that were associated with causality in LexConn but not in EXPLICADIS. Naturally, the absence of an association between a LC and a DR in EXPLICADIS does not question the information listed in LexConn. Such a study should be continued on a larger annotated corpus. We therefore decided to be as exhaustive as possible and to record all the causal LCs identified in EXPLICADIS and/or in LexConn, specifying if it was associated in each resource to:

- a content-level DR;
- an epistemic causal DR;
- an inferential causal DR;
- a speech-act (or pragmatic) causal DR;
- a non-causal DR.

The complete database includes 120 causal LCs, among which 67 LCs associated with Rh_Exp DRs and 53 with Rh_Res DRs. We provide in table 1 an excerpt that concerns the 52 LCs we identified in EXPLICADIS and associated with the expression of causality.

Table 1. Excerpt of the LEX-PLICADIS database

LC	Rh_Exp DRs				other DRs
	content-level	epistemic	inferential	speech-act	
à cause de	L- E+	L- E-	L- E-	L- E-	L- E-
à la suite de	L- E+	L- E-	L- E-	L- E-	L- E+
avec	L- E+	L- E-	L- E-	L- E-	L- E+

[3] It is interesting to note that LexConn had been partly built on the basis of the LCs listed in the ANNODIS annotation guide.

car	L- E+	L* E+	L* E+	L* E-	L- E-
comme	L+ E+	L* E-	L* E-	L* E-	L+ E+
conséquence de	L- E+	L- E-	L- E-	L- E-	L- E-
d'autant plus que	L+ E+	L- E-	L- E-	L- E-	L- E-
d'autant que	L+ E-	L- E+	L- E-	L- E-	L- E-
dans la mesure où	L- E-	L* E+	L* E-	L* E-	L+ E-
de	L- E+	L- E-	L- E-	L- E-	L- E+
dès que	L+ E+	L- E-	L- E-	L- E-	L+ E+
des suites de	L- E+	L- E-	L- E-	L- E-	L- E-
devant	L- E+	L- E-	L- E-	L- E-	L- E+
du fait de	L- E+	L- E-	L- E-	L- E-	L- E-
en+Verb-*ANT [gerund]*	L+ E+	L- E-	L- E-	L- E-	L+ E+
en effet	L- E+	L* E+	L* E+	L* E-	L- E+
en raison de	L- E+	L- E-	L- E-	L- E-	L- E-
en témoignage de	L- E+	L- E-	L- E-	L- E-	L- E-
étant donné	L- E-	L- E+	L- E+	L- E-	L- E-
étant donné que	L+ E-	L- E+	L- E-	L- E-	L- E-
faute de	L+ E+	L- E-	L- E-	L- E-	L- E-
grâce à	L- E+	L- E-	L- E-	L- E-	L- E-
le temps de	L- E+	L- E-	L- E-	L- E-	L- E+
par	L- E+	L- E-	L- E-	L- E-	L- E+
parce que	L+ E+	L* E+	L* E+	L* E-	L- E-
pour	L- E+	L- E-	L- E-	L- E-	L+ E+
pour des raisons (de)	L- E+	L- E+	L- E-	L- E-	L- E-
puisque	L+ E+	L* E+	L* E-	L* E-	L- E-
si... c'est que	L- E+	L- E-	L- E-	L- E-	L- E-
suite à	L- E+	L- E-	L- E-	L- E-	L- E+
vu	L- E-	L- E+	L- E-	L- E-	L- E-

LC	Rh_Res DRs				other DRs
	content-level	epistemic	inferential	speech-act	
à ce rythme	L- E-	L- E-	L- E+	L- E-	L- E-
ainsi	L+ E+	L- E+	L- E+	L- E-	L- E+
alors	L+ E+	L* E-	L* E-	L* E-	L+ E+
au point que	L+ E+	L- E-	L- E-	L- E-	L- E-
au prix de	L- E-	L- E+	L- E-	L- E-	L- E-
aussi [initial position]	L+ E-	L- E+	L- E-	L- E-	L- E-
avec pour conséquence	L- E+	L- E-	L- E-	L- E-	L- E-
c'est pourquoi	L+ E+	L- E+	L- E-	L- E-	L- E-
conduisant à	L- E+	L- E-	L- E-	L- E-	L- E-
de sorte que	L+ E+	L- E-	L- E+	L- E-	L- E-
dès lors	L- E+	L* E-	L* E-	L* E-	L- E-
donc	L+ E+	L* E+	L* E+	L* E-	L- E+
d'où	L+ E-	L- E-	L- E+	L- E-	L- E+

et	L- E+	L- E-	L- E-	L- E-	L+ E+
jusqu'à ce que	L+ E+	L- E-	L- E-	L- E-	L+ E+
pour	L- E+	L- E-	L- E-	L- E-	L+ E+
preuve que	L- E-	L* E+	L* E-	L* E-	L- E-
résultat(s)	L+ E+	L- E-	L- E-	L- E-	L- E-
si bien que	L+ E+	L- E-	L- E-	L- E-	L- E-
tant que	L- E+	L- E-	L- E-	L- E-	L+ E-
tel(les)... que	L- E+	L- E-	L- E-	L- E-	L- E-

"L-": LC absent in LexConn "E-": LC absent in EXPLICADIS
"L+": LC present in LexConn "E+": LC present in EXPLICADIS
"L*": LC associated in LexConn with a speech-act causal DR

The study of the repartition of each LC allowed us to test some hypotheses formulated in the literature. For example, we found, for Rh_Exp DRs, that the values originally associated with *parce que* and *car* (*because*) (Groupe λ-l, 1975; Degand and Fagard, 2008) still persisted: *car* is more subjective than *parce que* (Simon and Degand, 2007), ie it is more often associated with epistemic DRs than content-level DRs. And we found, for Rh_Res DRs, that *donc* (*therefore*) was specialized in inferential DRs. *Donc* forces some sort of inferential reading: in a content-level DR, the effect described is presented as an inevitable event, and in an epistemic DR, the conclusion is presented as an obvious and indisputable fact (Hybertie, 1996).

4 Conclusion

To build the new resource LEX-PLICADIS, onomasiological and semasiological approaches were used complementarily. Thanks to the onomasiological analysis, which consists in a sort of exhaustive exploration of the corpus, we got results that could not have been obtained otherwise, such as DR occurring without LC. It also enabled us to add to LexConn many associations between LCs and DRs that had not been envisaged. It was important to complete and test the LexConn proposals for the causality domain. The same work should be done with other domains in a method akin to the ASFALDA French FrameNet project's one (Djemaa et al., 2016).

However, as an onomasiological approach requires a corpus annotated with DRs and as such a corpus requires a long and hard work, it implies to work with small quantity of data and to accept that the corpus, because of its size, presents limitations. Therefore, the onomasiological study must be considered and adopted as a first exploratory and non-exhaustive phase of the analysis, which can be then completed by a semasiological study on a bigger corpus.

References

1. Afantenos, S. D., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, L.-M., Le Draoulec, A., Muller, P., Péry-Woodley, M.-P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M. et Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organization : the ANNODIS corpus. *In Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, pages 2727–2734, Istanbul, Turkey.
2. Asher, N., and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.
3. Atallah, C. (2014). *Analyse de relations de discours causales en corpus : étude empirique et caractérisation théorique*. Thèse de Doctorat, Université de Toulouse.
4. Atallah, C. (2015). La ressource EXPLICADIS, un corpus annoté spécifiquement pour l'étude des relations de discours causales. In *Actes de TALN 2015* (pp. 551–557). Caen.
5. Atallah, C., Vieu, L., Bras, M. (2016). Formal characterization of a new set of causal discourse relations. *NISM 2016 (New Ideas in Semantics and Modeling)*, Paris, 7-8 septembre 2016.
6. Degand, L. et Fagard, B. (2008). (Inter)subjectification des connecteurs : le cas de *car* et *parce que*. *Revista de Estudos Linguìsticos da Universidade do Porto*, 3(1):119–136.
7. Djemaa, M., Candito, M., Muller, P. and Vieu, L. (2016). Corpus annotation within the French FrameNet: a domain-by-domain methodology. *In* Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Goggi, and Grobelnik, editors, *Language Resources and Evaluation Conference (LREC)*, pages 3794–3801, Portoroz, Slovenia, 23-28 May 2016. *European Language Resources Association (ELRA)*.
8. Groupe λ-l (1975). *Car, parce que, puisque*. *Revue Romane*, 10(2):258–280.
9. Hybertie, C. (1996). *La conséquence en français*. L'essentiel français. Ophrys, Paris/Gap.
10. Péry-Woodley, M.-P. (2000). *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*. Mémoire d'HDR, Université Toulouse II Le Mirail, Toulouse.
11. Péry-Woodley, M.-P., Afantenos, S. D., Ho-Dac, L.-M. and Asher, N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *TAL*, 52(3):71–101.
12. Péry-Woodley, M.-P., Asher, N., Enjalbert, P., Benamara, F., Bras, M., Fabre, C., Ferrari, S., Ho-Dac, L.-M., Le Draoulec, A., Mathet, Y., Muller, P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M., Vieu, L. and Widlöcher, A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. *In Actes de TALN 2009*, Senlis, France.
13. Roze, C. (2009). *Base lexicale des connecteurs discursifs du français*. Mémoire de Master 2, Université Paris Diderot, Paris.
14. Roze, C., Danlos, L. and Muller, P. (2012). LEXCONN: A French Lexicon of Discourse Connectives. *Discours*, (10).
15. Simon, A. C. et Degand, L. (2007). Connecteurs de causalité, implication du locuteur et profils prosodiques : le cas de *car* et de *parce que*. *Journal of French Language Studies*, 17(03):323–341.
16. Vergez-Couret, M. (2010). *Etude en corpus des réalisations linguistiques de la relation d'Elaboration*. Thèse de Doctorat, Université Toulouse II Le Mirail.

Co-occurrence of Discourse Markers: From Juxtaposition to Composition

Ludivine Crible[1] and Maria Josep Cuenca[2]

[1] Université catholique de Louvain, Belgium
[2] Universitat de València, Spain
ludivine.crible@uclouvain.be
maria.j.cuenca@uv.es

Abstract. In this paper, we report on a qualitative analysis of co-occurring discourse markers in spoken English, that is sequences of adjacent discourse markers that belong to the same unit but may express different function(s). We examine several formal and functional features of these co-occurring strings on the basis of authentic corpus examples extracted from conversational data. In particular, we focus on scope, meaning-in-context, syntactic category and position. Our analysis reveals several degrees of integration, which are mainly distinguished by the scope and meaning-in-context of the markers. We pay particular attention to the variable case of *and then*, which instantiates different degrees in our cline of co-occurrence depending on the meaning that can be interpreted from the cluster (i.e. additive/temporal, consequential, or enumerative). We discuss the implications of such fine-grained distinctions for the perspective of corpus annotation.

Keywords: Co-occurrence, Compound Discourse Markers, Spoken Corpus Annotation.

1 Introduction

Among the vast literature on discourse markers (henceforth DMs) and discourse-relational devices in general, one aspect of their behaviour has been somewhat overlooked, namely their co-occurrence. It is frequently the case that two or more DMs co-occur, as in the case of and if, but when or so for instance if, where DMs only co-occur or are juxtaposed, or in the case of but actually, and so, and then, and in fact or but anyway, where they combine. DM co-occurrence is a multi-faceted phenomenon, since not all cases display the same degree of integration: most authors distinguish between at least two types of co-occurrence depending on a number of syntactic and functional criteria (see, e.g., Luscher 1993; Hansen 1998; Pons 2008, in press; Cuenca & Marín 2009).

Discourse analysis and corpus annotation show that this phenomenon is quite pervasive: 20% of all occurrences are coded as part of a co-occurring string in Crible's (2017) corpus study of spoken English and French. In fact, DM co-occurrence poses a challenge for corpus annotation since i) it is not always clear whether two co-occurring DMs remain independent from each other or whether they should be considered as one token, and ii) senses can be influenced by co-occurring DMs during disambiguation.

This study sets out to provide clear criteria for different degrees of co-occurrence on the basis of corpus-based examples.

2 State of the Art

Previous papers on the subject propose several criteria to distinguish different degrees of integration. Luscher (1993) uses syntactic and semantic scope to distinguish between "additive" and "compositional" sequences. He defines the latter as applying to two adjacent DMs which are semantically similar (e.g. French *mais pourtant* 'but however'), one of them being more restricted or specific in its meaning than the other. This latter type is the focus of Fraser's (2013) study targeting English contrastive connectives. Hansen's (1998) distinction between summative and combinatory sequences adopts a different perspective and depends on whether the elements in the sequence retain their individual meaning (French *ah bon* 'oh really') or form a new complex one (*eh bien* 'well'). She argues that most DM sequences are summative (or compositional), since it is always possible to reconstruct the meaning of each element. Similarly, Pons (2008) concludes from his analysis of the co-occurrences of the Spanish modal marker *bueno* with other discourse markers that discourse segmentation of oral discourse allows to differentiate two different configurations: the cases in which the two markers are simply adjacent from the cases in with they combine according to whether they apply to different or to a unique structural unit. More recently, Dostie (2013) and Crible (2015) consider other types of cues in DM use that provide evidence for stronger degrees of combination, such as phonological reduction (*eh bien* to *eh ben*), new spellings (*ou sinon* 'or else' to *aussi non*) and new contexts of use (initial to final position for *ou sinon*).

Cuenca & Marín (2009) discuss and illustrate a three-fold distinction in a corpus of spoken Spanish and Catalan, namely:

- juxtaposition, when the DMs do not combine syntactically nor semantically (typically two conjunctions);
- addition, when the DMs combine locally but their functions remain distinct (typically conjunctions followed by parenthetical connectives that jointly connect at a local level);
- composition, when the DMs function as one unit (typically two parenthetical connective units with a single global-level function).

Their analysis is very fine-grained and identifies recurrent formal and functional tendencies for each of these levels. Crible (2017) attempted to apply Cuenca & Marín's (2009) classification through systematic annotation and was confronted with problematic, borderline cases (e.g. *and so* or *et alors* 'and then') which raised concerns about some features, pointing especially at the fuzzy border between addition and composition. Crible also discusses the role of frequency in the definition of these levels, and suggests an additional degree to deal with cases of "reinforcement" (e.g. *but in fact*). Her study draws the attention to the consequences of an adequate treatment of DM co-occurrence for corpus annotation (token identification and sense disambiguation). Sim-

ilarly, in the guidelines of the Penn Discourse TreeBank 2.0, Prasad et al. (2007) mention that multiple (i.e. co-occurring) connectives should ideally be annotated as such and differentiated according to the (in)dependence of their elements in order to improve predictive features and classifiers.[1]

3 Method

The purpose of this study is to revisit Cuenca & Marín's (2009) three-fold classification and refine the criteria to distinguish each degree of co-occurrence, in order to be able to apply them systematically to corpus annotation. To this end, we used a sample of English conversational data from the DisFrEn dataset where DMs were already identified (Crible 2017). We considered as one DM multi-word DMs (e.g. *so that*, *even if*, *I mean*) and excluded cases where two DMs belong to different units (final position of the first unit, initial position of the second one, as in *I like winter actually but I prefer spring*) or are repeated due to performance effects.

For each cluster, we manually encoded the following features: number of elements in the cluster, syntactic category of each DM (based on Cuenca 2013: conjunction, parenthetical connective, pragmatic connective, interjection), scope (same or different), position (utterance initial or medial). We then discussed whether the elements of the cluster expressed the same meaning (or function) or not, and then decided on the degree of integration of the adjacent DMs.

4 Results

The qualitative analysis led to distinguish between criteria (necessary conditions) and features (quantitative tendencies): we found that considerations of scope and of function are criterial in the definition of the levels, whereas prosody (i.e. contiguous pause) and syntactic categories are mere tendencies. As a result, the revised cline of co-occurrence proposed is the following:

Juxtaposition, when the DMs take scope over different units (mostly when two or more conjunctions co-occur);

(1) he said he seemed quite quite happy to meet you (0.320) I'm I'll attempt not to turn this off // well I mean it's no problem [*because* [*if* he doesn't turn up if he doesn't turn up] I'll just uhm (0.020) you know go and get some sandwiches or something]

Combination, when the DMs have the same scope and their functions mix. Combination can lead to addition or to composition of markers:

Addition, when the DMs have the same scope but distinct compatible meanings that add so that the second DMs narrows down or reinforces the meaning of the first DM;

[1] The PDTB 2.0 distinguishes between "multiple" and "conjoined" connectives", the latter referring to a very restricted number of uses such as *if and when*, which are annotated as one item.

(2) was there a sister there // well uh he's got a (0.330) I don't know whether yeah I suppose so but I heard that // Emma (0.800) Josephine // Josephine I don't know (0.490) Uhm *but actually* he's got (0.920) he's got somebody living in his house

Composition, when the DMs have the same scope and jointly express one single meaning.

(3) the funny thing is that none of the sort of Nancy Mitford stuff (0.050) do I mean Nancy (0.020) I can never remember which Mitford is which *but anyway* none of the u and non-u stuff seems to have washed off on your mother at all

Lexicalization, when a new meaning arises from the co-occurrence which is not the sum of its parts and the instruction encoded by the cluster becomes conventional.

This proposal takes into account the dynamicity of language and phenomena such as layering and stratification, related to polyfunctionality and underspecification. For instance, in English the highly frequent cluster *and then* instantiates different configurations and degrees of integration depending on the semantic status of the temporal adverbial. The first (and most frequent) use of *and then* (1) is an addition of the additive conjunction and the temporal adverb. In another related use (2), the elements add to express consequence, a meaning which can be derived – but differs – from the temporal meaning of *then* ('at that time'). Lastly, *and then* (3) can express one global function of continuity or enumeration at discourse level (i.e. not temporality between facts) with contrastive nuances, in which case the co-occurrence is somewhere in-between the space of composition and lexicalization, since the meaning of the cluster is not (strictly) the sum of its parts.

(4) they buy the book say for a couple of pounds (1.420) *and then* return it and get half

(5) I've got people coming I'll get some salmon from the stall and when you get down there you find he hasn't actually got any *and then* it throws you into a complete quandary

(6) people do tend to describe themselves […] a lot of people describe people as jeal-ous […] *and then* there are the really bland ones

It can be concluded that a single co-occurrence *and then* can instantiate different categorical configurations and can also vary along the cline of co-occurrence, thus advocating for a flexible, context-bound approach to the issue in future annotation endeavours.

5 Discussion

These distinctions are subtle and highly context-bound, yet they can and should be systematically accounted for, especially since *and then* is also quite frequent in writing (cf. *but then* or *so for instance*, mentioned in the PDTB guidelines). Additional features

(e.g. prosody, length and type of host unit) can be investigated to further support this flexible portrait of and then.

To conclude, in line with Crible & Cuenca (2017), we suggest that DM annotation endeavours should consider including information about co-occurrence, minimally by identifying clusters, ideally by distinguishing between degrees of integration following the criteria that we have developed in this study. This is particularly crucial for sequences such as *and then* (and its cross-linguistic equivalents, e.g. French *et puis*), which do not display a unique functional profile depending on co-occurrence degree. Our criteria and analysis pave the way for fruitful comparisons across languages and also across spoken and written registers.

References

1. Crible, L. 2017. *Discourse Markers and (Dis)fluency across Registers: A Contrastive Usage-Based Study in English and French*. Doctoral thesis, Université catholique de Louvain.
2. Crible, L. 2015. Grammaticalisation du marqueur discursif complexe *ou sinon* dans le corpus de SMS belge : spécificités sémantiques, graphiques et diatopiques. *Le Discours et la Langue* 7(1): 181-200.
3. Crible, L. & Cuenca, M. J. 2017. Discourse markers in speech: Characteristics and challenges for corpus annotation. *Dialogue and Discourse* 8(2): 149-166.
4. Cuenca, M. J. 2013. The fuzzy boundaries between modal and discourse marking. In L. Degand, B. Cornillie & P. Pietrandrea (eds), *Discourse Markers and Modal Particles: Description and Categorization*. Amsterdam, John Benjamins: 191-216.
5. Cuenca, M. J. & Marín, M. J. 2009. Co-occurrence of discourse markers in Catalan and Spanish oral narrative. *Journal of Pragmatics* 41: 899-914.
6. Dostie, G. 2013. Les associations de marqueurs discursifs - De la cooccurrence libre à la collocation. *Linguistik Online* 62(5).
7. Fraser, B. 2013. Combinations of contrastive discourse markers in English. *International Review of Pragmatics* 5: 318-340.
8. Hansen, M.-B. M. 1998. *The Function of Discourse Particles. A study with special reference to spoken standard French*. Amsterdam: John Benjamins.
9. Luscher, J.-M. 1993. La marque de connexion complexe. *Cahiers de Linguistique Française* 14: 173-188.
10. Pons, S. 2008. La combinación de marcadores del discurso en la conversación coloquial : interacciones entre posición y función. *Estudos Linguísticos/Linguistic Studies*, 2. Lisboa: Edições Colibri/CLUNL: 141-159.
11. Pons, S. In press. The combination of discourse markers in spontaneous conversations: keys to undo a gordian knot. *Revue Romane*.
12. Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A. & Joshi, A. 2007. The Penn Discourse Treebank 2.0 Annotation Manual. *IRCS Technical Reports Series*, University of Pennsylvania, Institute for Research in Cognitive Science.

Disambiguating discourse relations with or without a connective: Does “and” really say nothing?

Ludivine Crible[1] and Vera Demberg[2]

[1] Université catholique de Louvain, Belgium
[2] Universität des Saarlandes, Germany
ludivine.crible@uclouvain.be

Abstract. This paper reports on the results of two crowdsourcing experiments, where pairs of sentences originally related by the additive connective “and” were disambiguated by naïve participants. In these connective insertion tasks, the original connective is visible in the first condition, whereas it has been removed in the second condition. The participants have to select the connective that best describes the meaning of the relation (either by substituting “and” or by filling the blank), from a list of options containing, e.g. “but”, “so”, or “however”. Our hypothesis is that the removal of “and” will lead to changes in the naïve participants’ disambiguations, thus showing that “and” provides some instructions for relation interpretation, in spite of its small informative value. Our results have implications not only for the methodology of connective annotation but also for theoretical considerations of polyfunctionality and semantic underdetermination.

Keywords: disambiguation, additive connective, implicitation.

1 Introduction

Coherent texts, whether written or spoken, are built upon discourse relations linking utterances together through causal, temporal or contrastive connections, among many other types (Mann & Thompson 1988). Different types of relations are signalled by different types of markers, although there is no one-to-one mapping. These markers often belong to the functional category of discourse-relational devices, or “connectives”, such as however, because or in fact. Writers and speakers also have the option to use other signalling devices (e.g. lexical or syntactic patterns) or even to leave a discourse relation implicit (e.g. Taboada 2009). This study focuses on another strategy for discourse marking, namely the use of underspecified connectives (Spooren 1997). More particularly, we investigate the role of the additive conjunction *and* to signal relations of addition but also of consequence, contrast and concession. In these cases, the discourse relation is more specific than the information strictly provided by the connective: a consequence or contrast is more informative than a mere additive relation. Despite the low informative value of *and*, it is quite often found in authentic contexts where such enriched interpretations were assigned to the discourse relation (6% of and express a result in the Penn Discourse TreeBank 2.0, Prasad et al. 2008),

which calls for more research on the conditions under which and can be used as an underspecified connective.

2 Crowdsourcing disambiguations

Our research objective is thus to compare the linguistic and contextual features of utterances linked by *and* which either express addition, consequence, contrast or concession. To do so, we first need corpus-based data where such discourse relations are reliably identified. Discourse relation annotation is extremely costly in time and human resources, it requires heavy training and, even so, agreement scores are often rather low (Spooren & Degand 2010). As a result, researchers have recently started to turn to crowdsourcing as an alternative method to gather discourse relation disambiguations through a low-cost, non-expert workforce (Kawahara et al. 2014; Rohde et al. 2016). Scholman & Demberg (2017) report on the results of a connective insertion task which they used as an indirect method to annotate discourse relations: the sense of a relation can be retrieved through the selection of unambiguous connectives from a list to fill in a blank between utterances, provided this task is repeated by a large number of participants (around 20). The authors discuss the validity of this method and conclude that crowdsourcing connective insertions is reliable enough as an alternative to expert annotations.

3 Connective insertion tasks : hypotheses

For connective insertion tasks, it's important to distinguish between originally implicit vs. explicit relations. For originally implicit relations, inserting a connective resembles the approach taken in PDTB annotation (except that the choice of connectives is more restricted in the crowdsourcing step in order to allow for disambiguation of relation type). In originally explicit relations, however, the meaning and interpretation may substantially change by removing the connective, see examples (1)-(3) below. In this study, where we investigate originally explicitly marked relations with the connective "and", we will therefore compare a crowdsourced connective insertion task with a crowdsourced connective replacement task, in which the original connective "and" is not removed from the stimulus.

(1) I am not going back to Germany. Therefore, I will not eat spätzle ever again. (consequence)

(2) I am not going back to Germany. In fact, I will not eat spätzle ever again. (addition)

(3) I am not going back to Germany. I will not eat spätzle ever again. (?cause)

Our working hypothesis is that such differences in interpretations, triggered by the connective (or absence thereof), apply to utterances containing *and* as well. In this respect, we challenge previous experimental research on *and* which showed that *and* has a very small informative value and little or no facilitating effect on reading times (Murray 1994) or comprehension (Cain & Nash 2011). By contrast, we expect that

connective insertion tasks will be affected by the presence of *and*, thus supporting the claim that *and* does trigger enriched pragmatic inferences (Blakemore & Carston 1999). We therefore propose to use crowdsourcing for the study of underspecified *and* by comparing stimuli with and without the original connective. In other words, we want to test whether connective elicitations will differ across stimuli which are identical except for the presence or absence of the conjunction and.

4 Method

To this end, we ran two crowdsourcing experiments on the Prolific Academic online platform. In the first one, we used 83 authentic pairs of utterances originally containing and. We collected them from the Loyola Corpus of Computer-Mediated Communication (Goldstein-Stewart et al. 2008), in order to avoid the high formality of existing corpora such as the Penn Discourse Treebank (economy newspaper articles). This corpus contains blogs and chat conversations between college students about topics such as gay marriage, gender discrimination or privacy rights. This data was pre-annotated by the first author as either expressing a relation of addition, contrast, concession or consequence. The stimuli are grouped in four lists of about 20 items each, which are balanced with respect to the pre-annotated relation type (about 10 addition, 6 consequence, 1 contrast, 3 concession in each list). The participants (paid 1€ per list) can choose from a list of eight connectives to fill in a blank between the two utterances (the original and has been removed). The connectives are in addition, plus, therefore, as a result, by contrast, whereas, nevertheless and yet.

The second experiment uses exactly the same lists of items, except that the stimuli now show the original *and* connecting the utterances, and the participants are therefore instructed to substitute this *and* with one of the connectives from the same list of options.

In the analysis, we first compare the connectives chosen by the participants with the relation type pre-identified by the expert annotator, in order to see whether they converge (e.g. therefore or as a result selected in case of a relation of consequence). This first step provides us with a dataset of utterance pairs with their disambiguated discourse relation, without resorting to costly (and partly subjective) expert annotations. The items thus classified into one of the four categories (addition, consequence, contrast, concession) will allow us to test the effect of additional variables (e.g. register) in further studies (Crible & Demberg 2017). We replicated this analysis with the results of the second experiment. We then compared whether the connectives selected by the participants are the same when *and* is present in the stimuli and when it is not.

5 Results and discussion

Preliminary results show that, when the relation was pre-annotated as additive, the participants tend to equally choose a consequence or an additive connective, with no significant difference, which suggests that consequence is often interpreted in the absence of a connective. For all other relation types (i.e. consequence, concessive and

contrast), the great majority of participants' choices match the pre-annotation, thus confirming that *and* can be used in contexts which express more than mere addition. The results from the second experiment are still pending, and should lead to interesting comparisons on the effect of *and* in connective elicitation (or connective substitution).

This study has a number of implications on the informative value of *and*, which may be higher than what previous studies have suggested, and on the use of connective elicitation with or without the original connective included in the stimuli. We argue that, when dealing with authentic corpus-based stimuli, including the original connective in the experiment is a more accurate representation of the data and better reproduces the interpretation mechanisms as they would be processed in natural conditions.

The experiments reported in this paper constitute the first step of a larger project on the contextual and cognitive constraints to the production and interpretation of underspecified connectives. They also relate to ongoing crosslinguistic projects on the meaning variation of *and* and its use across spoken registers (Crible, in press) and in translation (Abuzcki et al. 2017).

References

1. Abuzcki, A., Burkšaitienė, N., Crible, L., Furkó, P., Nedoluzhko, A., Oleškevičienė**,** G. V. & Zikánová, Š. 2017. "The underspecified connective *and* in a parallel TedTalk corpus: functions, translation and implicitation". Paper accepted at the DiscourseNet conference, May 17-19, Budapest, Hungary.
2. Blakemore, D. & Carston, R. 1999. The pragmatics of *and*-conjunctions: The non-narrative cases. *UCL Working Papers in Linguistics* 11.
3. Cain, K. & Nash, H. M. 2011. The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology* 103(2): 429-441.
4. Crible, L. (in press). Emplois sous-spécifiés des marqueurs discursifs *et / and* à l'oral : stratégie (inter)subjective et variation en genre. *Cahiers du FoReLL*.
5. Crible, L. & Demberg, V. 2017. The effect of genre variation on the production and acceptability of underspecified discourse markers in English. Paper accepted at the DiscourseNet conference, May 17-19, Budapest, Hungary.
6. Goldstein-Stewart, J. Goodwin, K. A., Sabin, R. E. & Winder R. K. 2008. Creating and using a correlated corpora to glean communicative commonalities. In *LREC2008 Proceedings*, Marrakech, Morocco.
7. Kawahara D., Machida Y., Shibata T., Kurohashi S., Kobayashi H. & Sassano M. 2014. Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proceedings of the International Conference on Computational Linguistics (COLING)*: 269–278.
8. Mann, W. & Thompson, S. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3): 243-281.
9. Murray, J. D. (1994). Logical connectives and local coherence. In R. F. Lorch & E. 1. O'Brien (Eds.), *Sources of Cohesion in Text Comprehension*. Hillsdale, NJ, Erlbaum: 107-125.
10. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. & Webber, B. 2008. The Penn Discourse TreeBank 2.0. *Proceedings of LREC, June 2008*: 2961-2968.

11. Rohde H., Dickinson A., Schneider N., Clark C., Louis A. & Webber B. 2016. Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the Linguistic Annotation Workshop (LAW X)*: 49-58.
12. Scholman, M. & Demberg, V. 2017. Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion tas. In *Proceedings of the 11th Linguistic Annotation Workshop*: 24-33.
13. Spooren, W. 1997. The processing of underspecified coherence relations. *Discourse Processes* 24(1): 149-168.
14. Spooren, W. & Degand, L. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory* 6(2): 241-266.
15. Taboada, M. 2009. Implicit and explicit coherence relations. In J. Renkema (Ed.), *Discourse, of Course*, Amsterdam, John Benjamins: 127-140.

Designing a corpus-based lexicon for spoken DRDs: semantic considerations

Ludivine Crible[1] and Amalia Mendes[2]

[1] Université catholique de Louvain, Belgium
[2] Universidade de Lisboa, Centre of Linguistics, Portugal
ludivine.crible@uclouvain.be

Abstract. This paper discusses issues related to the design of a lexicon for spoken DRDs on the basis of an annotated English-French corpus, where DRDs have been functionally disambiguated. While most existing lexicons of DRDs use taxonomies designed for written data, our proposal is based on a model which includes typical discourse relations but also functions that DRDs can express in spoken language, such as topic-shift, turn-taking or repair. We focus on semantic issues of polyfunctionality, of which we distinguish four types: polysemy (several related meanings), multifunctionality (several simultaneous meanings), underspecification (contextual enrichment) and multidimensionality (one meaning across several functional dimensions or domains). We discuss how each of these cases can be structured in a lexicon, and conclude on the limitations of corpus annotations as direct input for building a lexicon.

Keywords: lexicon, speech, polyfunctionality.

1 Introduction

Natural language, either spoken or written, is built upon relations of coherence amongst linguistic units of various types (Mann & Thompson 1988). These relations are often signalled by the functional class of discourse-relational devices (henceforth DRDs), also called "connectives" (e.g. van Dijk 1979) or "discourse markers" (e.g. Schiffrin 1987). DRDs are typically short and fixed expressions with a (primarily) procedural meaning whose function is to constrain the interpretation of their host unit and its relation to the context (Blakemore 2002; Crible 2017a). Most authors agree on a common core including conjunctions (*and*, *but*, *although*, etc.) and adverbials (*so*, *however*, *in fact*, *on the other hand*, etc.); other categories, such as verb phrases (e.g. *I mean*), interjections (e.g. *oh*), alternative lexicalizations (e.g. *It results that*) or even syntactic forms (e.g. gerund) can be included, depending on the definition (Fischer 2006).

DRDs are very varied in forms and functions and are not necessarily used in the same way across different languages. As a result, they can be particularly challenging to acquire (Evers-Vermeul & Sanders 2009) and to translate (Meyer et al. 2012). DRD lexicons are particularly useful in this respect: they provide a machine-readable resource that can be consulted or automatically implemented for a variety of applications. In this paper, we introduce the design of a new lexicon targeting DRDs as they

were annotated in a corpus of spoken English and French, viz. the *DisFrEn* dataset (Crible 2017b). We discuss in particular the semantic labels that should be used to describe DRD entries. It is argued that the polyfunctionality of some DRDs should be carefully encoded in the lexicon through a semantic framework where notions of ambiguity and polysemy are distinguished.

2 Existing DRD lexicons

Most lexicons of DRDs focus on written data and are created either by automatically extracting the information from annotated discourse banks, or by manually inspecting written texts (Roze et al. 2012; Mendes & Lejeune 2016; Stede 2002; Scheffler & Stede 2016, Feltracco et al., 2016). These lexicons take DRDs as expressing a two-place semantic relation that involves propositional arguments. An exception is the Diccionario de partículas discursivas del español - DPDE (Briz et al., 2003) in that it also includes information extracted from spoken data. As a result, it goes beyond the function of connection between two segments and also covers modal and interactional meanings (functions of modalisation and control of contact). Different typologies may be used to label the semantic relations of the DRDs: LEXCONN follows the SDRT set of relations, while DIMLex and LDM-PT use the PDTB 3.0 sense hierarchy, and the DPDE writes a lexicographic definition. Different solutions are used to encode the polyfunctionality of the DRDs: either a list of senses in a POS entry (DIMLex) or individual entries of form/meaning pairs (LEXCONN, LDM-PT). The lexicographic nature of the DPDE makes it possible to distinguish between distinct meanings, that are treated as homonyms, and contextual senses of a basic meaning, that are listed in the field "other uses".

3 The DisFrEn dataset

Our proposal of a lexicon for spoken DRDs is based on the annotations of the DisFrEn dataset. DisFrEn contains about 160,000 words (15 hours) of native English and French distributed across eight settings of spoken interaction (e.g. conversation, interview, political speech). In this corpus, a comprehensive, bottom-up selection of DRDs have been manually identified according to three major criteria: syntactic optionality, formal fixedness and procedural meaning. More than a hundred types of DRDs have thus been identified, such as and, so, because, actually, you know, well, for example, among many others. The selected items have then been disambiguated following a taxonomy of thirty senses (e.g. cause, concession, reformulation, topic-shift) and four domains of use (viz. ideational, rhetorical, sequential, interpersonal). This fine-grained annotation was carried out by one expert annotator, with reliable intra-annotator agreement ($\kappa = 0.779$).

The senses in *DisFrEn* were annotated in order to reflect the meaning-in-context or function of the items, not only based on the semantic information provided by the DRD but also by contextually enriched interpretations. More particularly, functions related to the management of speech turns, topics of speaker-hearer relationships are

annotated alongside more traditional senses for discourse relations. The methodological approach also allows for double labels in the case of simultaneous functions (e.g. consequence + topic-shift). In other words, the annotation does not only capture what the DRDs "mean" but also what they "do", following the assumption that (spoken) discourse is fundamentally multifunctional (Bunt 2011). As a result, the number and types of semantic labels assigned to one token can be very high, and not all labels are equally encoded in the semantics of the DRD: for instance, the 429 occurrences of *so* are distributed across 19 different (combinations of) labels, such as conclusion, specification, topic-shift, reformulation or topic-resuming. By contrast, the same DRD *so* is only given two labels (result, reason) in the PDTB 2.0 (Prasad et al. 2008). This difference is not only due to the data type (spoken vs. written) but also to the coverage of the taxonomy (cf. Crible & Cuenca 2017).

4 Dealing with polyfunctionality in the lexicon

Such rich information cannot be directly implemented in the lexicon and needs to be filtered, or at least structured. Too many semantic labels would be impractical for the various purposes of the lexicon: a language learner would not be able to know which contexts of use are typical or atypical; a highly polyfunctional DRD such as and would be a potential translation for virtually any other one in machine translation. In the applied perspective of building a lexicon, we argue that it is important to distinguish between different types of polyfunctionality. In particular, multiple senses for a single DRD can either relate to polysemy, multifunctionality, underspecification or multidimensionality (Crible 2017c).

A DRD is polysemous when it encodes more than one (related but distinct) meanings (e.g. but expresses both contrast and concession). In this case, the lexicon should reflect all of these meanings.

A DRD is multifunctional when, in a given context, it expresses two or more functions at the same time (e.g. temporal and consequence relation). Annotation instructions often specify how many different senses can be assigned simultaneously (only one; up to two in DisFrEn; up to two in the PDTB in theory, but the option is rarely used, Scholman & Demberg 2017). Double labels are not a practical option for lexicons, which thus requires to either split them in two, or to choose the more prominent sense, if any.

A DRD is underspecified when it expresses a meaning that is richer or more informative than its basic meaning (e.g. and in a consequence or concessive relation). Underspecification mainly concerns and: this basic conjunction only encodes addition but can be used in contexts where enriched interpretations of, e.g., consequence or contrast can be construed from the context. It may be considered that these additional senses are not part of the semantic spectrum of and and should therefore not be included in the lexicon. Another position would be to include these uses, so as not to lose any information in the possible uses of and, but to distinguish them from the core meaning in the structure of the lexicon.

Lastly, a DRD is multidimensional if one or several of its senses can be expressed in more than one dimension or "domain". This latter notion is inspired by Crible & Degand's (in press) annotation model, where the number of senses from Crible's (2017a) taxonomy is reduced to 11 and where functions and domains are independent. With this model, the meaning variation of DRDs is reduced to one core meaning (or several, in the case of polysemy), which can then instantiate one or several domains: for instance, so mainly expresses consequence, and this consequence can relate facts (ideational consequence), conclusions (rhetorical consequence) or topics (sequential consequence). Providing this dual information of domains and functions in the structure of the lexicon would help dealing with some cases of multifunctionality, all the while maintaining a large coverage of the functional spectrum of DRDs in spoken language.

5 Conclusion

In the presentation, we will discuss how these different types of polyfunctionality relate to the existing annotations in the DisFrEn dataset and how they can be formalized in the lexicon. It will become apparent that such an annotated resource may not be directly usable to build a lexicon (e.g. need to merge sense labels or redefine the relation between domains and functions) and that a semantic framework is needed to structure and describe the meaning variation of DRDs in speech and in general. Building a corpus-based lexicon is a complex process (not only because of semantic considerations), especially if the input corpus was not designed for this specific application. Our paper thus stresses the importance of the purpose and research question behind any annotation endeavor.

References

1. Blakemore, D. 2002. *Relevance and Linguistic Meaning. The Semantics and Pragmatics of Discourse Markers*. Cambridge : CUP.
2. Bunt, H. 2011. Multifunctionality in dialogue. *Computer Speech and Language* 25: 222-245.
3. Crible, L. 2017a. Towards an operational category of discourse markers: A definition and its model. In C. Fedriani & A. Sansó (eds), *Discourse markers, Pragmatics Markers and Modal Particles: New Perspectives*: 101-126. Amsterdam: John Benjamins.
4. Crible, L. 2017b. Discourse markers and (dis)fluencies in English and French: Variation and combination in the DisFrEn corpus. *International Journal of Corpus Linguistics* 22(2): 242-269.
5. Crible, L. 2017c. "Ambiguity, multifunctionality and underspecification: coming to terms with French *et, mais, donc*". Paper presented at the *5th International Conference on Discourse Markers in the Romance Languages (DisRom 2017)*, November 8-10, Louvain-la-Neuve, Belgium.
6. Crible, L. & Cuenca, M. 2017. Discourse markers in speech: Characteristics and challenges for corpus annotation. *Dialogue and Discourse* 8(2): 149-166.

7. Crible, L. & Degand, L. In press. Reliability vs. granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory*.
8. Dijk, T. A. van. 1979. Pragmatic connectives. *Journal of Pragmatics* 3: 447-456.
9. 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: 829–854.
10. Feltracco, A., Jezek, E., Magnini, B. and Stede, M. (2016) LICO - A Lexicon of Italian Connectives. *Proceedings of the Third Italian Conference on Computational Linguistics* (CLiC-it 2016), 2016, Napoli, December 5-7, 2016.
11. Fischer, K. 2006. Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume. In K. Fischer (Ed.), *Approaches to Discourse Particles*: 1-20. Amsterdam: Elsevier.
12. Mann, W. & Thompson, S. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3): 243-281.
13. Mendes, A. & Lejeune, P. 2016. LDM-PT. A Portuguese Lexicon of Discourse Markers. In L. Degand, C. Dér, P. Furkó & B. Webber (eds.), *Conference Handbook of TextLink – Structuring Discourse in Multilingual Europe Second Action Conference*, Budapest, 11-14 April 2016: 89-92.
14. Meyer, T., Popescu-Belis, A., Hajlaoui, N. & Gesmundo, A. 2012. Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*.
15. 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 08)*, Marrackech, Morroco: 2961-2968.
16. Roze, C., Danlos, L. & Muller, P. 2012. LexConn: a French lexicon of discourse connectives. *Discours* 10.
17. Scheffler, T. & Stede, M. 2016. Adding semantic relations to a large-coverage connective lexicon of German. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroz, Slovenia.
18. Schiffrin, D. 1987. *Discourse Markers*. Cambridge: CUP.
19. Scholman, M. & Demberg, V. 2017. Examples and specifications that prove a point: Identifying elaborative and argumentative discourse relations. *Dialogue and Discourse* 8(2): 56-83.
20. 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.

Functions and domains of discourse markers across languages: Testing a two-dimensional annotation scheme

Liesbeth Degand[1], Ludivine Crible[1] and Karolina Grzech[2]

[1] Université catholique de Louvain
[2] SOAS University London
liesbeth.degand@uclouvain.be

Abstract. We propose a corpus-based annotation scheme for discourse markers in spoken language use aiming to cover their wide spectrum of uses through the combination of four discourse domains (*ideational, rhetorical, sequential, interpersonal*) with eleven discourse functions (*contrast, cause, specification, punctuation, topic, …*). We evaluate the multilingual validity of the annotation scheme on English, French and Polish, and we discuss its advantages and disadvantages in comparison to alternative annotation schemes.

Keywords: Discourse Markers, Discourse Annotation, Spoken language.

1 Towards a multilingual annotation scheme

Discourse markers are the focus of an abundant research field investigating the many aspects of their behavior, either from a syntactic, semantic, prosodic or other approach. One crucial aspect is their polyfunctionality, which has been explained and modeled under several different theoretical frameworks (see Fischer 2006 for an overview). These models include, among others, the notion of multidimensionality in the Dynamic Interpretation Theory (Petukhova & Bunt 2009), the five "planes of talk" in Schiffrin (1987), the concept of "meaning potentials" (Norén & Linell 2006; Aijmer 2013), the constructionist approach by Fischer (2010, 2015) or the three components of discourse structure in Redeker (1990) (see also González 2005). Each approach provides a different (yet partially overlapping) account of the many dimensions of meaning that discourse markers can express, following different theories and agendas (e.g. discourse analysis, cognitive linguistics, computational applications).

Combining theoretical and methodological considerations, we propose a corpus-based annotation scheme for discourse markers in spoken language use, where their functional spectrum is seen as the interface between two independent dimensions, namely a domain and a function (Crible & Degand in press). Our four domains (viz. ideational, rhetorical, sequential and interpersonal) are rooted in the tradition of cognitive models of discourse structure (e.g. Redeker 1990; Sweetser 1990; Sanders 1997) and correspond to different layers of discourse which speakers can address: content relations (ideational), subjective and metalinguistic meanings (rhetorical), discourse structure (sequential) and speaker-hearer relationship (interpersonal). Functions, on the

other hand, are more specific interpretations of the type of operation which a discourse marker is performing in a given context (eleven types, e.g. causal relation, topic-shift, specification, etc.).

Following Bunt (2011), we distinguish between "general-purpose" functions, which can activate any of the four domains (e.g. a relation of contrast can be either ideational, rhetorical, sequential or interpersonal), and "dimension-specific", here, domain-specific functions, which pertain to one domain only (e.g. topic-shift is always sequential). We consider domains and functions as two orthogonal dimensions of meaning which each correspond to a type of semantic variation, viz. polyfunctionality and polysemy, respectively. By polyfunctionality, we mean the possibility for a single invariant meaning to be expressed across several domains (e.g. ideational vs. rhetorical contrast). Polysemy, in turn, refers to the multiple functions a discourse marker can fulfil, regardless of the domain (e.g. so to express a consequence, an exemplification or a topic shift). In this sense, polysemy is different from "simultaneous multifunctionality" (Bunt 2011), which rather targets the joint expression of more than one meaning at a time in a given context, be it different functions (polysemy) or the same function in different domains (polyfunctionality). Our integrated approach is compatible with an inclusive definition of the discourse marker category as adopted and annotated by Crible (2017), especially since issues of categorization and functional classification are strongly interrelated (cf. Degand et al. 2013).

The challenge of any discourse annotation scheme is to be reliable and valid. Reliability of the annotation rests on the assumption that it should be optimally objective and replicable so that other researchers would be able to reproduce them. Validity, then, aims at optimal coverage of the linguistic phenomenon under scrutiny. In our case, the aim is that the discourse markers can be validly described in their full functional spectrum and be distinguished from one another, within and between languages. In other words, we aim at providing an annotation scheme that can be cross-linguistically applied.

1.1 Contrasting discourse markers in English, French and Polish

The aim of this presentation is to test the assumed cross-linguistic validity of the approach and the ensuing annotation scheme by analyzing the variation in use and functions of a broad bottom-up selection of DMs across three languages from different typological families, namely French (Romance), English (Germanic) and Polish (Slavic). The taxonomy was applied to a sample of ca. 30 minutes (between 5000 and 6000 words) of spoken unplanned dialogues in each of the languages making use of available corpora (LOCAS-F corpus (Degand et al. 2014) for French, ICE-GB corpus (Nelson et al. 2002) for English, and Polish data (Pęzik 2015). Discourse markers were identified in a bottom-up approach, without any closed list of pre-selected items, following a broad definition: any expression, which is syntactically optional, has a (partly) procedural meaning and performs a discourse-level pragmatic function was selected. The items were then manually annotated according to the functional classification introduced above. The French data was double-coded for the purpose of inter-annotator

agreement. Annotations were then extracted for contrastive analyses of distribution and variation of DMs and their functions.

The data and annotation scheme will be briefly presented, focusing on challenges of multilingual annotation. In this discussion, we will include the results of an ongoing annotation campaign on additional languages (Slovenian, Spanish, L2 English). We will then report on quantitative findings of the distribution of domains and functions across French, English and Polish conversations, looking for cross-linguistic differences and similarities in the functions DMs can express in each language, both from a categorical and a DM-specific point of view.

2 First results

Our results show that the interpersonal domain is much more frequent in Polish than in the other two languages: this can be explained by the different nature and status of question tags (English *isn't it*, Polish *nie*, French *hein*) in the three systems. This major result led us to further investigate interpersonal DMs. Each language has a different number of typically interpersonal DMs, which display a different frequency in the data: for instance, the typical *tu vois* in French is highly infrequent in the sample. There are many more different types of interpersonal DMs in Polish, which suggests that these items are at the core of the DM category. In addition, interpersonal meanings can be expressed by other DMs as well: we identified a cline from "purely interpersonal" to "sometimes interpersonal" DMs: the latter is not attested in the English data. Interpersonal uses of otherwise adversative markers in French (*mais*) and Polish (*przecież*) will be discussed. This analysis shows that semantic equivalence of DMs attested in different languages does not necessarily lead to functional and distributional similarities between them.

Overall, we observed a higher similarity between English and French than with Polish. However, we cannot conclude at this stage whether this observation is due to family resemblance or to annotators' bias. Regarding the two-dimensional taxonomy, Crible & Degand's (in press) revised version seems to reach higher inter-annotator agreement (compared to Crible's (2017) original). It is applicable to a large range of languages from different typological families, and enables interesting analyses, both at a comprehensive level over the whole DM category and at a more DM-specific level.

References

1. Aijmer, K. 2013. Understanding Pragmatic Markers. A Variational Pragmatic Approach. Edinburgh: Edinburgh University Press.
2. Bunt, H. 2011. Multifunctionality in dialogue. *Computer Speech and Language* 25: 222-245.
3. Crible, L. 2017. Towards an operational category of discourse markers: A definition and its model. In C. Fedriani & A. Sanso (eds), *Discourse Markers, Pragmatics*

Markers and Modal Particles: New Perspectives, Amsterdam, John Benjamins: 101-126.
4. Crible, L. & Degand, L. In press. Reliability vs. granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory*.
5. Degand, L., Cornillie, B. & Pietrandrea, P. (eds). 2013. *Discourse Markers and Modal Particles. Categorization and Description*. Amsterdam: John Benjamins.
6. Degand, L., Martin, L.J. & Simon, A.-C. 2014. Unités discursives de base et leur périphérie gauche dans LOCAS-F, un corpus oral multigenres annoté. In *Proceedings of CMLF 2014 – 4ème Congrès Mondial de Linguistique Française 2014, Berlin, Germany: EDP Sciences*.
7. Fischer, K. 2006. Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume. In K. Fischer (Ed.), *Approaches to Discourse Particles*, Amsterdam, Elsevier: 1-20.
8. Fischer, K. 2010. Beyond the sentence. Constructions, frames and spoken interaction. *Constructions and Frames* 2(2): 185-207.
9. Fischer, K. 2015. Conversation, Construction Grammar, and cognition. *Language and Cognition* 7(4): 563-588.
10. González, M. 2005. Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse Studies* 77(1): 53-86.
11. Krzeszowski, T.P. 1981. Tertium Comparationis. In J. Fisiak (Ed.), *Linguistics: Prospects and Problems*, Berlin, Mouton de Gruyter: 301-312.
12. Nelson, G., Wallis, S. & Aarts, B. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.
13. Norén, K. & Linell, P. 2006. Meaning potentials and the interaction between lexis and contexts: An empirical substantiation. *Pragmatics* 17(3): 387-416.
14. Pęzik, Piotr. 2015. "Spokes – a Search and Exploration Service for Conversational Corpus Data." In Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands, 99–109. Linköping Electronic Conference Proceedings. Linköping University Electronic Press, Linköpings Universitet.
15. Petukhova, V. & Bunt, H. 2009. Towards a multidimensional semantics of discourse markers in spoken dialogue. In *Proceedings of the 8th International Conference on Computational Semantics*: 157-168.
16. Redeker, Gisela. 1990. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics* 14: 367-81.
17. Sanders, Ted. 1997. Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes* 24: 119-47.
18. Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.
19. Sweetser, Eve. 1990. *From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge: Cambridge University Press.

Naïve annotations of French *et* and *alors*: comparison with experts and effect of implicitation

Ivana Didirková, George Christodoulides, Ludivine Crible and Anne Catherine Simon

Université Catholique de Louvain, Belgium
anne-catherine.simon@uclouvain.be

Abstract. We present the results of an experiment where naïve participants were asked to annotate discourse relations in sequences S1 – discourse marker *alors* / *et* (French so / and) – S2. The experiment was divided in two conditions in order to verify the influence of absence / presence of the discourse marker on perceived discourse relation. 176 sequences were annotated by 44 participants in each condition. Participants had to choose one of four pre-defined discourse relations in a forced-choice task. Responses are analysed in terms of inter-annotator agreement, comparison with expert annotators and implicitation. For both DMs, results show that removing the DM leads to a decrease of identification identification of the discourse relation at stake. Furthermore, analyses of agreements between the naïve and expert annotations raise the question of the importance of using naïve annotations in discourse studies.

Keywords: Discourse Markers, Discourse Relations, Naïve Annotations, Implicitation.

1 Introduction

The process of building a mental representation of discourse in real time largely depends on the identification and interpretation of discourse relations that link utterances. Such discourse relations include addition, causality or temporality ([1]) and are often signalled by so-called discourse markers (henceforth DMs). DMs can be defined as "sequentially dependent elements which bracket units of talk" ([2]), and include expressions such as *et*, *pourtant*, *donc* or *en effet* in French. Discourse markers may be used systematically to signal a specific discourse relation, and thus have a strong core meaning (e.g. French *néanmoins* 'nevertheless' for concession). However, it has been established that a single DM can be used to express several discourse relations, leading to different possible interpretations (e.g. [3, 4]). This is the case for DMs such as French *et* 'and' or *alors* 'then/well', which are notoriously multi-functional and variable in meaning ([5, 6]). This creates challenges for corpus-based studies that aim at disambiguating and annotating the meaning-in-context of DMs, even when such annotation is performed by experts ([7-9]).

1.1 Objectives and hypothesis

In this study we investigate how people without specific training or expertise in linguistics (naïve subjects) identify discourse relations introduced by French *et* and *alors*, as compared to experts' annotations. Our hypothesis is that subjects without any experience in annotating discourse relations will tend to associate each DM with its core meaning rather than its alternative meanings. For example, we expect that subjects will tend to identify a connector such as "and" as inducing an addition rather than a relation of specification. Naïve annotators may better reflect natural discourse processing than careful linguistic annotation (e.g. [10]).

2 Method

The experiment was split into two different conditions in order to test our hypothesis. In the first condition, 44 naïve participants were asked to annotate discourse relations in 176 sequences of utterances, using a multiple-choice procedure. Results from one participant were excluded due to technical issues; thus, further analyses are based on 43 responses. Sequences were divided into four groups, so that every participant had 44 sequences to annotate. Each sequence contained a first segment (S1), a DM (*et* or *alors*) and a second segment (S2). For each DM, four different discourse relations were proposed: two of them were shared by both DMs (consequence and specification); in addition, *alors* could also convey topic shift or concession, while *et* could express addition or temporality. All S1s and some of the S2s were extracted from original spoken data (LOCAS-F corpus, [11]). The original sequences had been annotated by two of the authors who had reached a consensus after discussion of disagreements (see [12] for more details). A S2 was always associated to the S1 and its DM respecting the following procedure: when the original S2 was judged suitable for the purposes of our study (e.g. in terms of length), it was included in the stimuli. Three other S2s were then constructed for the S1, in order to represent additional discourse relations. When the original S2 did not suit the expectations, four new S2s were constructed by the authors for the S1. All the S2s were controlled for syntactic structure.

A document explaining and illustrating the six discourse relations was sent to subjects prior to the annotation. They were then asked to choose one of the four proposed relations for each sequence (forced choice). This choice was made in another document sent to participants, where all the sequences were explicitly split into three parts: S1 – DM – S2. Results were compared to the experts' annotations.

The first condition was then identically replicated a second time by 44 different naïve participants with the notable difference that, this time, the sequences did not contain any DM. Thus, participants were given a document with every sequence split in two parts (S1 and S2) without the original DM. This second condition aims at testing whether originally explicit discourse relations (i.e. signalled by a DM) can be disambiguated without a DM, and what effect this implicitation has on inter-annotator agreement.

Lastly, in a rating task, a different group of 19 naïve participants was asked to judge each sequence for acceptability using a Likert scale from 1 (not acceptable at all) to 5 (perfectly acceptable).

All participants were recruited on Facebook from the Participants' Pool of the Psychology Department of the Université catholique de Louvain, and had no previous experience in annotating discourse relations. They all have declared to be university students. Participants in the annotation experiment took part in the study by way of email exchanges with the first author. Participants in the judgment task were asked to complete an online survey using Lime Survey.

3 Results

3.1 Acceptability task

The results from the acceptability judgment task are presented in the **Fig. 1**. We see that sequences containing the DM *et* are considered as perfectly acceptable in more than 40% per cent of sequences independently from the discourse relation. This is not the case for *alors*, where only consequence is judged as being perfectly acceptable by annotators. These scores shall be interpreted keeping in mind that original sequences were produced in spoken language, whereas this study re-uses them in their transcribed (i.e. written) form.

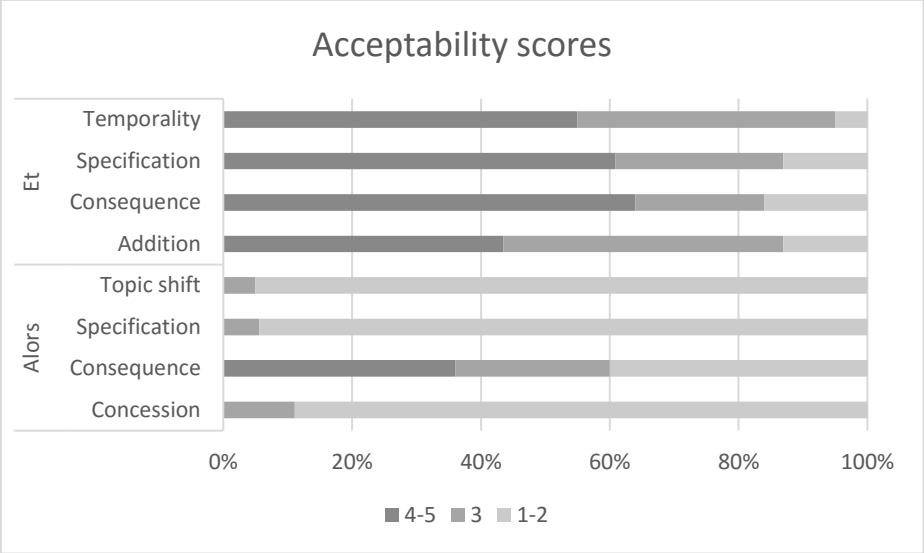


Fig. 1. Mean acceptability scores for the original sequences (i.e. with the DMs)

3.2 Inter-annotator agreement

Annotations were analysed based on several criteria. First, we examined the inter-annotator agreement using the Fleiss' κ measure for multiple annotators ([13]). Values are presented in the Table 1.

Only small tendencies can be observed across conditions, where removing *alors* would slightly enhance the inter-rater agreement, while deleting *et* would make it more difficult for subjects. However, differences between the two conditions can not be considered as substantial. This result suggests that the overall interpretation of discourse relations would not be dependent on presence / absence of the DM in the studied sequences. Interestingly, while sequences containing *alors* were judged more severely in the acceptability task, we can see that in the annotation task, their interpretation seem to be less difficult compared to the *et* sequences.

Table 1. Fleiss' κ for each DM in each condition.

DM	With DM	Without DM
alors	0.429	0.498
et	0.328	0.295

3.3 Expert *vs.* naïve annotations

First condition (with visible DM). When compared to the original expert annotations performed by the authors, the percentage of naïve annotators that chose the same discourse relation exceeds 50% for each of the six discourse relations. However, these scores differ depending on the relation, ranging from 50.9% for the temporality relation expressed by the DM *et* to 75.8% for the consequence relation expressed by *alors*. These two extreme values seem to be linked with the core meaning of these two connectors, in that *alors* often can induce a relation of consequence, whereas *et* alone (i.e. without any other temporal cues) is not likely to be perceived as conveying a temporal relation between two segments.

However, the results relating to the expression of addition by *et* were unexpected. Naïve annotators were somewhat reluctant to identify *et* as conveying the relation of addition between S1 and S2, even though it corresponds to the core meaning of this marker. Furthermore, this result also seems to refute the hypothesis that some of the discourse relations would be more or less transparent for non-experts. In cases where *et* was used to express consequence, there was agreement between the expert and naïve annotators in 64% of the cases.

We then analysed cases of disagreement between naïve and the expert annotators, in order to detect regularities in the identification of relations in these sequences. Remember that naïve annotators were given a substituable DM for each discourse relation and a paraphrase they could use in order to disambiguate the discourse relation. However, we do not know how each naïve annotator did proceed in order to choose one discourse relation out of 4 possible relations. Results show that sequences with *alors* annotated as expressing specification and concession by the experts tend to be annotated as expressing consequence by naïve annotators (in 18.59% and 21.82% respectively), and topic shift tends to be interpreted as a concession (in 17.32% of all annotations). In cases where there was no agreement on *alors* signalling a consequence relation, it was mainly interpreted as concession. In the case of *et*, subjects were inclined to annotate other relations as being either an addition (namely consequence annotated as addition in 15.81%, specification in 25.30% and temporality in 22.53%) or a consequence (sequences identified as addition by the experts were annotated as consequence by the naïve annotators in 18.18% of all cases). Detailed results can be seen in the Table 2.

Table 2. Experts' annotation of discourse relations (rows) and percentage of naïve annotations choosing to annotate the sequence as expressing each of the four discourse relations, with DM. In **bold**, inter-group agreement (in %). In *italics*, the second choice of naïve annotators. ADD – addition, CCS – concession, CSQ – consequence, SPE – specification, TOS – topic shift, TMP - temporality

Alors	CCS	CSQ	SPE	TOS	Et	ADD	CSQ	SPE	TMP
CCS	**60.90%**	*21.82%*	7.73%	9.55%	ADD	**55.73%**	*18.18%*	15.02%	11.07%
CSQ	*11.25%*	**75.76%**	9.96%	3.03%	CSQ	*15.81%*	**63.64%**	13.04%	7.51%
SPE	8.68%	*18.59%*	**65.29%**	7.44%	SPE	*25.30%*	11.86%	**56.13%**	6.72%
TOS	*17.32%*	9.09%	6.49%	**67.10%**	TMP	20.95%	*22.53%*	5.53%	**50.99%**

Second condition (without visible DM). In the second condition (i.e. without the original DM in the sequences), results show some modifications as for the agreement between naïve and the expert annotators. First, the scores of agreement (in percentages) seem to increase for utterances originally containing the DM *alors* except for the consequence relation where a loss of 11 per cent has been noticed. Thus, the consequence relation tends to be more difficult to identify when the *alors* DM is deleted. On the other hand, removing the same DM from sequences carrying out other relations seems to enhance their interpretation, reinforcing the idea of consequence being the core meaning of this DM and leading us to suppose that the use of *alors* in other situations can be troubling for the naïve annotators, especially in written stimuli (without audio).

Utterances originally containing the DM *et* exhibit the opposite behaviour in that the agreement between the two groups of annotators decreases systematically, except for the discourse relation of consequence. This relation is more or less stable with a gain of 2 per cent compared to the annotation with the original DM, which points to a strong tendency to infer cause-effect relations even in the absence of an explicit marker (cf.

causality-by-default hypothesis, [14]). By contrast, the temporality score of inter-group agreement falls down to 42.69%.

Turning to the cases where no agreement was observed between the naïve and the expert annotators, the tendencies for *alors* do not change from the results obtained in the first part of this experiment (e.g. where sequences contained the DM). Again, the only exception concerns the consequence relation, which tend to be interpreted as specification when the DM is removed. As for *et*, when excluding the identical annotations, subjects mostly identified addition as being specification, whereas all the other relations were mostly annotated as addition (Table 3).

Table 3: Experts' annotation of discourse relations (rows) and percentage of naïve annotations choosing to annotate the sequence as expressing each of the four discourse relations, without visible DM. In **bold**, inter-group agreement (in %). In *italics*, the second choice of naïve annotators. ADD – addition, CCS – concession, CSQ – consequence, SPE – specification, TOS – topic shift, TMP - temporality

Alors	CCS	CSQ	SPE	TOS	Et	ADD	CSQ	SPE	TMP
CCS	**72.27%**	*10.91%*	9.55%	7.27%	ADD	**52.96%**	15.02*%*	*18.18%*	13.83%
CSQ	9.52%	**64.07%**	*19.48%*	6.93%	CSQ	*16.21%*	**65.61%**	10.67%	7.51%
SPE	6.20%	*10.74%*	**78.86%**	6.20%	SPE	*32.81%*	7.51%	**52.17%**	7.51%
TOS	*17.32%*	3.03%	6.49%	**73.16%**	TMP	*28.46%*	20.16%	8.70%	**42.69%**

4 Discussion

These results lead to the conclusion that, in some cases, naïve subjects do not base their judgment exclusively on the core meaning or on the assumed transparency of discourse markers, but on other elements as well (in line with the results of [15]). Some of the additional elements that affect the interpretation of an utterance may include syntactic or lexical patterns as well as prosody. This study also discusses the informative value of the DM on the construal of discourse relations by comparing annotations with or without the original DM. We have shown that the two DMs do not seem to have the same impact on the annotation of discourse relations: deleting *alors* improves the inter-group agreement scores, whereas deleting *et* deteriorates the same scores. For the two DMs, however, the consequence relation behaved differently from other relations. Moreover, inter-group differences observed in annotations underline the importance of using naïve participants to such tasks in order to compare alternative interpretations of linguistic phenomena (using a methodology similar to the one in [16], for example). The use of interpretations by naïve subjects also raises the question whether we, as experts, do not overestimate the possibility of reaching unambiguous interpretation of discourse relations, since psycholinguists have convincingly shown that “that language processing is sometimes only partial and that semantic representations are often incomplete” ([17]).

The present preliminary study is carried out in the larger context of a research project investigating the contribution of prosody to the online interpretation of discourse relations. The next steps of the project involve production and perception studies on the discriminating value of some prosodic parameters in the disambiguation of the discourse relations and DMs investigated in the present paper.

References

1. Mann, W., Thompson, S.: Rhetorical Structure Theory: Toward a functional theory of text organization. Text 8(3): 243-281 (1988).
2. Schiffrin, D.: Discourse Markers. Cambridge, UK: Cambridge University Press (1987).
3. Couper-Kuhlen, E., Kortmann, B. (Eds.).: Cause - Condition - Concession - Contrast : Cognitive and Discourse Perspectives. Berlin New York: Mouton de Gruyter (2000).
4. Asher, N., Lascarides, A.: Logics of Conversation. Cambridge, UK: Cambridge University Press (2003).
5. Luscher, J.M., Moeschler, J.: Approches dérivationnelles et procédurales des opérateurs et connecteurs temporels: les exemples de et et de enfin. Cahiers de Linguistique française 11: 77-104 (1990).
6. Hansen, M.B. M.: Alors and donc in spoken French: a reanalysis. Journal of Pragmatics 28: 153-187 (1997).
7. Spooren, W., Degand, L.: Coding coherence relations: Reliability and validity. Corpus Linguistics and Linguistic Theory 6(2): 241-266 (2010).
8. Zufferey, S., Degand, L.: Representing the meaning of discourse connectives for multilingual purposes. Corpus Linguistics and Linguistic Theory 10 (2013).
9. Crible, L. Degand, L. (In press): Reliability vs. granularity in discourse annotation: What is the trade-off? Corpus Linguistics and Linguistic Theory (In press).
10. Scholman, M., Evers-Vermeul, J., Sanders, T.. A step-wise approach to discourse annotation: Towards a reliable categorization of coherence relations. Dialogue and Discourse 7(2): 1-28 (2016).
11. Degand, L., Martin, L., Simon, A.C.: LOCAS-F: un corpus oral multigenre annoté. Proceedings of Congrès Mondial de Linguistique française: 2613-2626 (2014).
12. Degand, L., Simon, A.C.: Variation of Discourse Markers across a multi-genre corpus of spoken French: Annotating function and meaning. Presentation at Variation et changement pragmatico-discursif 3 (DiPVaC), Toronto, 4-6 May 2016 (2016).
13. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5): 378-382 (1971).
14. Sanders, T.: Coherence, causality and cognitive complexity in discourse. In Aurnague, M., Bras, M. (eds.), Proceedings of the First International Symposium on the Exploration and Modelling of Meaning: 31-46. Toulouse: Université de Toulouse-le-Mirail (2005).
15. Mak, W. M., Tribushinina, E., Andreiushina, E.: Semantics of Connectives Guides Referential Expectations in Discourse: An Eye-Tracking Study of Dutch and Russian. Discourse Processes, 50(8): 557–576 (2013).
16. Scholman, M., Demberg, V.: Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task (pp. 24–33). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-0803 (2017).
17. Ferreira, F., Bailey, K. G. D., & Ferraro, V. Good-Enough Representations in Language Comprehension. Current Directions in Psychological Science, 11(1), 11-15 (2002).

Identifying Discourse Markers through Automated Semantic Annotation - Using the UCREL Semantic Analysis System as a Pre-annotation Tool

Furkó, B Péter
Károli Gáspár University of the Reformed Church, Hungary
furko.peter@gmail.com
https://orcid.org/0000-0002-9650-4785

Abstract. In this paper a comparison of automated and manual annotation of oral discourse markers (DMs) is presented. Firstly, we outline the criterial features of DMs that are relevant to the disambiguation of DM and non-DM tokens. Secondly, the UCREL Semantic Analysis System (USAS) and its disambiguation methods are briefly presented.

The following research questions are addressed: (1) Are the disambiguation methods USAS uses adequate for filtering out non-DM tokens of the most frequent DM types? (2) Does the margin of error reported to apply in general apply to the identification of DMs as well? (3) Are individual DMs identified / tagged with a similar margin of error? (4) If individual DMs are tagged with varying precisions by USAS, what formal-functional properties of the relevant DMs might explain the differences?

Keywords: Discourse Marker, Automated Semantic Annotation, Manual Annotation, D-function Ratio, Disambiguation

1 Introduction

There is a rapidly growing body of research on primarily oral discourse structuring devices referred to as discourse markers (henceforth DMs), discourse connectives, discourse operators, discourse particles, cue phrases, pragmatic markers, framing devices; several terms have been used as the function of the number of theoretical frameworks that have been applied (Relevance Theory, coherence-based studies, sentence grammar, interactional sociolinguistics, etc.). It is widely agreed that such expressions play a vital role in discourse structuring and / or utterance interpretation, there is, however, disagreement on the criteria one can use to delimit this class of linguistic items.

Several lists have been provided of the formal, functional and stylistic criteria that are associated with DMs as a functional class, cf. e.g. [1-3], still, few authors provide (and many claim it is impossible to provide) an exhaustive list of criterial features that can be used to identify all instances of DMs in a given corpus. An even more challenging task is to develop annotation software that can automatically identify DMs in oral discourse and filter out non-DM tokens of lexical items that are frequently used as DM types (e.g. adverbial uses of *well* or *now*, prepositional uses of *like*, etc.).

Accordingly, the present paper will explore the utility of using an automated semantic tagging software, USAS as a pre-annotation tool for the identification of oral DMs,

including interpersonal as well as textual markers. The paper will argue that automated semantic annotation (ASA) can be an effective tool depending on the scope of the inquiry and with regard to certain DMs, but needs to be complemented by extensive manual error correction.

2 Oral discourse markers - criterial features

Formal criterial features that have been identified so far include syntactic heterogeneity, non-propositionality and a consequent syntactic optionality, quasi-initiality, phonological reduction and comma intonation, while functional criteria include functional (as well as variable) scope, semantic bleaching and a consequent pragmatic enrichment, (extreme) context-dependence and multifunctionality. Moreover, frequency, orality, gender-specificity and negative attitudes (stigmatization) have also been identified as stylistic features of the functional class of DMs. A comprehensive account is beyond the scope of the paper, for a detailed discussion cf. e.g. [1, 3, 4]

Crible [5] argues that three of these features (lack of syntactic integration, functional scope and multifunctionality) allow for a consistent and extensive definition of DMs, while the combination of the three can also be used as an operationalization for empirical analyses. Accordingly, in the following analysis these three criteria will be primarily applied in the course of the manual annotion of DMs with a view to testing the automated identification of individual lexical items' DM and non-DM tokens.

3 Automated semantic annotation: disambiguation methods and precision

There are a variety of computerized semantic tagging (CST) systems, including artificial intelligence-based, knowledge-based, corpus-based, and semantic taxonomy-based systems (for an overview, cf. e.g. [6]). The present analysis draws on the results gained from the UCREL Semantic Analysis System (USAS), which has the advantage of combining these approaches. Furthermore, USAS groups lexical items in terms of a taxonomy of semantic fields and assigns semantic categories to all words, including grammatical and other procedural (non-propositional) items, which is relevant for the present study in view of the fact that the lexical items under scrutiny are highly procedural and semantically bleached, cf. [1].

USAS system uses an automated coding scheme of 21 semantic fields, subdivided into 232 sub-categories. For reasons of brevity, only the tags that have been associated with the DM types under analysis will be discussed, the complete coding scheme can be found at http://ucrel.lancs.ac.uk/usas/. USAS uses disambiguation methods including part-of-speech tagging, general likelihood ranking, multi-word-expression extraction, domain of discourse identification, and contextual rules, for a detailed discussion cf. [7]. Previous evaluations of the accuracy of the system reported a precision value of 91%, cf. [7], i.e. a 9% margin of error applying to lexical items across the board (including propositional and non-propositional items).

The research questions are as follows:

1. Are the disambiguation methods USAS uses adequate for filtering out non-DM tokens of the most frequent DM types?
2. Does the margin of error reported to apply in general apply to the identification of DMs as well?
3. Are individual DMs identified / tagged with a similar margin of error?
4. If individual DMs are tagged with varying precisions by USAS, what formal-functional properties of the relevant DMs might explain the differences?

4 Methodology

In the course of the research, two sub-corpora of the same size (100,000 words each) have been used:

- a corpus of the official transcripts of 37 confrontational type of mediatized political interviews (henceforth MPI sub-corpus) selected from BBC's *Hard Talk* and *Newsnight* (available at http://bbc.co.uk);
- a corpus of the official transcripts 50 celebrity interviews (henceforth CI sub-corpus) downsampled from CNN's *Larry King Live* (available at http://www.cnn.com).

The two sub-corpora have been extensively studied in previous research, thus, the results of automated tagging have been compared to findings based on manual annotation and a combination of quantitative and qualitative methods, cf. [8–9]. Previous research was aimed at finding genre-specific patterns of DM use in the two sub-corpora, which has informed the present paper in terms of the D-values (see section 5 below).

The research process has been as follows: in order to identify and compare the USAS tags of oral DMs in the two sub-corpora, we looked up the semantic tags assigned to frequent DMs, such as *I mean, you know, in other words, so, well* etc. and then used those semantic tags to identify further types and tokens relevant to discourse marking. As a result, 95.1% of the instances of DMs we trawled from the two sub-corpora through this method were found to be either tagged with Z4, described in the USAS manual as the "discourse bin" (including items such as *oh, I mean, you know, basically, obviously, right, yeah, yes*) or with A5.x, described as "evaluative terms depicting quality" (including DMs such as *well, OK, okay, good, right, alright*). Subsequently, we put together a list of the most frequently Z4/A5.x-tagged lexical types, and calculated the ratio between DM-relevant tags (i.e. Z4 and A5.x) and non-DM relevant tags (e.g. B2, I1.1, T1.3, etc., see below for details) in the case of each item on the list.

In the second stage, a representative sample of 400 tokens in the MPI sub-corpus were manually annotated using a numeric code of 1 for DM and 2 for non-DM tokens with a view to comparing the results of automated and manual tagging. When deciding if an individual token is a DM or not, the three criterial features identified by Crible [5] (see section 2 above) were applied by a single expert annotator. The tokens that were selected for the sample were weighted for their frequency in the corpus, while DM and non-DM tokens were included in equal proportions. For example, the 429 tokens of

well comprise 19.6% of all Z4/A5.x-tagged items in the corpus, thus, 78 tokens, (39 A5.1-tagged and 39 non-A5.1 tagged) were included in the sample.

5 Research findings

Since both sub-corpora were compiled in a way that they are of the same size of 100,000 words, table 1 below summarizes the raw frequency of the relevant lexical items' DM and non-DM related USAS tags.

lexical item	**raw frequency of DM-related tag in the MPI**	**raw frequency of DM-related tag in the CI**	**raw frequency of non-DM-related tag in the MPI**	**raw frequency of non-DM-related tag in the CI**
well (429)	360xA5.1	312xA5.1	14xI1.1, 55xN5	1xA7, 2xB2, 24xN5
sort (38)	14xZ4	25xZ4	21xA4.1, 3xA1.1.1	10xA4.1
now (299)	4xZ4	1xZ4	288xT1.1.2, 7xZ5	229xT1.1.2, 6xZ5
(you) know (346)	205xZ4	455xZ4	140xX2.2, 1xZ6	307xX2.2
like (97)	6xZ4	17xZ4	51xZ5, 40xE2+	238xZ5, 139xE2+
(I) mean (141)	114xZ4	201xZ4	27xQ1.1	30xQ1.1, 5xS2.2.2
(in other) words (11)	4xZ4	13xZ4	7xQ.3	7xQ.3
actually (165)	165xA5.4	72xA5.4	0	0
(I) think (549)	126xZ4	121xZ4	423xX2.1	319xX2.1
right (114)	55xZ4, 53xA5.3	211xZ4, 98xA5.3	6xT1.1.2	12xN3.8, 16xS7.4, 15xT1.1.2

table 1. summary of DM and non-DM-related semantic tags assigned to the most frequent DM types in the MPI and CI sub-corpora

As a first step, we compared the ratio of DM and non-DM tokens of individual items with the results of previous research [8], in the course of which DMs in the same sub-corpora were manually annotated. In order to gauge the categorial multifunctionality of DMs the measure of D-function ratio or D-value, a term proposed by Stenström [10], was used. An individual item's D-value is calculated as a quotient of the number of tokens that fulfill discourse-pragmatic functions and the total number of occurrences in a given corpus. The D-value of *oh*, for example, is 1 (100%) in the London-Lund Corpus, since it is used exclusively as a DM, whereas *well* showed a D-value of 0.86, as 14% of its tokens serve non-DM (adverbial, nominal, etc.) functions [10].

After calculating the D-values of individual DMs based on the above values and comparing them to the findings of [8], the results of automated annotation and manual annotation yielded converging values. The lexical item *mean*, for example, has a D-value of 0.808 in the MPI corpus based on automated annotation (calculated as the number of Z4 tags divided by all tokens of *mean*), while manual annotation yielded a D-value of 0.797 (cf. [8]). Similarly, manual annotation yielded a D-value of 0.82 for *well* in the MPI corpus in [8], while table 1 yields a D-value of 0.839 for this lexical item (360 Z4 tags divided by the total number of tokens, i.e. 429).

The table also correctly predicts that most of the lexical items under scrutiny have higher D-values in the CI sub-corpus than in the MPI sub-corpus, which is explained by the fact that there is a higher degree of conversationalization in celebrity interviews, i.e. they are more similar to spontaneous, informal, face-to-face conversations (cf. [11]). For example, the D-value of *well* is 0.92, the D-value of *mean* is 0.851 in the CI sub-corpus based on automated annotation (312 A5.1 tags divided by a total of 339 tokens, 201 Z4 tags divided by a total of 236 tokens, respectively).

In the second stage of the research a representative sample of tokens in the MPI were manually annotated using the numeral 1 for DM tokens and 2 for non-DM uses. With a view to comparing the results of automated and manual annotation, all DM-related tags (Z4 and A5.x) yielded by USAS were re-coded as 1, while non-DM tags (B2, I1.1, T1.3, etc.) were re-coded as 2. Consequently, the extracted list of the corresponding manual and automated tags was entered into a reliability calculator (Freelon's ReCal 2 for 2 coders) in order to calculate inter-annotator agreement statistics. Table two below shows the result.

	Percent Agreement	Scott's Pi	Cohen's Kappa	N Agree-ments	N Disa-greements	N Cases	N De-cisions
Variable 1 (cols 1 & 2)	92.75	0.854519	0.854527	371	29	400	800

Table 2. Inter-annotator agreement between automated and manual tagging of DM / non-DM tokens

Although the above inter-coder agreement values appear high (cf. [12]), it is important to note that there is a great degree of variation in the precision with which individual DMs are tagged by USAS. On the one hand, there are DMs such as *I mean* and *you know* whose DM and non-DM uses are disambiguated with high precision, resulting in a kappa score of <.98, i.e. close to perfect intercoder agreement between USAS and the human annotator. This is expected to be due to two of the disambiguation methods USAS applies: firstly, its multi-word-expression extraction algorithm and its core component of MWE lexicon (cf. [7]), secondly, the fact that POS tagging enables the parser to differentiate between syntactically integrated tokens that are monotransitive (and are thus followed by their nominal or clausal complements) and syntactically non-integrated ones that are marked by the absence of complements. On the other hand, there are lexical items that are invariably tagged with the same (sometimes DM-

relevant, other times non-DM relevant) tags regardless of their syntactic (non-)integration and functional scope. For space considerations, only two examples will be given, one for DM-relevant invariant tagging, and one for non-DM relevant invariant tagging.

An example for the former is *actually*, which might be used as a DM that has the ensuing discourse unit in its scope (1) or as an adverbial modifier that has scope over the verb it modifies as in 2 below (all extracts are from the USAS-tagged CI corpus, emphases are mine):

(1) No_Z4 ,_PUNC that_Z8 was_A3+ n't_Z6 exactly_A4.2+ the_Z5 reason_A2.2 ._PUNC ***Actually_A5.4+*** **,**_PUNC what_Z8 it_Z8 was_A3+ ,_PUNC is_Z5 I_Z8mf felt_X2.1 that_Z5 films_Q4.3 were_Z5 getting_A9+ they_Z8mfn started_T2+ to_Z5 be_Z5 repeating_N6+ ._PUNC

(2) They_Z8mfn 're_A3+ one_T3 of_Z5 the_Z5 few_N5- cats_L2mfn in_Z5 the_Z5 world_W1 that_Z8 can_A7+ ***actually_A5.4***+ swim_M4 under_M4[i619.2.1 water_M4[i619.2.2

An example for non-DM relevant invariant tagging is *now*, which can be used as a DM that marks topic shift (3) or as a circumstance adverb (4). However, as table 1 shows, USAS has assigned a Z4 tag to only 5 DM uses of *now*, and in 530 out of 535 cases both DM and non-DM uses are labelled as T1.1.2, i.e. as 'general terms relating to a present period/point in time':

(3) Good_Z4[i297.2.1 heavens_Z4[i297.2.2 ,_PUNC such_Z5 an_Z5 intelligent_X9.1+ man_S2.2m is_Z5 excited_X5.2+ about_Z5 a_Z5 movie_Q4.3 star_W1 ?_PUNC ***Now_T1.1.2*** what_Z8 about_Z5 her_Z8f and_Z5 the_Z5 Kennedy_Z1mf 's_Z5 ?

(4) Somebody_Z8mfc explain_Q2.2/A7+ to_Z5 Paris_Z2 and_Z5 Nicole_Z1f ,_PUNC live_L1+ means_X4.2 we_Z8 're_A3+ on_Z5 television_Q4.3 right_T1.1.2[i7.2.1 ***now_T1.1.2***[i7.2.2 ._PUNC

6 Conclusions, utility and limitations of using USAS as a pre-annotation tool

In answer to the research questions posed in section 3 above, the following can be observed. The disambiguation methods USAS uses are efficient for calculating the ratio between DM and non-DM tokens of the most frequent DM types: using USAS enables the researcher to obtain an adequate global picture of the D-values of most of the lexical items under scrutiny. In addition, the margin of error reported to apply in general also applies to the identification of DMs collectively, and, in the case of multi-word units such as *you know* and *I mean*, individually as well. However, we find a great degree of variation in the precision / margin of error with which non-multi word DMs are tagged. The varying precisions are due to DMs' criterial features of source category layering, syntactic non-integration, variable / functional scope. These features provide challenges

to the disambiguation methods USAS applies, such as general likelihood ranking, and multi-word-expression extraction.

References

1. Schourup L (1999) Discourse markers: tutorial overview. Lingua 107: 227–265. doi:10.1016/S0024-3841(96)90026-1
2. Fraser B (1999) What are discourse markers? Journal of Pragmatics 31: 931–952. doi: 10.1016/S0378-2166(98)00101-5
3. Beeching, K (2016) Pragmatic Markers in British English: Meaning in Social Interaction. Cambridge University Press, Cambridge. doi: 10.1017/CBO9781139507110
4. Furkó BP (2014) Cooptation over grammaticalization: The characteristics of discourse markers reconsidered. Argumentum 10: 289–300.
5. Crible, L (2017) Towards an operational category of discourse markers: A definition and its model. In: Pragmatic Markers, Discourse Markers and Modal Particles: New perspectives. John Benjamins, Amsterdam, pp 99–124. doi: 10.1075/slcs.186
6. Prentice S (2010) Using automated semantic tagging in Critical Discourse Analysis: A case study on Scottish independence from a Scottish nationalist perspective. Discourse & Society 21(4): 405–437. doi: 10.1177/0957926510366198
7. Rayson P, Archer D, Piao S, McEnery T (2004) The UCREL Semantic Analysis System. Paper given at Beyond Named Entity Recognition Semantic Labeling for NLP Tasks in LREC'04, Lisbon.
8. Furkó BP, Abuczki Á. (2014) English Discourse Markers in Mediatised Political Interviews. Brno Studies in English 40: 45-64. doi: 10.5817/BSE2014-1-3
9. Furkó BP, Kertész A, Abuczki Á (forthcoming) Discourse Markers in Different Types of Reporting. In: Capone A (ed) Indirect Reports - Perspectives in Pragmatics, Philosophy and Psychology. Springer, Heidelberg.
10. Stenström AB (1990) Lexical items peculiar to spoken discourse. In: Svartvik J (ed) The London-Lund Corpus of Spoken English: description and research. Lund University Press, Lund, pp 137–175.
11. Furkó BP (2017) Manipulative uses of pragmatic markers in political discourse. Palgrave Communications 3/ 17054. doi:10.1057/palcomms.2017.54
12. Spooren W, Degand L (2010) Coding coherence relations: reliability and validity. Corpus Linguistics and Linguistic Theory 6(2): 241–266. doi:10.1515/cllt:2010.009

Aligning connective lexicons for a multilingual database

Yulia Grishina, Peter Bourgonje, and Manfred Stede

Computational Linguistics
UFS Cognitive Science
University of Potsdam, Germany
firstname.lastname@uni-potsdam.de

The identification and classification of discourse connectives plays a central role in many discourse processing approaches. When dealing with key challenges of discourse connectives (syntactic heterogeneity, functional ambiguity (connective vs. non-connective reading) and sense ambiguity), specific discourse connective resources in the form of lexicons can help. Several such monolingual lexicons exist; for German (Stede, 2002), French (Roze et al., 2012), English (Prasad et al., 2008), Portuguese (Mendes and Lejeune, 2016) and Italian (Feltracco et al., 2016) for example.

While valuable resources for their respective languages, in this submission we propose and evaluate a method to go from a mono-lingual lexicon to a bilingual lexicon on the basis of a case study for contrastive connectives in German and Italian (Bourgonje et al., 2017). The purpose of this is two-fold: First we want to map German connectives to their Italian counterparts and in reverse. Understanding the different translations of connectives and their contexts can help translators and also language learners. The second goal of this approach is to find gaps in the mono-lingual lexicons, which appear when a frequent alignment of some Italian connective is not yet in the German lexicon, or the other way round. We have recently started with a similar approach and setup for German-English and German-Dutch (for Dutch there is no lexicon available yet, so we extend our setup to evaluate its suitability for the construction of a new lexicon). But as this is ongoing work, in this submission we focus on the procedure and results for German-Italian.

First, we extracted the German contrastive connectives from DiMLex[1] (Scheffler and Stede, 2016), a connective lexicon containing 275 entries. The set of Italian contrastive connectives comes from LICo (Feltracco et al., 2016), a similar lexicon for Italian containing 170 entries[2]. Both lexicons share the same structure, thus including orthographical variants, syntactic type, discourse sense, and usage examples for each of the entries. The sense annotations are based on the Penn Discourse Treebank (PDTB) senses (Miltsakaki et al., 2008) in its latest version 3.

For the parallel German/Italian corpus we used Europarl (Koehn, 2005), as it still appears to be the biggest resource of this kind, and it is, conveniently,

[1] https://github.com/discourse-lab/dimlex
[2] https://hlt-nlp.fbk.eu/technologies/lico

already sentence-aligned. From the 1,832,053 sentences in the German-Italian part of the corpus we extracted the word alignments using MGIZA++ (Gao and Vogel, 2008).

To arrive at the alignments for connectives, we approach the problem from two sides; once using Italian as the source and German as the target and once the other way round. We locate every connective in the source language lexicon in the word-aligned corpus and store its alignments in a key-value structure, where the key is the position in the sentence and the value the corresponding word. This is necessary because a single word in the source language can align to multiple words in the target language. The resulting structure is ordered by position (ascending) and their corresponding words are joined, resulting in the target language word or phrase. Note that this procedure works for the same for null-alignments, single-word, multi-word and discontinuous connectives, where for the latter we extract only the connective words in order of appearance, not the content in between. The results of this lookup procedure have to be manually evaluated, due to the potential functional ambiguity and sense ambiguity of connectives. If the alignment is the result of a word appearing in its non-connective reading, or of a word appearing not in the desired discourse sense (contrast in the case of our German-Italian case study), we want to discard this alignment. However, because there is no such classifier available, which automatically distinguishes for the languages at hand here, we have to manually check the list of alignments for correctness. In our case study, we first applied a frequency filter (discarding all alignments below some frequency threshold) and then had the remaining list checked by a native speaker. A visual example of the result of this procedure is shown in Figure 1.

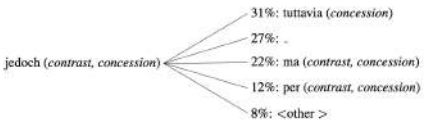


Fig. 1. Most frequent alignments of *jedoch*

The results of our experiment are two-fold: (a) We obtained valid bilingual mappings between German and Italian connectives based on their alignment frequency and (b) identified gaps in both lexicons in terms of new connective candidates and discourse senses for the already described connectives. To exemplify the latter, we found that *anstelle dessen* from DiMLex is aligned to *invece* in LICo, and by mapping the occurrences of *invece* back to their German correspondences, we found that *anstelle* (in isolation, without *dessen*) is also present among the set. Therefore, we considered *anstelle* as a valid connective candidate that needs to be added to DiMLex. Some other connective candidates found for German include *umgekehrt* and *(ganz) im Gegenteil*. Similarly, we found several

connective candidates for Italian that were not present in LICo, such as *al contempo* or *solo che*. Overall, while some of the found candidates can clearly serve as connectives and could be added to the corresponding lexicon right away, for other cases more corpus evidence is required to decide whether they can indeed function as connectives in the language in question.

Furthermore, considering discourse senses, we found that several Italian connectives only had the Concession sense, while the corresponding German connectives also had the Contrast sense, such as *comunque*, for which we found the German alignments *aber*, *allerdings* and *doch*, for example. Additionally, we observed that Italian connectives with a sense Contrast or Concession are frequently aligned to their German counterparts with a sense Substitution, such as *anstelle-invece.* Having examined the parallel examples more closely, we conclude that assigning both senses would be valid for both German and Italian, although they are placed distantly in the PDTB hierarchy of senses. These findings are confirmed by Feltracco et al. (2016), who acknowledge that the distinction between the two senses was one of the main cases of the inter-annotator disagreement. We conclude that both lexicons could benefit from adding additional senses gained via comparing parallel translations.

Having recently developed a GUI for the existing (monolingual) connective lexicons (available at `http://connective-lex.info`), we plan to include the mappings between languages in this platform, so that in addition to the current search functionality (string search, search by syntactic type and by the connective's discourse sense), the user can also search for the alignments of the connective to that in different languages.

In sum, we present, to the best of our knowledge, the first mapping of Italian-German contrastive connectives across parallel corpora. Specifically, we were able to establish correspondences between the two monolingual lexicons as well as identify several missing entries for both languages. Once the information is organized in a complete bilingual database, it can assist translation, and conclusions can be drawn regarding connective distribution, sense distribution and ambiguity in the different languages. Our next steps include the disambiguation of connective- and non-connective readings and the implementation of more sophisticated filtering strategies to retrieve more reliable connective candidates. In addition, we are interested in extending this study for different languages pairs, starting with German-English and German-Dutch.

Bibliography

Bourgonje, P., Grishina, Y., and Stede, M. (2017). Toward a bilingual lexical database on connectives: Exploiting a German/Italian parallel corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it)*, Rome, Italy.

Feltracco, A., Jezek, E., Magnini, B., and Stede, M. (2016). Lico: A lexicon of italian connectives. In *Proceedings of the 3rd Italian Conference on Computational Linguistics (CLiC-it)*, Napoli, Italy.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Mendes, A. and Lejeune, P. (2016). Ldm-pt. a portuguese lexicon of discourse markers. In *Conference Handbook of TextLink Structuring Discourse in Multilingual Europe Second Action Conference*, Budapest, Hungary.

Miltsakaki, E., Robaldo, L., Lee, A., and Joshi, A. (2008). *Sense annotation in the Penn Discourse Treebank*, pages 275–286. Springer Berlin Heidelberg, Berlin, Heidelberg.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The penn discourse treebank 2.0. In *In Proceedings of LREC*.

Roze, C., Danlos, L., and Muller, P. (2012). Lexconn: a french lexicon of discourse connectives. *Discours - Revue de linguistique, psycholinguistique et informatique*.

Scheffler, T. and Stede, M. (2016). Adding semantic relations to a large-coverage connective lexicon of German. In et al., N. C., editor, *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, Portoro, Slovenia.

Stede, M. (2002). DiMLex: A Lexical Approach to Discourse Markers. In *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria.

Identification of Thematic Discourse Relations on the Data from an Annotated Corpus of Czech

Eva Hajičová and Jiří Mírovský

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

Abstract. In the present contribution we analyze the data of the Prague Discourse Treebank 2.0 (PDiT 2.0; M. Rysová et al., 2016) as for the text coherence based on the so-called thematic progressions, that is links between sentences with regard to their topic–focus articulation (information structure). For this purpose, we work with two ingredients of the PDiT annotation, namely (i) the annotation of the anaphoric relations ("proper" coreference and some basic types of bridging) between sentence elements (both at short and at long distance), and (ii) the bipartition of the sentence into Topic (T) and Focus (F) based on the annotation of contextual boundness.

Keywords: thematic progressions, topic–focus articulation, anaphoric relations.

1 Related Work

1.1 Centering Theory

One of the most deeply elaborated and best known theory of discourse (local) coherence is the so called *centering theory* (Grosz, Joshi and Weinstein, 1995) based on the model of the local attentional states of speakers and hearers as proposed by Grosz and Sidner (1986). Each utterance in discourse is considered to contain a *backward looking center* which links it with the preceding utterance and a set of entities called *forward looking centers*; these entities are ranked according to language-specific ranking principles stated in terms of syntactic functions of the referring expressions. The *transitions* from one utterance to the following one are then specified by rules that capture their ordering: the most preferred are '*continue*' and '*retain*' (the backward looking center of a given utterance equals the backward looking center of the preceding utterance) followed by '*smooth shift*'and '*rough shift*' (the backward looking center of a given utterance differs from the backward looking center of the preceding utterance). The intuition which is behind this ranking of transitions is very close to those behind the notion of the low cost effort (Fais 2004, p.120).

Interesting experiments investigating the effects of utterance structure and anaphoric reference on discourse comprehension examined in the context of utterance pairs with parallel constituent structure (e.g., *Josh criticized Paul. Then Marie insulted him*) are reported in Chambers (1998). The results reveal several limitations in

centering theory and suggest that a more detailed account of utterance structure is necessary to capture how coreference influences the coherence of discourse.

A corpus-based evaluation of the preferences proposed in centering theory is given by Poesio et al. (2000). The study has reached some interesting results. As for the 'shifts' rule stating that (sequences of) continuations are preferred over (sequences of) retains, which are preferred over (sequences of) shifts, the tests revealed that there are more shifts than retains.

1.2 Thematic Progressions

To our knowledge, the first comprehensive treatment of the dynamic development of discourse, though clad in psychological rather than linguistic considerations, was given by Weil (1844, quoted here from the 1978 E. transl.). Weil recognized two types of the "movement of ideas", *marche parallèle* and *progression*: "If the initial notion is related to the united notion of the preceding sentence, the march of the two sentences is to some extent parallel; if it is related to the goal of the sentence which precedes, there is a progression in the march of the discourse" (p. 41). He also noticed a possibility of a reverse order called 'pathetic': "When the imagination is vividly impressed, or when the sensibilities of the soul are deeply stirred, the speaker enters into the matter of his discourse at the goal." (p. 45.)

In Czech linguistics, this idea is later reflected in Daneš' notion of *thematic progressions* (Daneš 1970; 1974), explicitly referring to the relation between the theme and the rheme of a sentence and the theme or rheme of the next following sentence (a simple linear thematic progression and a thematic progression with a continuous theme), or to a 'global' theme (derived themes) of the (segment of the) discourse.

2 Corpus Based Study

In our present corpus-based analysis we focus our attention on the issue of local coherence as established by links between the thematic (Topic) and rhematic (Focus) parts of sentences in different genres of discourse. For this purpose, we use the data from the Prague Discourse Treebank 2.0, which offers a good testing bed as it provides – in addition to the dependency underlying (deep) syntactic relations – annotation of (i) contextual boundness from which the Topic–Focus bipartition of the sentence can be derived, and (ii) basic anaphoric relations, incl. some types of bridging. Such an annotation has allowed us to follow the occurrence of the two basic types of thematic progressions mentioned above, namely (i) continuous theme (Topic), i.e. the Topic of the given sentence is anaphorically related to the Topic of the previous sentence, and (ii) the "progressive" rheme (Focus), i.e. the Topic of the given sentence is anaphorically related to the Focus of the previous sentence.

2.1 Small Sample

For the first step, in which we wanted to test whether our research methodology and the corpus material available may lead to some interesting and representative results,

we have randomly chosen 6 documents of 5 genres with the total of 150 sentences and applied the (already implemented) algorithm for the division of the sentence into Topic and Focus based on the values of the TFA attribute (with values non-contrastive contextually bound, contrastive contextually bound and contextually non-bound).[1] As a result, we had at our disposal the total of 150 dependency trees with marked (binary) division into Topic and Focus and with the annotation of coreference and basic bridging relations between referring expressions of the adjacent sentences.

On this sample, we have followed four possible "thematic" relations between neighbouring sentences (the boundary between Topic and Focus is indicated in our examples by a slash):[2]

(i) (some element of the) Topic of the sentence *n* refers to (some element of the) Topic of the sentence *n-1* (denoted below as $T_{n-1} \leftarrow T_n$):

Myšlenka stručného ústavního zákona, který by prostě stanovil, že výdaje státního rozpočtu mají být kryty příjmy téhož roku, / se vyskytla v řadě zemí. Nejrozsáhlejší diskuse na toto téma / se odehrála v 80. letech ve Spojených státech.

The idea of a concise constitutional law, which would simply state that the state budget expenditures are to be covered by the same year's income, / has occurred in a number of countries. The most extensive discussion on this issue / took place in the 1980s in the United States.

(ii) (some element of the) Topic of the sentence *n* refers to (some element of the) Focus of the sentence *n-1* (denoted below as $F_{n-1} \leftarrow T_n$):

Dnes je každý / pod novinářskou diktaturou. Diktatura jest / nehlučná, ale jest.

Today everybody is / under a journalist dictatorship. Dictatorship is / not noisy, but it is.

(iii) (some element of the) Focus of the sentence *n* refers to (some element of the) Focus of the sentence *n-1* (denoted below as $F_{n-1} \leftarrow F_n$):

Barevný terčík / usnadňuje nakládání pošty do kontejnerů. Během přepravy barva / zlepšuje přehled o tom, zda se zásilka nezpožďuje.

The coloured disc / makes easier the loading of the mail into containers. During the transport the colour / makes the information easier whether the article is not delayed.

[1] The Topic-Focus bipartition of the sentence has been carried out automatically based on the primary opposition of contextually bound and non-bound items reflected in the PDiT by a manual assignment of one of three values of the attribute of TFA. The distinction of contextual boundness should not be understood in a straightforward etymological way: an *nb* element may be 'known' in a cognitive sense (from the context or on the basis of background knowledge) but structured as non-bound, 'new', in Focus. The overall accuracy of the algorithm, measured on the assignment of tectogrammatical nodes either to Topic or Focus of the sentence, is 0.93 (Rysová et al., 2015).

[2] The examples in this section are original sentences from the PDiT.

(iv) (some element of the) Focus of the sentence *n* refers to (some element of the) Topic of the sentence *n-1* (denoted below as $T_{n-1} \leftarrow F_n$).

Novináři jsou / hlídací psi společnosti. Taková je / všeobecně sdílená představa o poslání novinářů.

Journalists are / watching dogs of the society. This is / a generally shared image of the mission of journalists.

"An element x refers to an element y" means that there is an anaphoric link (be it a proper coreference or a bridging relation) between the referring expressions x and y in adjacent sentences. As for the genres of the more closely studied documents, in this first step our attention was focussed on the essay and letter genre.

Our starting assumption was that if the sentence is to be "about" something (i.e. about the Topic of the sentence), this "something" has to be somehow established (anchored) in the memory of the addressees. This is why we first examined the types (assumed as prototypical) $T_{n-1} \leftarrow T_n$ and $F_{n-1} \leftarrow T_n$, that is the pairs of sentences in which Topic refers to the Topic of the previous sentence ("continuous Topic") or in which the Topic refers to the Focus of the previous sentence ("progression of Focus"). This assumption has been confirmed in both genres, but there was a difference which of the two types prevails in which genre: $T_{n-1} \leftarrow T_n$ occurred twice as often than $F_{n-1} \leftarrow T_n$ in the letter document, while in the essay genre, $F_{n-1} \leftarrow T_n$ occurred three times as often than $T_{n-1} \leftarrow T_n$. With the non-prototypical relations, that is with the types $F_{n-1} \leftarrow F_n$ and $T_{n-1} \leftarrow F_n$, both types occurred rather rarely in the letter genre but the type $F_{n-1} \leftarrow F_n$ was surprisingly frequent in the essay type (13 occurrences as compared to 20 of $F_{n-1} \leftarrow T_n$ and 8 of $T_{n-1} \leftarrow T_n$). Under a more detailed inspection, it has been found that in most of these cases the anaphoric relation of an element in F_n leads from a contextually bound element of Focus. This finding is in an agreement with the assumption (made explicit in Hajičová, Partee and Sgall, 1998) of the theory of TFA we subscribe to that the recursive character of this articulation makes it possible (or even necessary) to distinguish between the "overall" bipartition of the sentence into its Topic and Focus and the local partitioning within these two parts into what may be called "local Topic" and "local Focus".

2.2 Large Data

To obtain a more general picture of the distribution of the different types of "thematic" relations as attested in larger data, we applied the analysis onto a collection of 10 genres, namely (i) advice, (ii) comment, (iii) description, (iv) essay, (v) invitation, (vi) letter, (vii) news, (viii) overview, (ix) review and (x) survey. We put under scrutiny documents containing more than 20 sentences and looked for anaphoric chains globally, that is we did not restrict our search to adjacent sentences. Taking into account anaphoric chains consisting of two elements only, the results obtained for all these genres are as follows: as for the relations leading from the Topic of the given sentence to some preceding sentence, the $F_{n-x} \leftarrow T_n$ sequences prevailed considerably (3 436 cases) over

Table 1. Anaphoric chains.

Frequency	Anaphoric chain
3 436	F – T
3 307	F – F
1 863	T – T
1 439	T – F
643	F – T – T
597	F – F – F
432	T – T – T
…	
184	F – F – F – F
…	
36	F – T – T – T– F
…	
9	F – T – T – F – F – T
etc.	

the $T_{n-x} \leftarrow T_n$ type (1 863 cases); the total number of these typical relations was 5 299. This result indicates that continuous topic, i.e. the anaphoric relations between Topics of two sentences, are considerably less frequent than the progression of focus, i.e. anaphoric reference from the Topic of the given sentence to an element in the Focus of (some of) the preceding sentence(s).

2.3 Non-Typical Cases

However, the relations we consider to be non-typical (leading from the Focus of a given sentence to an element in the Topic or in the Focus of (some of) the previous sentence(s)) occurred surprisingly frequently (the total of 4 746 cases, out of which $F_{n-x} \leftarrow F_n$ type was found in 3 307 cases and the type $T_{n-x} \leftarrow F_n$ was found in 1 439 cases). These figures have led us to a deeper analysis of these non-typical cases. For this purpose we have sorted the material obtained in this step according to the length of the coreference chains, i.e. according to the "course" ("progression") of the given anaphoric relation throughout the document. In this way, we obtained a list (and frequences) of two-element chains, three-element chains etc. sorted by the four above mentioned "directions" of anaphoric relations. Table 1 is an illustration of the resulting data, where in the first column there is the frequency of the given relation, and F(ocus) and T(opic) denote the part of the sentence in which there occur the referring expressions linked by the given anaphoric link. (The first four lines of the Table are those mentioned in Sect. 2.2 above.)

We have put under a more detailed scrutiny the cases of what might be called "continuous foci" (i.e. the type F – F – F etc.) to see under which conditions they

arise. For this purpose we have analyzed 40 examples in which the length of the "continuous foci" was 4 and more. Here again, in 29 cases the anaphoric link leads from a contextually bound element of F which supports the necessity to distinguish local topics and local foci with the overall Topic and Focus. The rest of the cases include (i) bridging relations rather than proper coreference, (ii) a list in Focus (e.g. list of exhibitions in a locality), (iii) change of speakers of sentences in the Focus of which the referring expression occurs.

The obtained data have allowed us also to follow the distance between the referring expressions in terms of the number of sentences in between them. The starting hypothesis is that the longer the chain, the more probable is the re-occurrence of the referring expression in the Focus of the sentence. A perfunctory look at the collected data indicates that this is an important factor: e.g. in the above mentioned chain, the distance (indicated by numbers of intervening sentences) is as follows: F -*1*- F -*3*- F -*3*- F. One of the points of our future inquiry will be to investigate the dynamism of discourse in terms of the necessity to re-introduce an item in the Focus part of the sentence based on the "distance" and also in terms of the form of the referring expression, e.g. when a reference by a pronoun (or even a zero pronoun) is possible and when it is necessary to refer to some "fading" item by a noun or a nominal group. For the overall framework and hypotheses for such an inquiry, see Hajičová and Vrbová (1982), Hajičová (2003) and Hajičová and Hladká (2008).

3 Conclusions

In the present contribution we have focused on the intersentential relations based on coreferential chains (both proper coreference and some basic types of bridging relations) with regard to the bipartion of the sentences into their Topic and Focus. We first verify the accepted methodology on a small sample of texts from two genres of the annotated texts from the multi-layered Prague Discourse Treebank 2.0, followed by an analysis of a more representative sample of annotated texts from nine genres. We have also taken into account the length of the anaphoric chains and the length of the segments (in terms of the number of sentences) in between two expressions referring to the same item.

The following observations have been reached:

(a) among the four possible types of the relations between anaphoric links and the Topic–Focus bipartition of the sentence, the most frequently occurring type is a link between the Topic of the sentence to the Focus of the previous sentence; this is in contrast to the assumption of Fais (2004) based on the low cost and Chamber's (1998) assumption of structural parallelism, but in favour of Poesio et al.'s (2004) finding on the prevailance of shifts to retain relation.

(b) If compared with the studies on thematic progressions in English carried out by Czech linguists (see e.g.Dušková 2008), the structural parallelism seems to be valid for English, thanks to the function of English subject in the grammatically fixed word order. Our observations seem not to support such a parallelism for Czech, a language the word order of which is guided by communicative factors

rather than by grammatical rules.

(c) In case there is an anaphoric link leading from the Focus of a sentence to the Topic or Focus of the preceding sentence:

(i) this link frequently leads from a contextually bound element of the Focus of the given sentence, which supports the assumption that it is convenient to distinguish between the "overall" Topic and Focus and the local Topic and Focus; and/or

(ii) the anaphoric relation is of the type of bridging, which is often intepreted as a contrast.

Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (projects GA17-03461S and GA17-06123S). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

Chambers, C. (1998). Structural Parallelism and Discourse Coherence: A Test of Centering Theory, *Journal of Memory and Language*, Volume 39, Issue 4, November 1998, Pages 593-608

Daneš, F. (1970), Zur linguistischen Analyse der Textstruktur. *Folia linguistica* 4:72-78.

Daneš, F. (1974). Functional Sentence Perspective and the organization of the text. In: Daneš, Ed. *Papers on Functional Sentence Perspective*. Prague: Academia, 106-128.

Dušková, L. (2008). Theme movement in academic discourse. In: M. Procházka and J. Čermák, Eds., *Shakespeare between the Middle Ages and Modernity.* From translators art to academic discourse. Prague, FF UK, 221-247.

Fais, L. (2004). Inferable centers, centering transitions, and the notion of coherence. *Computational linguistics 30*, 119-150.

Grosz, B. and C. L. Sidner (1986). Attention, Intentions and the structure of discourse. *Computational Linguistics*, 12, 175-204.

Grosz, B. J., Joshi, A. K. and S. Weinstein (1995). Centering: A Framework for modeling the local coherence of discourse. *Computational Linguistics, 21*, 203-225.

Hajičová, E. (2003). Aspects of Discourse Structure. In: *Natural Language Processing between Linguistic Inquiry and System Engineering* (ed. by W. Menzel and C. Vertan), Iasi, pp. 47-56.

Hajičová, E. and B. Hladká (2008). What does sentence annotation say about discourse? In *18th International Congress of Linguists*, Abstracts , The Linguistic Society of Korea, Seoul, Korea, pp. 125-126

Hajičová, E. and J. Mírovský (in prep.). Discourse Coherence Through the Lens of an Annotated Text Corpus: A Case Study. Accepted for LREC 2018.

8

Hajičová, E., Partee, B. H. and P. Sgall (1998). *Topic-Focus Articulation, Tripartite Structures, and Semantic Content,* Dordrecht , Kluwer Academic Publishers.

Hajičová, E. and J. Vrbová (1982). On the role of the hierarchy of activation in the process of natural language understanding. In: Horecký J., Ed. , *Coling 82 – Proceedings of the Ninth International Congress of Computational Linguistics.* Amsterdam: John Benjamins. 107-113.

Poesio, M., Stevenson, R., Di Eugenio, B.and J. Hitzeman (2004). Centering: a parametric theory and its instantiations. *Computational Linguistics* 30, 309-363

Rysová, K., Mírovský, J. and E. Hajičová (2015). On an apparent freedom of Czech word order. A case study. In: *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015),* IPIPAN, Warszawa, Poland, ISBN 978-83-63159-18-4, pp. 93-105.

Rysová, M., Synková, P., Mírovský, J., Hajičová, E., Nedoluzhko, A., Ocelák, R., Pergler, J., Poláková, L., Pavlíková, V., Zdeňková, J. and Š. Zikánová (2016). *Prague Discourse Treebank 2.0.* Data/software, ÚFAL MFF UK, Prague, Czech Republic, http://hdl.handle.net/11234/1-1905, Dec 2016

Weil, H. (1844). *De l'order des mots dans les langues anciennes comparées aux langues modernes,* Paris: Joubert. Translated by Charles W. Super as *The order of words in the ancient languages compared with that of the modern languages,* Boston: Ginn, 1887, reedited and published by John Benjamins, Amsterdam 1978.

The linguistic marking of coherence relations

The interaction between segment-internal elements and connectives

Jet Hoek,[1] Sandrine Zufferey,[2] Jacqueline Evers-Vermeul,[1] and Ted J.M. Sanders,[1]

[1] Utrecht University
[2] University of Bern

When readers or listeners are presented with a text, they do not treat the individual clauses and sentences in that text as independent and unrelated. Instead, they try to relate each part of the text, or each discourse segment, to the rest of the discourse by inferring coherence relations between the discourse segments. Language users can be facilitated in inferring coherence relations by the presence of connectives or cue phrases (from now on referred to as 'connectives') that provide explicit processing instructions on how to relate two discourse segments to each other (cf. Sanders & Spooren 2007), but many coherence relations have to be inferred without the help of a connective.

Traditionally, relations with a connective, as in (1), have been labeled 'explicit' coherence relations; relations without a connective, as in (2), as 'implicit' relations. Although this distinction seems very straightforward, it is not without its problems. Connectives can for instance signal a relation that is less specific than the relation that is constructed by language users, as in (3), where the relation is marked by after, a temporal connective, but the inferred relation is causal. The relation in (3) is therefore less explicitly signaled than the relation in (2). In addition, a relation without a connective may contain strong other cues that help language users infer the appropriate relation. The semantic opposition between loves and despises, for instance, could be argued to function as a signal for the contrastive coherence relation in (4). This relation is then more explicitly signaled than the relation in (2), even though neither fragment contains a connective.

(1) [Kate missed last night's dinner]$_{S1}$ because [she had to take all three of her dogs to the emergency vet.]$_{S2}$
(2) [The day after Christmas Mark always makes chocolate mousse.]$_{S1}$ Ø [It is a great way to get rid of all the leftover holiday candy.]$_{S2}$
(3) [Paul was banned for life from the bowling alley]$_{S1}$ after [getting drunk and hiding all the bowling pins.]$_{S2}$
(4) [Harry loves Easter.]$_{S1}$ Ø [His sister despises all big holidays.]$_{S2}$

Connectives are the most prototypical linguistic elements that signal how discourse segments should be related to each other, which is why research on discourse coherence has mostly been focused on connectives as markers of coherence relations. However,

while connectives are the only linguistic elements that by definition express relational meaning, that does not necessarily mean they are the only indicators for coherence relations. By limiting our attention to connectives, we are likely missing out on important other cues readers and listeners use when establishing coherence relations.

The most elaborate research effort to identify other signals for coherence relations has been the recently released RST Signalling Corpus (Das, Taboada, & McFetridge 2015), in which linguistic cues that signal coherence relations annotated in the RST Treebank (Carlson, Okurowski, & Marcu 2002) are identified. While the RST Signalling Corpus is an extremely valuable inventory of potential signals for coherence relations, it does not (yet) draw a systematic link between signals and specific relation types, and does not comment on how or why the indicated signals function as cues for coherence relations. In addition, since the annotation was mostly focused on relations without connectives, the RST Signalling Corpus does not identify potential additional signals in relations that contain a connective. In this presentation, we will explore the marking of coherence relations by connectives on the one hand, and other types of cues on the other. Specifically, we will investigate how linguistic elements within the segments of a coherence relation, i.e., segment-internal elements, can contribute to the marking of the relation, and how the presence of segment-internal signals relates to the presence of connectives. We consider elements to be segment internal if they are integrated in and are part of the propositional content of the clauses that are, or are part of, the segments of a coherence relation.

Within the field of discourse, there are several segment-internal features that have been linked to particular types of coherence relations. These segment-specific elements include a wide range of linguistic categories, such as complex phrases, lexical items, modal markers, and verbal inflection. The features can either occur in one of the segments or in both of the segments. It seems, however, that not all linguistic elements that have been associated with a specific type of coherence relation signal the relation in the same way, and there appear to be differences in the way in which the presence of a specific linguistic element in the segments of a relation can impact the marking of that relation by means of a connective. In this presentation, we will argue that there are, at least, three distinct ways in which segment-internal elements systematically interact with the connective that marks a coherence relation. We label these interactions division of labor, agreement, and general collocation. In division of labor types of interactions, as between the negation element and the connective *instead* in (5), the connective and the other signal overlap in the meaning they encode and the presence of one is likely to make (part of) the other redundant; in agreement types of interactions, as between the explicit evaluation and the subjective causal connective in (6), the connective and the other signal overlap in the meaning they encode, but they are commonly used in addition to each other. In general collocation types of interactions,

as between the implicit causality verb and the causal connective in (7), there is no overlap in the meaning signaled by the connective and the other signal.

(6) [Bob did not go to the park.] Instead, [he went to the cinema.]

(7) [Dat is echt een belachelijke beschuldiging,]$_{S1}$ want [dat zou ik nooit doen.]$_{S2}$
'That is a ridiculous accusation, since I would never do such a thing.'

(8) [Jared congratulated Gail]$_{S1}$ because [she won the pie eating competition.]$_{S2}$

We base our categorization on several combinations of segment-internal elements and connectives that have been linked to each other on the basis of monolingual corpus data, theoretical explorations, or experimental work, e.g., Webber (2013) for CHOSEN ALTERNATIVE relations like the one in (6), Sanders (1997) for subjective causal relations like the one in (7), and Au (1986) for causal relations following implicit causality verbs, like the one in (8). We then use parallel corpus data to show that the three types of interactions we formulated are observable in translation. We opt for a translation corpus because even though monolingual corpora are extremely valuable resources for language research, when studying meaning they require researchers to rely on their own interpretations, since "meaning is not directly observable," (Noël 2003:758). When it comes to the interaction between segment-internal elements and connectives, it is not necessarily obvious what and how each element contributes to the overall interpretation of a relation. A proposal for an alternative method to research meaning is to make use of parallel corpora, which consist of a source text and one or multiple translations (cf. Dyvik 1998, Noël 2003). In this approach, the translator is treated as a naive 'annotator,' whose main purpose was to accurately convey the meaning of the source text fragment in the target language. The three types of interactions between segment-internal elements and connectives we identify make different predictions for translation and our corpus data show that we can indeed distinguish distinct translation patterns for each type of interaction. Identifying a way in which linguistic elements other than connectives can be systematically linked to coherence relations is an important step toward fully understanding the marking of coherence relations.

References

1. Au, Terry K. 1986. A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. *Journal of Memory and Language* 25(1). 104-122.
2. Carlson, Lynn, Mary Ellen Okurowski, & Daniel Marcu, 2002. RST Discourse Treebank. Philadelphia: Linguistic Data Consortium.

3. Das, Debopam, Taboada, Maite, and McFetridge, Paul, 2015. RST Signalling Corpus LDC2015T10. Web Download. Philadelphia: Linguistic Data Consortium.
4. Dyvik, Helge, 1998. A translational basis for semantics. In: Stig Johansson & Signe Oksefjell (eds.), *Corpora and cross-linguistic research: Theory, method and case studies*, 51-87. Amsterdam: Rodopi.
5. Noël, Dirk, 2003. Translations as evidence for semantics: An illustration. *Linguistics* 41(4), 757-785.
6. Sanders, T.J.M. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes* 24(1), 119-147.
7. Sanders, Ted J.M. & Wilbert P.M.S. Spooren, 2007. Discourse and text structure. In: Dirk Geeraerts & Hubert Cuyckens (eds.), *The Oxford Handbook of cognitive linguistics*, 916-941. Oxford: Oxford University Press.
8. Webber, Bonnie L. 2013. What excludes an alternative in coherence relations? *Proceedings of the 10th International Workshop on Computational Semantics (IWCS2013)*. 276-287.

TextLink Web Portal

Murathan Kurfalı[1], Ahmet Üstün[1], and Bonnie Webber[2]

[1]Informatics Institute, Middle East Technical University (ODTÜ)
{kurfali,ustun.ahmet}@metu.edu.tr
[2]School of Informatics, University of Edinburgh, bonnie@inf.ed.ac.uk

December 15, 2017

1 Introduction

This paper introduces the TextLink Web Portal - an online web service for searching discourse-annotated text, that stands as one of the promised outcomes of the TextLink Cost action[1]. As implemented, the portal serves two purposes: (1) to enable researchers to display and filter discourse annotations according various parameters[2] and then, if desired, download the resulting set; (2) to provide access to the growing multi-lingual TED-MDB corpora, allowing researchers to examine cross-lingual parallel discourse annotation. As a web service, the TextLink Web portal requires no installation and is easy to master, yet it is capable of performing complex queries. The rest of the paper explains its capabilities in detail.

2 TextLink Web Portal

The TextLink Web Portal is being developed as a publicly available web-site for examining, annotating and summarizing discourse annotation in either monolingual text or parallel cross-lingual bi-texts.

Currently, the portal has four sections:

- *Home*: The home page contains links to other pages of the web portal as well as to the main web site of TextLink. Users can also download sample files and the user manual from the home page.
- *Upload Annotations*: This page allows users to upload text files and their (stand-off) discourse annotation, on which they want to perform searches. Users can access this page any time they want to upload additional files. When uploading files, users should indicate the language of the annotations so that, whenever necessary, portal can retrieve information from the relevant DIMLex corpora.
- *Search (Monolingual)*: This page allows users to display their annotations and filter them through various search options. Section 2.1 gives a detailed description of the search page.
- *TED-MDB Search (Multilingual)*: This page hosts aligned annotation of files from the TED-MDB, to allow researchers to make cross-lingual comparison in discourse level annotations among the covered languages (see Section 2.2).

In the near future, we plan to implement an online annotator, similar to the PDTB annotator[3], where users can upload new text files and annotate their discourse relations without having to have a local copy of the PDTB annotator. In addition to existing abilities of PDTB annotator, the online annotator will be able to provide information regarding annotated tokens based on the previous annotations and (where available) DimLex lexicons.

[1]www.textlink.ii.metu.edu.tr
[2]Currently, portal only accepts annotations produced by either PDTB Annotator(Lee et al., 2016) or DATT(Aktaş, Bozsahin, & Zeyrek, 2010)
[3]http://www.seas.upenn.edu/ pdtb/annotator.html

2.1 Monolingual Search

2.1.1 User Interface

The user interface consists of the following three main blocks (Figure 1):

- *Search panel*: The search panel resides on the top of the page. It allows a user to select annotation files, determine the search parameters and, provided that a connective is selected, display the list of the senses conveyed by the selected connective using connective-lex.
- *Annotation list*: The list on the left-hand side presents all annotation tokens for the file being searched. The list is updated when a user selects another file or performs a search. The selected annotation token becomes highlighted on the text. In the list, each annotation is represented with the discourse connective, if any, along with the type of the relation and the senses, if any, it conveys.
- *Text Panel*: The main panel displays the text file which was annotated. When an annotation is selected, it automatically scrolls to the annotation.

Figure 1: A screenshot of the TextLink Web Portal search page

2.1.2 Search Facilities

The portal offers various filtering options. The search can be performed via interface without needing to write any SQL queries. The search options are listed below:

- *Sense Search*: One can search for tokens with a particular sense or senses by selecting them from the drop-down menu or typing or typing them in. Users can select as many senses as they want. The selected senses are combined by *or*, meaning that all relations which involve at least one of the selected senses will be included in the result set.

  Furthermore, as it is not uncommon for a discourse relation to convey two senses simultaneously, users can specify two senses, as well as how they should be combined using the operator menu. There are two possible operator, 'and'– 'not'. When the latter is selected, all the relations conveying the first sense but not the second one are retrieved.

- *Type Search*: Users can specify the type of the relation they want by selecting the check-boxes provided. Users can select as many check-boxes as they want.

- *Connective Search*: Finally, users can filter the annotations according to the discourse connective they possess. Users can select as many connective as they want by either typing or

using the drop down list. Although PDTB anchors implicit discourse relations to a connective, which is referred as 'implicit connective ', the connective search is limited to Explicit and AltLex relations, as insertion of implicit connectives is prone to inconsistency.

All search parameters can be combined. That is, users can perform searches of the following kind: retrieve all '*Explicit*' relations which convey an '*Expansion*' sense but '*not*' any of the '*Temporal*' senses via the connective '*and*'.

For the time being, although users can upload as many files as they want, search can be performed on only one file at a time. However, we plan to extend this so that the specified search criteria will be applied to all the files uploaded.

Finally, the annotations uploaded by users are only stored in the portal during their session. That is, the portal does not store any files permanently, in order to provide confidentiality to users. Therefore, anyone can use the portal without risking their annotations to be publicized without their permission.

2.1.3 Download Facility

Portal enables users to download their search results for further processing. The results can be saved in the original file format (e.g. pipe delimited file) or as a CSV file. The difference is, CSV file contains text spans rather than byte spans (or XML tags, as in the case of DATT), rendering the results more readable, as well as suitable to process with office tools such as Excel. On the other hand, anyone who wishes to further process the results through the portal (or the Annotator where the annotations are prepared) can save them in the original pipe-delimited format.

2.1.4 DIMLex Facility

The TextLink Web Portal also incorporates available DIMLex-style lexicons through the connective-lex web-site[4]. Briefly, connective-lex provides an interface to perform search on available DIMLex corpora. Among others, connective-lex provides the list of the senses conveyed by any given connective. In the portal, when a user performs a search containing a connective , portal displays that list on the search page enabling to compare between the annotations included in the result set and all possible senses conveyed by the given connective according to DIMLex. Portal periodically retrieves DIMLex files from connective-lex through its API in order to avoid any inconsistency between two sites.

2.2 Multilingual Search (TED-MDB)

TED-MDB is a multilingual corpus of TED-talks, annotated in the style of the PDTB, currently covering six languages (English, European Portuguese, German, Polish, Russian, Turkish). Recently, the discourse relations of two talks in each language have been aligned with respect to English through semi-automatic means. Multilingual search option of the portal enables to access aligned tokens of the TED-MDB corpus.

A sample view of the page is provided in Figure 2. The search options are the same as those of monolingual search; however, there are two text panels where the rightmost one always displays the English tokens. Users can select among the languages and the files using the drop-down menus in the File/Language menu. Only the tokens in the selected language are filtered, as it is likely that aligned tokens convey different senses or may be of different types across languages. Whenever, a token is selected on the left annotation list, its English counterpart, if any, is automatically selected and highlighted.

3 Implementation Details

The portal currently resides on a Amazon EC2 server. The server-side logic is implemented using Django Framework(Version 1.11.6, http://www.djangoproject.com/) making use of Structured Query Language (SQL) in order to retrieve the desired annotation tokens from the databases. All queries are generated automatically from the user' selections via interface. The interfaces, which are referred as *templates* in Django, are coded in HyperText Markup Language (HTML) with JavaScript, including AJAX, in order to provide necessary functionality.

[4]http://connective-lex.info/

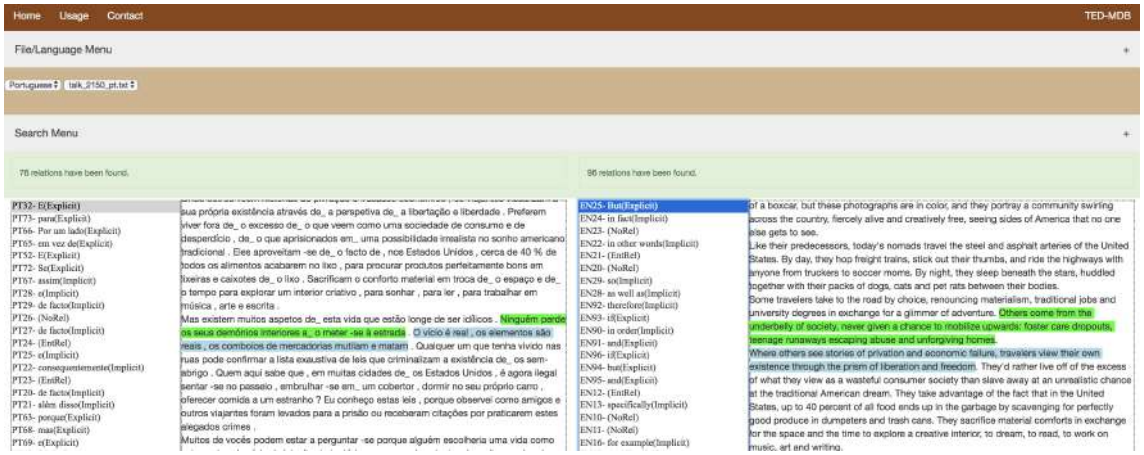


Figure 2: A screenshot of the TextLink Web Portal TED-MDB search page

References

Aktaş, B., Bozsahin, C., & Zeyrek, D. (2010). Discourse relation configurations in turkish and an annotation environment. In *Proceedings of the fourth linguistic annotation workshop* (pp. 202–206).

Lee, A., Prasad, R., Webber, B. L., & Joshi, A. K. (2016). Annotating discourse relations with the pdtb annotator. In *Coling (demos)* (pp. 121–125).

A bottom-up analysis of sentence-initial DRDs in the Finnish Internet

Veronika Laippala[1], Aki-Juhani Kyröläinen[2], Filip Ginter[1], Jenna Kanerva[1], Johanna Komppa[3] and Jyrki Kalliokoski[3]

[1] University of Turku, Finland
[2] McMaster University, Canada
[3] University of Helsinki, Finland

Keywords: DRDs, Web-as-corpus, Dependency syntax.

1 Introduction

DRDs, such as 'moreover' and 'thus' in English, challenge the methods and materials commonly applied in linguistics: they include a variety of syntactic categories, and they are often polysemous [12,21,22]. Typically, they are examined by analyzing the DRD and the coherence relation(s) it signals. Studies are often based on the manual interpretation of the coherence relation in the context, either by the researcher or by annotators of a ready-made corpus. The coherence relations are described in several taxonomies adapting various backgrounds: for instance, Sanders et al. [23] with a cognitive motivation and Mann and Thompson [20] with a more computationally oriented perspective.

This presentation explores the use of 24 Finnish DRDs by adopting an alternative, data-driven approach. Instead of manually interpreted coherence relations, the analysis is based on 1) automatically estimated typical usage patterns associated with the DRDs and 2) their grouping into clusters composed of DRDs with similar usage patterns. This method allows for the study of DRDs in very large datasets and in languages for which no manually annotated discourse treebank exist. In this presentation, we take the first steps in the analysis of the linguistic outcome of this method. We analyze the similarities and differences between the DRDs that the usage contexts reveal and the distinguishing patterns that these contexts associate with the DRDs. What do they tell about the DRDs, about their use and the coherence relation they signal? How do these data-driven groupings relate to theory-based solutions? The methodological aspects of the study and the process of constructing the co-occurrence patterns of the DRDs are explained in detail in [16].

The study is inspired by recent work in cognitive linguistics which apply co-occurrence information to study, e.g., lexical semantics [1,6]. These studies follow the underlying assumption of usage-based studies, which considers that the semantic and functional properties of linguistic expressions can be analyzed based on their distributional characteristics [7,10]. Co-occurrence information has been applied also in the study of discourse connectives. For instance, Sanders and Spooren [25] analyzed Dutch

causal DRDs and their functions by examining co-occurrence information on, e.g., modality, the specific DRD and the coherence relation, and Levshina and Degand [17] aimed at distinguishing between the subjective and objective meanings of *because* based on morphological, syntactic and semantic co-occurrence patterns.

In the previous studies [1,6,25], the co-occurrence patterns were composed of manually or semi-manually [17] annotated patterns. To adapt the method to larger corpora and be able to examine a larger group of DRDs, as suggested by Gries [8], we apply detailed, automatically produced syntactic co-occurrence information in the form of unlexicalized syntactic n-grams: subtrees of dependency analysis with the lexical information deleted. While syntactic n-grams can be generated in different lengths, we applied *biarcs*, i.e. constructions composed of three tokens related by two dependency relations (see Figure 1) [9,13].

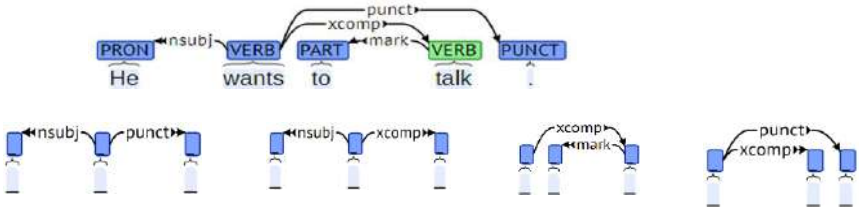


Figure 1: A dependency syntax analysis and the subsequent unlexicalized syntactic n-grams. *nsubj:* nominal subject, *punct:* punctuation, *xcomp:* clausal complement.

2 Corpus and methods

The corpus of the study consists of the Finnish Internet Parsebank, a 3.7 billion token collection of Finnish crawled from the Internet. The Parsebank has automatic syntax analyses produced with the Finnish Dep parser with a labeled attachment score of 82.1% [18]. The examined DRDs were chosen from the list of 100 most frequent sentence-initial words tagged as coordinating conjunctions or adverbs in the corpus. The DRDs were extracted from the Parsebank with a context of two full sentences: the sentence with the DRD and the preceding one. Altogether, this gave us a corpus of 469,997,522 words. At this point of the process, we also deleted the studied DRDs from the data to ensure that they do not affect the subsequent analysis. Then, we transformed the two-sentence chunks into unlexicalized syntactic biarcs using the tool by Kanerva et al. [13]. Finally, we performed several post-processing steps in order to clean the data (see [16] for details). To construct the co-occurrence patterns for the analysis, we counted the frequency of co-occurrence between each DRD and a biarc. Moreover, we distinguished the co-occurrence patterns according to the position of the biarc in the two-sentence window: *hit* biarcs occurred in the sentence with the DRD and *context* biarcs in the preceding one.

To examine the similarities between the DRDs and the groupings that this data-driven approach reveals, we applied clustering to the co-occurrence patterns. The clustering was done with the function hclust in R (version 3.3.1) [14]. A solution of six

clusters offered the best fit to the data and is presented in Figure 2. The translations are based on the definitions of the DRDs in *CGF* [4].

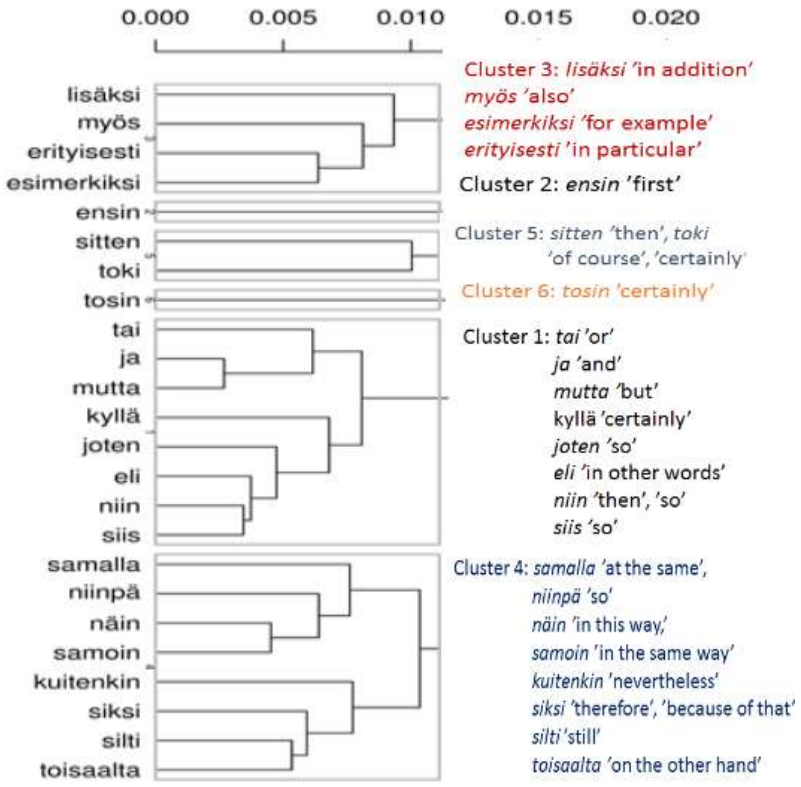


Fig. 1. The six clusters produced by the data-driven clustering method.

The cluster solution consists of two larger clusters and four smaller ones. Importantly, some of the data-driven groupings have similarities with the groupings presented in theory-based solutions. For instance, the additive and elaborative connectives are all placed in cluster 3. This validates our method and proves that distributional information on syntax can be used as a basis for examining similarities between DRDs without manually annotated data. Further, it shows that DRDs signaling similar coherence relations share similar contexts. By grouping all the coordinating conjunctions in cluster 1, the cluster solution also suggests that these conjunctions and the other cluster DRDs share similar contexts. To understand what these contexts are, and why the other clusters are formed as they are, we need to examine the important usage patterns of the clusters and the DRDs in them. To this end, we applied supervised machine learning, namely a support vector machine (SVM). Its task was to predict the cluster label based on the syntactic n-grams generated from the two-sentence chunk with the DRD. Additionally, the SVM estimates the important n-grams of the clusters that contribute to their identification. These reflect the linguistic characteristics of the clusters.

The SVM was fitted on a subset of the corpus consisting of 149,999 random two-sentence chunks for each DRD. The fitting was repeated 1000 times, and the corpus was randomly split to training and testing (80%/20%) on each run. The average precision of the SVM was 41% and recall 42%. The final list of important n-grams for each cluster was formed based on the n-grams that appeared in the 30 top ranking n-grams in at least 75% of the rounds. This consists of 105 unique n-grams that reflect the linguistic characteristics of the clusters and their DRDs. These include 62 n-grams with a positive co-efficient value, thus having a positive association with the cluster DRDs, and 43 n-grams with a negative value, thus having a negative association with the DRDs. A large majority, 83 of the important n-grams were tagged as *hits*, i.e. they were placed in the sentence with the DRD, while only 22 were placed in the preceding sentence. This highlights the role of the local linguistic content in the functioning of the DRDs, even if the related elements and the coherence relation concern larger parts of the text.

3 Linguistic motivation behind the clusters

The important n-grams estimated by the SVM give a clear picture about the distinguishing usage contexts of the DRDs and about the linguistic motivation behind the cluster solution. Some of the usage contexts reflect patterns that are associated with a particular coherence relation signaled by the DRDs. Namely, this is the case for the smaller clusters 2,3,5 and 6. These patterns confirm previous theoretically oriented descriptions of coherence relations and are illustrated below.

For the two larger clusters, the important usage patterns reflect characteristics typical of spoken-like (cluster 1) and written-like (cluster 4) internet genres. These give novel information about their uses and explain why the clusters group together DRDs signaling different coherence relations. Conjunctions in cluster 1 are used in more dialogic contexts and internet genres which rely on features of spoken discourse, whereas the additive and elaborative connectives in cluster 4 represent lexically more specified DRDs typical of written discourse and more formal genres.

Finally, some of the typical usage patterns associated with the clusters denote subjective expressions and argumentative patterns. This suggests that the cluster solution may also reflect the division of DRDs to those that express *subjective* coherence relations relating speech acts, and those that express *objective* relations relating real-world events [20,22,23]. This, however, is a very complex issue that the current presentation cannot discuss thoroughly. We will, however, give examples on these usage patterns in order to illustrate the question.

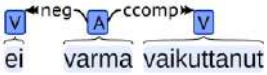
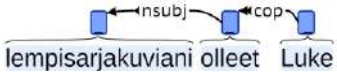
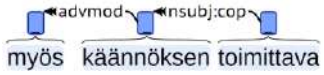
3.1 Smaller clusters: usage patterns relating DRDs with specific coherence relations

The important usage patterns estimated by the SVM for the smaller clusters reflect characteristics of the coherence relations signaled by the DRDs. For example, clusters 5 and 6 include two concessive DRDs, *tosin* and *toki*, which both can be translated as

'certainly', 'although'. The important usage patterns associated with these clusters include frequent negations. This is motivated by the semantics of the concessive relation, which relates states of affairs which are both valid but at the same time incompatible. The expected outcome of the first state of affairs would be the negation of the other [4,15]. In Example 1, this means that 1) the basement is cold, but 2) there is still one small room that is isolated.

In Cluster 3, all the DRDs are connectives that signal addition, specification or exemplification [4]. While these coherence relations may at first seem different, in fact, they are very similar, as they all point to a member of a larger entity, either by adding a new one or by specifying one. The 15 important n-grams reflect various usage patterns related to these relations. Example 2 illustrates coordination in the *context* part of the two-sentence window from which the n-grams were generated. This coordination presents the elements from which the DRD in the next sentence specifies one. Example 3 shows another typical co-occurrence pattern of cluster DRDs, where they co-occur with other similar DRDs in the *context* part.

Table 1. Examples on usage contexts reflecting specific coherence relations. The examples include a lexicalized and a graph-based version of the illustrated n-gram, lexicalized examples of the n-gram and its translation. The tokens that are part of the n-gram are in bold.

Example 1. neg/2 ROOT/0 ccomp/2_hits  *Hän olisi tahtonut nähdä jotain kaunista, ei harmaata ja masentavaa. Tosin hän **ei** ollut **varma**, miten auringonpaiste olisi mihinkään **vaikuttanut**.* 'He would have wanted to see something beautiful, not grey and depressing. Although he was **not** su**r**e how sunshine would **have had on effect** on anything..'
Example 2. nsubj/2 cop/3 conj/0_context  *Ajatella, **lempisarjakuviani** ovat **olleet** Tintti, Lucky **Luke**, Mustanaamio, Aku Ankka (on se kyllä edelleenkin , mutta ei niin hyvä kuin virtanen), Korkeajännitys, Tex Willer ym ym. Erityisesti tämä strippi B. Virtasesta on jäänyt mieleen...* 'Think about it, my favorite **comics** have **been** Tintin, Lucky **Luke**, Phantom, Donald Duck (it is still my favorite, but not as good as virtanen), Commando, Tex Willer, etc etc. In particular I remember this strip from B. Virtanen...'
Example 3. advmod/2 nsubj/3 acl:relcl/0_context  *Kyseessä on runo, joten **myös käännöksen** on **toimittava** kohdekielisessä kulttuurissa runona. Erityisesti käytettävien sanojen ja runon tunnusmerkkien valinta on vaativaa.* 'It is a poem, so **also** the **translation has to work** as a poem in the culture of the target language. In particular, the choice of the words and poetic constructions is demanding.'
Example 4. nsubj:cop/3 ROOT/0 xcomp/2_hits  *Taloustutkimus Oy:n Suomi Syö 2010 -tutkimuksen mukaan 65 prosentille suomalaisista on tärkeää, että he syövät suomalaista ruokaa. Erityisesti **leivän**, maidon ja lihan **halutaan** olevan **kotimaista**.* 'According to the Finland Eats 2010 research done by Taloustutkimus Oy, it is important for 65% of the Finns that they eat Finnish food. In particular, they **want bread**, milk and meat to be **Finnish**.'

Finally, three out of the 10 n-grams with a positive association for this cluster include copular constructions and adjectives (see. Example 4). The most frequent lexicalizations of the verb are *tulla* 'must', *todeta* 'note' *kokea* 'experience' and of the predicative adjective *hyvä* 'good', *erinomainen* 'excellent', *vaikea* 'difficult', *helppo* 'easy'. These constructions thus express typically subjective expressions. This may suggest that these

DRDs are typically used to express *subjective* coherence relations. This, however, requires further investigations.

3.2 DRDs in spoken-like genres and written-like internet genres

The explanation for the distribution of the DRDs to clusters 1 and 4 can be found in the differences between spoken-like and written-like internet genres. Cluster 1 includes the coordinating conjunctions, the concessive *kyllä* 'certainly' and the causals *joten, niin, siis* that could all be translated as 'so'. Sentence-initial coordinating conjunctions have already in previous research been defined as typical of informal and spoken-like discourse, and referred to as *forbidden first words* [5,19]. The usage patterns reflected by the most important biarcs for this cluster support this.

Out of the 13 most important n-grams with a positive co-efficient value estimated by the SVM for cluster 1, seven include an initial coordinating conjunction and two a discourse particle. Examples 5 and 6 illustrate constructions, where these items are attached to the cluster DRDs. Example 5 begins with 'but hey', an interjection, and Example 6 with 'yeah and', where the first DRD presents the writer's reaction, which is explicitly linked to the previous discourse by 'and'. According to [4], these particles guide the interaction between the writer and the reader. These co-occurrence patterns thus suggest that the cluster DRDs are typically used in interactive settings. This is confirmed by further important n-grams, which reflect patterns that are not grammatically correct. In Example 7, the sentence-initial coordinating conjunction is repeated. The repetition of a word reflects intensity and affection [4]. In Example 7, the repetition of 'but' can be interpreted as hesitation intensifier. The repetition of words seems to be typical e.g. in spoken and literary contexts [4]. In standard written Finnish, it is very prominent feature, and in many written contexts and genres unacceptable.

In Example 8, the important n-gram and the 'goeswith' dependency denote orthographic errors. Specifically, it relates the parts of a construction that should be a compound noun. Orthographic errors refer to genres, which allow non-standardized forms and combination of spoken and written language, such as chat communication [11].

Table 2. Examples on the typical usage patterns of cluster 1.

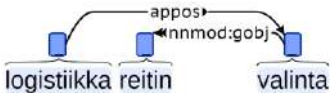
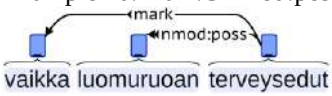
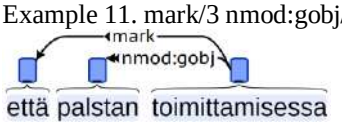
Example 5. discourse/2 ROOT/0 nsubj/2_hits  *Rahalla saa ja hevosella pääsee. Mutta **hei**, **saahan pojat** sinne mikron ja kahvinkeittimen, ja limpparia!* 'With money one gets, with a horse one goes. But h**e**y, the **boys** can **get** there a microwave and a coffeemaker, and soda!'
Example 6. cc/2 root/0 dobj/2_hits  *Niin, **ja** tässä auto ei **tarkoita taksia**.* 'Yes, and there car does not mean a taxi.
Example 7. cc/2 ROOT/0 dobj/2_hits  *Olen onnellisesti naimisissa, 8kk tytön äiti ja odottelemme toista lastamme jonka olisi määrä syntyä toukokuussa. Mutta **mutta**, **kaipailen** muita **äitiystäviä** joiden kanssa jakaa asioita ja vaikka kahvitella.* 'I am a happily married mother of an eight-month old girl, and we are waiting for our second child, who will be born in May. But **but**, I **miss** other **mothers** as friends with whom I could share things such as having coffee.'
Example 8. goeswith/2 dobj/3 ROOT/0_hits  *Kyllä tätä meidän pa**p**pi **asiaa** olikin **murehdittu** koko alkutalvi.* 'Yeah our **minister question** we have **worried** about it the whole winter.'

Finally, Cluster 4 presents a variety of DRDs. First of all, the important usage patterns include complex noun phrases ('nmod:gobj)', which have been associated with written discourse [2,3]. All of the Examples 9, 10 and 11 illustrate the co-occurrence of the cluster DRDs with these; Example 9 in the *context* part and Examples 10 and 11 in the *hit* part. This suggests that the cluster DRDs are typically used in written-like genres.

Another frequent usage pattern in the important n-grams of this cluster is subordinators and that-clauses ('mark'). In Example 10, the DRD marker of manner 'this is so' is attached to the concessive subordinator 'even if', and in Example 11, the causal DRD 'because' is followed by 'that' offering an explanation to the question presented in previous discourse. In both examples, the writer argues for something. The combination of the sentence-initial DRD to the other conjunctions seems to allow for more complex argumentative patterns than what would be possible with the DRD alone. In previous research, Biber [2] relates subordinators with persuasion, and Sanders [23] notes that subjective relations typically involve argumentation. Based on this, we can

ask whether the marking of subjective relations or argumentation would be typical of all the cluster DRDs. This will be explored during the presentation and in future work.

Table 3. Examples on the typical usage patterns of cluster 4.

Example 9. nsubj/0 nmod:gobj/3 appos/1_context    *Tärkeässä osassa on myös kuljetusyritysten **logistikk**a: **reitin valinta**, aikataulutus, lastaus ja purku tulisi suunnitella siten, että turhaa ajamista vältetään. Samoin ...* 'Also the carrier companys **logistics** is important: **choosing the route**, scheduling, loading and unloading. In the same way...'
Example 10. mark/3 nmod:poss/3 root/0_hits    *Kuluttajatutkimusten mukaan Suomessa puhtaus ja terveellisyys ovat tärkeimmät perusteet luomutuotteiden valitsemiseen. Näin **vaikka** ympäristöedut ovat paljon yksiselitteisemmin todetut kuin **luomuruoan terveysedut** ...* 'According to consumer surveys, in Finland cleanliness and healthiness are the most important motivations for selecting organic products. This is so, **even if** environmental benefits are reported much more unambiguously than **health benefits of organic food**...'
Example 11. mark/3 nmod:gobj/3 ROOT/0_hits    *Miksi ei blog, miksi blogumni? Siksi, **että** tässä oman **palstan toimittamisessa** keskusteluforumin rinnalla (...) on hyvä liittää muu nimitys.* 'Why not blog, why blog-column? Because **of the fact that having** a place of my own for writings in addition to the discussion forum (...) it is good to find an alternative name

4 References

1. Berez, A., Gries, S.: In defense of corpus-based methods: a behavioral profile analysis of polysemous get in English. In: Moran, S., Tanner, D., Scanlon, M. (eds.) Proceedings of the 24th Northwest Linguistics Conference. University of Washington Working Papers in Linguistics, vol. 27, pp. 157-166. Department of Linguistics, Seattle, WA (2009).
2. Biber, D.: Variation across speech and writing. Cambridge University Press, Cambridge (1988).
3. Biber, D.: Dimensions of register variation: A cross-linguistic comparison. Cambridge University Press, Cambridge (1995).

4. *The Comprehensive grammar of Finnish. (CGF)* Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T., Alho, I.: Iso suomen kielioppi [The comprehensive grammar of Finnish]. Suomalaisen Kirjallisuuden Seura, Helsinki: (2004).
5. Chang, Y., Swales, J.: Informal elements in English academic writing: threats or opportunities for advanced non-native speakers? In: Candlin, C., Hyland, K. (eds.) Writing: Texts, processes and practices, pp. 143–167. Longman, London (1999).
6. Divjak, D., Gries, S.T.: Ways of trying in Russian: clustering behavioral profiles. Corpus linguistics and linguistic theory 2(1), 23–60 (2006).
7. Firth, J.R.: A synopsis of linguistic theory 1930–1955. In Studies in linguistic analysis. Blackwell, Oxford. Reprinted in Palmer, F. (ed.) Selected papers of J. R. Firth (1952–59), pp. 168–205. Longman and Indiana University Press, London and Bloomington (1957).
8. Gries, S.: Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. In: Jarema, G., Libben, G., Westbury, C. (eds.) Methodological and analytic frontiers in lexical research, 57-80. John Benjamins, Amsterdam & Philadelphia (2012).
9. Goldberg, Y., Orwant, J.: A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books. Second Joint Conference on Lexical and Computational Semantics (*SEM), 1: Proceedings of the Main Conference and the Share d Task: Semantic Textual Similarity, pp. 241-247. Association for Computational Linguistics (2013).
10. Harris, Z.: Mathematical structure of language. Wiley, New York (1968).
11. Herring, S.C.: Grammar and electronic communication. The Encyclopedia of Applied Linguistics (2012).
12. Jääskeläinen, A., Koivisto A.: Konjunktio, partikkeli vai konnektiivi? [Conjunction, particle or connective?.] Virittäjä 116(4), 591-601 (2012).
13. Kanerva, J., Luotolahti, M.J., Laippala., V., Ginter. F.: Syntactic N gram collection from a large-scale corpus of Internet Finnish. Proceedings of the sixth international conference Baltic HLT, 184–191 (2014).
14. Kaufman, L., Rousseeuw. P.: Finding groups in data: An introduction to cluster analysis. John Wiley, New York (1990).
15. König, E.: Conditionals, concessive conditionals and concessives. Areas of contrast, overlap and neutralization. In: Closs Traugott, E., Meulen, A., Snitzer Reilly, J., Ferguson, C. (eds.) On conditionals, pp. 229–246. Cambridge University Press, Cambridge (1986).
16. Laippala, V., Kyröläinen, A-J., Ginter. F.: Dependency profiles in the large-scale analysis of discourse connectives. Corpus linguistics and linguistic theory (forthc.).
17. Levshina, N., Degand, L.: Just because: In search of objective criteria of subjectivity expressed by causal connectives. Dialogue and Discourse 2017(1), 132-150 (2017).
18. Luotolahti, M.J., Kanerva, J., Laippala, V., Pyysalo, S., Ginter, F.: Towards universal web parsebanks. Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), 211–220 (2015).
19. Makkonen-Craig, H.: The forbidden first word: Discourse functions and rhetorical patterns of AND-prefacing in student essays. Text & Talk 37(6), 713–734 (2017).
20. Mann, W.C., Thompson, S.: Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3), 243–81 (1988).
21. Mosegaard Hansen, M-B.: A comparative study of the semantics of enfin and finalement. Journal of French Language Studies 15, 153-171 (2015).
22. Redeker, G.: Ideational and pragmatic markers of discourse structure. Journal of Pragmatics 14, 305–319 (1990).
23. Sanders, T.: Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. Discourse Processes 24, 119–147 (1997).

24. Sanders, T., Spooren., W., Noordman, L.: Toward a taxonomy of coherence relations. Discourse Processes 15, 1–35 (1992).
25. Sanders, T., Spooren, W.: Causality and subjectivity in discourse: The meaning and use of causal connectives in spontaneous conversation, chat interactions and written text. Linguistics 53(1), 53-92 (2015).

Correlating DRDs with other types of discourse phenomena

Cross-linguistic analysis of the interplay between DRDs, coreference and bridging

Anna Nedoluzhko [1] and Ekaterina Lapshinova-Koltunski [2]

[1] Charles University, Prague, Chechia
[2] Saarland University, Germany
[1]nedoluzko@nedoluzko@ufal.mff.cuni.cz, [2]e.lapshinova@mx.uni-saarland.de

Abstract. The paper presents a cross-linguistic analysis of the interplay between several types of discourse phenomena such as relational local adverbs, conjunctions, multi-word discourse phrases, coreference and bridging relations. The analysis is based on an empirical corpus study of parallel texts containing transcribed TED talks. The selected dataset contains English original texts and their translations into German and Czech.

Keywords: parallel data, multilingual, translation, contrastive study, discourse, English, German, Czech, Russian.

1 Introduction

The present contribution describes a cross-linguistic analysis of the interplay between discourse-relational devices (DRDs) and other discourse-related phenomena, such as coreference and bridging relations. The difference between the phenomena under analysis lies in the type of relations, which is expressed by a corresponding device. DRDs express logico-semantic relations between propositions, such as contrast, time, addition and others). Coreference serves the task of linking identical objects or events (i.e. complex anaphors, see Zinsmeister et al. 2012) and bridging anaphora expresses non-identical or near-identical relations between referents, linking them with semantic interconnection.

All of them (DRDs, coreference and bridging) contribute to the construction of meaningful discourse. These phenomena exist in all languages, but their realisations depend on the different preferences that languages have (both systemic and context-based).

For instance, German pronominal adverb *dabei* (which is a fusion of the preposition *bei* and the definite article in Dative *dem*) in example (1) below can function as a

referring expression (*beim Betrügen* - ‘while cheating‘) expressing at the same time a temporal meaning. English does not have a direct equivalent for this form. So, the English corresponding example (which is the source) from our parallel dataset does not contain this kind of coreference chain. Another possible reading of *dabei* is the meaning of contrast/concession. Again, the English source does not have any explicit marker for this relation. However, the Czech translation (CZ) contains the connective *ale* (‚but‘) which has the meaning of contrast and in this case also concession.

(1)
EN: *We've learned that a lot of people can cheat.They cheat just by a little bit*.
DE: *Wir haben gelernt, daß viele Leute betrügen können. Der Einzelne betrügt [**dabei**] nur ein bißchen*.
CZ: *Zjistili jsme, že hodně lidí je ochotno podvádět. Podvádějí [**ale**] pouze po troškách.*

German pronominal adverbs like *dabei* in Example (1) often represent an interplay between DRDs, coreference and, in some cases, bridging. We aim at describing such cases and analysing various transformation patterns that are possible between German, English and Czech. The knowledge of these patterns is important for contrastive linguists, language learners and translators, as they have to be aware of the full range of linguistic options that exist in the analysed languages. In our analysis, we address German, English and Czech.

Since the analysis is performed on translations, we are also interested in the impact translation process may have on the choice of transformation patterns. For instance, in Example (1), the Czech translator decided to explicate the implicit contrastive meaning with a contrastive marker *ale*. The German translator prefers to use an ambiguous element *dabei*, which can be also interpreted contrastively but not obligatorily. In one scenario, this decision might have been influenced by the adverbial *just* in the English original sentence that can be transfered into German with *dabei nur* and express a contrastive meaning in this case. Another possibility is the correspondence of *just* with *nur*, so *dabei* is coreferential in this case, and it was inserted by the translator to create a link between the two sentences, for a stylistic purpose. The reason for the translator‘s choice remains unknown, as we do not have any information on the translation process. So, we want to find out the possible signals in the English source that trigger the usage of these explicit constructions in the corresponding translations into German and Czech. We assume that some cases are induced by language-specific constructions that do not have direct equivalents in a target language. Apart from that, they can be attributed to the phenomenon of explicitation (Blum-Kulka, 1986) or implicitation (Becher, 2011) which are specific for translation process.

2 Data and Methods

For our analysis, we use a corpus-based approach, extracting the corresponding data from a trilingual parallel corpus containing English original texts and their translations into German and Czech. This data is from the International Workshop on Spoken Language Translation (IWSLT) and contains TED talks. As the cases of German pronominal adverbs represent the interplay we are interested in, we extract the parallel sentences with these adverbs only. First, we compile a list of such adverbs (*daran, darauf daraus, dabei, dadurch, dafür, dagegen, dahinter, darin* and so on) using a grammar of German (Duden Online Wörterbuch). Then, we randomly select a number of parallel TED talks and extract the corresponding parallel sentences where the aligned German sentence contains one of the pronominal adverbs from the list. After that, we perform a manual alignment of the discourse phenomena in the sentence triples, e.g. connecting *dabei* with *ale* and zero as it was done for Example (1) above. Then, we manually analyse the created dataset for transformation patterns that reflect language differences and the impact of translations process (explicitation/implicitation). The findings are then further interpreted from the point of view of theories, trying to answer the following research questions: How are these discourse phenomena realised in the parallel data at hand? What are the transformation patterns across the three languages under analysis? Which are most frequent? What are the usage constraints / reasons for these realisations / transformations?

3 Analysis

Our analysis has been performed on 98 parallel sentences extracted from 10 random parallel TED talks (60849 ws, 1490 parallel sentences). The analysis of transformation patterns shows that in most cases (38), German differs from English and Czech (DE⇔EN+CZ). However, 22 sentence triples have the same structures in all the languages under analysis. English differs from Czech and German in 16 cases, whereas Czech differs from both Germanic languages less frequently (10 cases in our data). Seven sentence triples show cross-lingual discrepancies between all the three languages.

Analysis of corresponding triples shows that German sentences tend to be more explicit than their equivalents in Czech and English. Interestingly, out of 38 cases, where German differs from English and Czech, pronominal adverbs were not used in the source language and were used only in German in 35 cases (i.e. more than 90%), i.e. they were not inserted by the translators when translating into Czech, see Example (2). The rest three cases represent rewording, i.e. translations into German and a different syntactic structure, which caused the use of pronominal adverbs.

(2)
EN: *As soon as you win, suddenly stop*.
DE: *Sobald Sie gewonnen haben, hören Sie plötzlich [**damit**] auf.*
CZ: *Jakmile vyhrajete, zastavte*.

Examples where English differs from Czech and German (EN⇔DE+CZ) mostly represent English sentences (often with gerundial clauses) that are translated into target languages with correlative constructions (12 out of 16 cases). The rest four examples show the translation process effect: there are two implications and two explicitations there. Besides that, translation explicitation tends to occur when an English construction does not exist in the target languages, as in Example (3).

(3)
EN: *You know, I'm sick and tired of us not living up to our potential*.
DE: *Ich hab die Nase voll davon, dass wir unser ganzes Potenzial nicht nutzen*.
CZ: *Víte, mám dost toho, že nevyužíváme naplno svůj potenciál*.

The cases when Czech differs from both Germanic languages (10 instances, CZ⇔EN+DE) are not really structured. These include the cases where Czech examples are less explicit, i.e. pronominal reference has been translated from English into German but it was omitted in Czech (seven instances), see an illustration in Example (4). Another three examples show the opposite effect of explicitation: pronominal expressions are translated into Czech as full nominal groups. None of these cases contains correlative structure which is not surprising as correlative structures with pronominal adverbs are rather typical for Czech and German making them different from English.

(4)
EN: *People would get very excited about this when they read these articles*.
DE: *Die Leute waren begeistert darüber, als sie diese Artikel lesen*.
CZ: *Lidé velmi snadno podléhají nadšení, když čtou takové články*.

Table 1 summarises the results for how pronominal adverbs in German are mapped in English and Czech in the parallel TED talks.

Table 1. Realisation of German pronominal adverbs in English and Czech based on TED talks

Type and # in DE	Mapped to	EN abs.	EN in %	CZ abs.	CZ in %
anaphoric (63)	zero	26	41.27	26	41.27
	preposition + pronoun	27	42.86	21	33.33
	pronoun	2	03.17	5	07.94
	adverb	4	06.35	6	09.52
	NP	3	04.76	3	04.76
	pronominal adverb	1	01.59	2	03.17
correlative (26)	zero	26	100.00	15	57.69
	preposition + pronoun	0	00.00	11	42.31
connective (2)	connective	2	100.00	2	100.00
other (7)		not analysed[1]			

[1] Pronominal adverbs marked as other meanings are different cases of phraseologisations and accidental uses. They are not analysed further.

As seen from Table 1, pronominal adverbs used as a connective seem to always correspond to connectives both in the source and the other target language in our data (although their number in our data is very low). In most of the observed cases (52), the English source does not contain any structure corresponding to the German pronominal adverb (which is realized either as a referring expression or a correlative pronoun). The Czech translations seem to more often keep the English source (41 cases of zero) than the German translations. At the same time, pronominal adverbs in function of a referring expression are more commonly used when the English sentence contain a preposition and a pronoun (ca. 43% of all observed cases), which is not surprising. However, the number of realisations from zero to an anaphoric element (possibly explicitation) is also high (41%). Alternative forms, e.g. pronouns, adverbs, etc. trigger the German pronominal adverb less frequently (16% of all the observed cases), whereas in Czech, they more often serve as equivalents (ca. 25%). We find it surprising that while being more common than in English, pronominal adverbs are rarely used in the observed cases of the Czech translations. This might be explained by the fact that translators rather kept closer to the English sources.

These quantitative results, although delivering interesting information on the existing transformation patterns, are limited in several aspects. First of all, since we extracted the parallel sentences only, whose German part always contains a pronominal adverb, we are not able to observe all possible kinds of realisations that may happen in German translations. At the same time, this limitation was deliberately chosen, as this structure represent an interplay between several cohesive devices. Another problem is the limitation of the translation direction: we deal with English sources only, which means that we are not able to observe all possible equivalents of the German pronominal adverbs that would, for instance, appear in german were the source language. However, having just one source language allows us to make judgements about the interference or ``shining through'' effects (Teich, 2003) in translated texts, i.e. the structures of the source language having traces in the target texts.

4 Conclusion

Since translations are influenced by various factors of translation process, it is often difficult to explain the real reasons of a certain construction used in translation data. For instance, the usage may be constrained by shining through (influence of the source language, see Teich, 2003), normalisation, translator style, and other factors. So, we realise that a description of contrastive patterns often requires comparative data. At the same time, it is difficult to find comparative data required for such an analysis (with aligned discourse structures). Another shortback of our approach is the usage of one translation direction which can, again, be explained by practical reasons – it is difficult to find trilingual tridirectional translation data. However, we consider our study to be innovative, as there are no further studies on the interplay between different kind of discourse phenomena across languages known to us. The results of our analysis are valuable for both contrastive linguists and language learners. Besides,

the information on transformation patterns are of great interest for translation trainees and trainers, as well as translation scholars.

Our future work will include analyses on an extended set of parallel data with a more systematic description. We will also support a more detailed description of the observed cases, as well as extension of the analysed data. Besides that, we would like to have a look at texts translated from German into other languages to be able to make claims about equivalents of pronominal adverbs in these languages.

Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (project GA16-05394S).

References

1. Becher, V.: Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts. PhD thesis, Universität Hamburg (2011).
2. Bisiada, M. Lösen sie Schachtelsätze möglichst auf: The impact of editorial guidelines on sentence splitting in german business article translations. Applied Linguistics, 3 (2014).
3. Blum-Kulka, S. Shifts of Cohesion and Coherence in Translation." Juliane House, Shoshana Blum-Kulka (eds): Interlingual and Intercultural Communication. Tübingen: Narr, 17-35 (1986).
4. Duden Online Wörterbuch, https://www.duden.de/woerterbuch, last accessed 2017/12/1.
5. Li, J. J., Carpuat, M., Nenkova, A.: Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classification system. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 577–587, Dublin, Ireland (2014).
6. Meyer, T., Webber, B.: Implicitation of discourse connectives in (machine) translation. In: Proceedings of the Workshop on Discourse in Machine Translation, pp. 19–26, Sofia, Bulgaria. Association for Computational Linguistics (2013).
7. Prasad, R., A. Joshi, Webber, B.: Realization of discourse relations by other means: Alternative lexicalizations. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1023–1031, Beijing, China (2010).
8. Teich, E.: Cross-Linguistic Variation in System und Text. A Methodology for the Investigation of Translations and Comparable Texts. Mouton de Gruyter, Berlin (2003).
9. Zinsmeister, H., Dipper, S., Seiss, M.: Abstract pronominal anaphors and label nouns in german and english: selected case studies and quantitative investigations. Translation: Computation, Corpora, Cognition, 2(1) (2012).
10. Zikánová Š., Hajičová E., Hladká B., Jínová P., Mírovský J., Nedoluzhko A., Poláková L., Rysová K., Rysová M. Václ, J.: Discourse and Coherence. From the Sentence Structure to Relations in Text. Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Praha, Czechia (2015).

Annotating Discourse Markers in the MULTINOT corpus:
The case of elaborating connectives in English and Spanish

Julia Lavid and Estefanía Avilés
Universidad Complutense of Madrid

The study of discourse markers (DM) in the context of translation is crucial due to the idiomatic nature of these structures (Aijmer 2007, Beeching 2013]. In the field of Machine Translation (MT), and more precisely Statistical Machine Translation (SMT), recent work has pointed out the need for findings and studies that address divergences in DM usage in order to improve SMT output quality (Steele 2015). Current SMT systems often focus on translating single sentences with clauses being treated in isolation, leading to a loss of contextual information, ignoring the fact that DMs are vital contextual links between discourse segments and that they are often translated in ways that differ from how they are used in the source language (Hardmeier, 2012; Meyer and Popescu-Belis, 2012). In addition, although an extensive literature has already reported language-specific traits of these events (Fraser 1990, 1999; Beeching and Detges 2014; Fisher 2000; Ghezzi and Molinelli 2014, *inter alia*), there are no systematic studies which address their cross-language behavior in the context of translation between English and Spanish. The current study is a preliminary step in the context of a larger project aimed at the creation of a bilingual (English-Spanish) corpus annotated with DMs as part of the activities of the Textlink Cost Action. Focusing on elaborating connectives (ECs) as a case study, the paper addresses the following research questions: what are the explicit relationships between different subtypes of ECs in English and Spanish? How do semantic fields of ECs in English and Spanish relate to one another? Are there genre- specific uses of ECs in these two languages? The theoretical tools used are the classifications proposed in the Systemic-Functional approach (Halliday and Matthiessen 2004) for the English elaborating connectives, and the typologies on reformulation and exemplification markers proposed in the Spanish linguistic community (Portolés, J., MA Martín Zorraquino 1999; del Saz 2006, Cuenca 2001, *inter alia*). These include 'appositive' (i.e. expository and exemplifying) and 'clarifying' connectives (i.e. corrective, distractive, dismissive, particularising, resumptive, summative and verificative).

The sample used for the study consists of a total of two hundred texts, divided into two directional pairs from five different domains of the bilingual English-Spanish MULTINOT Corpus (Lavid et al 2015), i.e.: fiction, essays, expository, legal procedures from webpages and speeches. The methodology used consisted of the alignment of the source and the target texts and the annotation of the translation correspondences between the ECs occuring in original texts and their translations, looking at the meaning of these connectives as mirrored in their bidirectional translations (Dyvik 1998). The results of the annotation point to some general translation correspondences between ECs in English and Spanish which describe their paired lexico-semantic fields. They also show some genre-specific preferences in the use of these connectives, as a result of the different communicative purposes of the texts where they are used. Future work will focus on investigating translation correspondences by annotating more texts not only from the MULTINOT corpus but also from other parallel sources, including not only texts from the written medium but also from the spoken one.

References

Beeching, K. (2013): A parallel corpus approach to investigating semantic change ,

Advances in corpus-based contrastive linguistics, pages 103–125. Studies in honour of Stig Johansson. John Benjamins, Amsterdam,

Beeching, K. and Ulrich Detges, editors (2014): Discourse Functions at the Left and Right Periphery. *Crosslinguistic Investigations of Language Use and Language Change*. Brill, 2014.

Cuenca, MJ. (2001): Anàlisi contrastiva dels marcadors de reformulació i exemplificació. *Caplletra* 30 (2001), pp. 47-72.

Dyvik, H.: A Translational Basis for Semantics. In: Johansson, Oksefjell (eds.): *Corpora and Crosslinguistic Research: Theory, Method and Case Studies*, Rodopi, pp. 51–86 (1998).

Fischer, K (2000): *From Cognitive Semantics to Lexical Pragmatics: The Functional Polysemy of Discourse Particles*. Mouton de Gruyter, Berlin/New York, 2000.

Fraser, B. (1990): An approach to discourse markers. *Journal of Pragmatics*, 14:383–395, 1990.

Fraser, B. (1999): What are discourse markers? *Journal of Pragmatics*, 31:931–952, 1999.

Ghezzi, C. and Piera Molinelli, ed. (2014): *Discourse and Pragmatic Markers from Latin to the Romance Languages*. Oxford Studies in Diachronic and Historical Linguistics 9. Oxford University Press, 2014.

Hardmeier, C. (2012) *Discourse in Statistical Machine Translation: A Survey and a Case Study*. Elanders Sverige, Sweden.

Halliday, Michael, and Matthiessen, M.I.M (2004): *Introduction to Functional Grammar*. Hodder Headline Group. London, Great Britain.

Portolés, J., MA Martín Zorraquino (1999), "Los marcadores del discurso", en Bosque, I. y Demonte, V. (dirs.), *Gramática descriptiva de la lengua española*, Madrid, Espasa Calpe, capítulo 63.

Lavid, Julia, Arús, Jorge, DeClerck, B and Hoste, Veronique (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, pages 249–256.

Meyer, T. and Andrei Popescu-Belis (2012) Using sense-labelled discourse connectives for statistical machine translation. In: *EACL Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT* (ESIRMTHyTra), pages 129-138.

Steele, D. (2015): Improving the Translation of Discourse Markers for Chinese into English. *Proceedings of NAACL-HLT 2015 Student Research Workshop* (SRW), pages 110–117, Denver, Colorado, June 1, 2015. Association for Computational Linguistics.

Discourse relations with explicit and implicit arguments: The case of European Portuguese *aliás*

Pierre Lejeune[1] and Amália Mendes[2]

[1] Faculdade de Letras da Universidade de Lisboa
[2] Centro de Linguística da Universidade de Lisboa
lejeunepierre@hotmail.com, amaliamendes@letras.ulisboa.pt

Abstract. We analyse the discourse values of the Portuguese discourse marker *aliás* (besides/indeed) and focus on cases where a pragmatic implicature arises that involves an implicit argument. We discuss the challenges it poses for discourse annotation, namely the PDTB-style annotation. We further contrast the values of the Portuguese DM with its counterparts in English by extracting contexts from the Europarl corpus.

Keywords: discourse markers, pragmatic implicature, discourse annotation.

1 Introduction

Within the PDTB annotation scheme, little attention has been given so far to discourse markers (DMs) that do not necessarily link directly two explicit discourse segments Arg2 («the argument that appears in the clause that is syntactically bound to the connective» [1]) and Arg1 («the other argument») but may include an instruction to recover retroactively a pragmatic implicature of Arg 1.[1] Yet those DMs contribute to the construction of discourse meaning, and as such cannot be excluded from annotation as mere markers of modality.

Portuguese *aliás* (rough French and English equivalents: *d'ailleurs*; *besides/indeed*) belongs to that category of DMs.

Working with excerpts from two corpora, one monolingual (CRPC - Corpus de Referência do Português Contemporâneo[2] (Généreux *et al.* [15]) and one multilingual (Europarl[3]), we will try to reach a unified description of *aliás* that accounts for the

[1] Forbes-Riley *et al.* [2] mention that issue for the adverbials *actually*, *in fact* and *surprisingly*, questioning whether they can be treated as connectives.
[2] http://alfclul.clul.ul.pt/CQPweb/crpcfg16/
[3] As far as comparative linguistic analysis is concerned, Europarl has two important drawbacks: references to the original language are not systematic and it is not guaranteed that the other languages versions are direct translations, as English is often used as a pivot language. In our examples, Portuguese is not necessarily the original language. We need to confirm our findings with checked true translations.

diversity of its uses as a DM, after which we will discuss what treatment it should be given within the PDTB annotation system.

2 Discourse values of *aliás* and implicit arguments

In all of its usages, *aliás* seems to have the propriety described by Hannay *et al.* [3] for *besides* (after Traugott [4]) of «signalling an 'afterthought'», «entailing two properties: finality (since the afterthought comes after all preceding considerations) and tangentiality (since the afterthought is not integrated into the conceptual structure that has gone before)». This corresponds to the description of *d'ailleurs* by Luscher [5] («*D'ailleurs* signals that an utterance that was first presented as complete, has to be reevaluated as part of a whole» - referring not only to its explicit content but also to its contextual implicatures), and by Paillard [6], who, based on the meaning of *ailleurs* (somewhere else, neutralizing the contrast between *ici* and *là*) sees the segment introduced by *d'ailleurs* as a sort of final word neutralizing the alterity between what was asserted in the first place and an implied complementary viewpoint.

The only usage of *aliás* which seemingly does not correspond to French *d'ailleurs*, is **self-correction** (I have said X, I should have said Y):

(1) No entanto *o Mestre não é francês* (Arg 1), ALIÁS não é natural de lado nenhum ou talvez se pudesse dizer que não tem naturalidade nenhuma (Arg 2). (CRPC)
However, the Master is not French, rather he is native of nowhere or perhaps he has no place of birth. (our translation)

In this kind of example, which might be processed in PDTB 2 as Expansion/ Substitution/Arg2-as-substitute (at speech act level), *aliás* can be considered as a connective.

Aliás can also have an **argumentative value**, signalling a piece of evidence that reinforces a thesis. Like additive connectives such as *além disso* (*furthermore*), *aliás* does not introduce a first argument, but contrary to what happens with those connectives, with *aliás* the former arguments may be implicit. This argumentative value would be processed in PDTB 2 as Contingency / Cause + Belief / Reason, with the thesis as Arg 1 and the piece of evidence as Arg 2, so failing to capture the existence of (an) implicit argument(s). In (2), a first argument (the fact that the Swedish principle of transparency strengthens democracy) is recoverable from the earlier context. The English version does not render that part of the meaning.

(2) Na Suécia, temos um princípio de acesso aos documentos que constitui um reforço da democracia e gera um bom clima de diálogo entre os cidadãos, os decisores e as autoridades. *Gostaríamos muito que a UE também adoptasse este princípio* (Arg 1) que, ALIÁS, está consagrado no Tratado de Amesterdão (Arg 2). (Europarl)
We have a principle of transparency in Sweden which strengthens democracy and ensures that there is a worthwhile dialogue between

citizens, decision makers and authorities. We are very anxious indeed that the EU, too, should move in this direction, and this is also stated in the Treaty of Amsterdam.

In this other example of the Cause + Belief / Reason relation, the specific contribution of *aliás* to discourse meaning is to make the argument from authority superfluous, forcing the contextual recovery of (an) implicit one(s) (this element is absent from the English version).[4]

(3) No tocante às acções de controlo - uma questão fundamental - *devemos começar por examinar os resultados da aplicação da legislação existente* (Arg 1), como ALIÁS afirmou o senhor deputado Hatzidakis (Arg 2). (Europarl)
As for the inspections, very much a key question, the first thing we have to consider is how the current legislation has worked, as Mr Hatzidakis has said.

The main non-argumentative uses of *aliás* belong to the broad category of what Lopes (2014) calls **parenthetical comments**, which in PDTB would be treated as Expansion / Conjunction, missing here again on an important contextual effect: the retroactive blocking of inferences drawn from Arg1. In (4), the blocked inference is that the attention paid is only incidental.

(4) *Prestaremos portanto muita atenção à redacção da acta* (Arg1). ALIÁS, prestamos sempre (Arg 2). (Europarl)
We shall pay particular attention to the wording of the Minutes, as we always do, of course.

This specific inferential role of *aliás* is particularly visible when it coexists with a conjunction connective as in (5) (blocked inference here: the only intolerable accident we are thinking of is that of the Erika).

(5) É verdade, *este acidente* do Erika (Arg 1), como ALIÁS o do navio russo na Turquia (Arg 2), é inaceitável e intolerável no momento em que a alta tecnologia está no zénite. (Europarl)
Indeed this disaster involving the Erika, like that of the Russian vessel in Turkey, moreover (*sic*), is unacceptable and intolerable at a time when the ultimate hi-tech technology is available.

In (4) and (5) *aliás* blocks inferences that narrow the scope of Arg 1. Sometimes it is the contrary, as in (6) (blocked inference: the Flechard affair has already been sorted out):

(6) Os arquivos da Comissão deixam muito a desejar. Demo-nos conta do facto quando *tivemos de investigar o caso Flechard* (Arg 1), cuja investigação, ALIÁS, ainda não terminou (Arg 2). (Europarl)

[4] Another possible treatment possible in PDTB would be to annotate this example as attribution: this just shows that the discourse relation and the attribution layer are intertwined.

The Commission' s records leave a great deal to be desired. We noticed this when we had to investigate the Flechard affair, which, as it happens, has still not been sorted out.

It can be seen from those argumentative and parenthetical uses of *aliás* that its annotation in PDTB as a marker of a cause or conjunction relation does not account for the triggering of new inferences or the blocking of existing ones, which have a significant impact on the general discourse meaning, and, as such, should be taken into account when computing that meaning, arguably one of the main outputs of DMs annotation.

We also believe that this quite common phenomenon, which applies to a number of adverbial DMs, cannot be discarded when it comes to trying to find cross-linguistic equivalents of connectives. The same could be said of translations, which may risk neutralizing either the contextual effects of *aliás* (examples 2 and 3) or the difference between *aliás* and other additive connectives, as in (5), where *como aliás* does not link two arguments in favour of a thesis but two statements (one restrictive, the other more generic), or (7), where the first argument is implicit, making in both cases *moreover* look inappropriate:

(7) Com efeito, *está previsto, por escrito, que a Comissão mantenha o controlo e a orientação central do novo sistema* (Arg 1). ALIÁS, o relatório von Wogau, que o Parlamento Europeu acaba de aprovar, encoraja-a neste sentido. (Arg 2) (Europarl)
Indeed, the White Paper envisages that the Commission will retain the supervision and central direction of the new system. MOREOVER, the von Wogau report, which the European Parliament has just voted on, supports it in so doing.

3 Conclusion

The analysis of the contexts of the Portuguese DM *aliás* and its counterparts in English highlight argumentative and parenthetical uses that trigger or block inferences and pose a challenge to discourse annotation. We plan to enlarge our analysis to other DMs that have similar properties, and to further explore the concept of implicit arguments, in a contrastive approach to English.

References

1. PDTB Research Group: *The Penn Discourse TreeBank 2.0 annotation manual*. https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf (2008).
2. Forbes-Riley, K., Webber, B., Joshi, A.: Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. Journal of Semantics 23, 55-106 (2006).
3. Hannay, M., Martínez, Caro E., Mackenzie, J.L.: *Besides* as a connective. In: Gómez González, MA, Ruiz de Mendoza Ibáñez, F., Gonzálvez García, F., Downing, A. (eds) The

Functional Perspective on Language and Discourse: Applications and Implications, pp 223-242. John Benjamins, Amsterdam (2014).
4. Traugott, E. C.: The role of the development of discourse markers in a theory of grammaticalisation. Paper presented at ICHL XII, Manchester, 1995. http://www.wata.cc/forums/uploaded/136_1165014660.pdf (1997).
5. Luscher, J. M.: Connecteurs et marques de pertinence. L'exemple de *d'ailleurs*. Cahiers de Linguistique Française 10, 101-145 (1989).
6. Paillard, D.: *D'ailleurs* ou comment enchaîner l'un à l'autre: essai de traitement lexicologique. *Le Gré des Langues* 2, 60-66 (1991).
7. Lopes, A. C. M.: *Aliás*: a contribution to the study of a Portuguese discourse marker. In: Molinelli, P., Ghezzi, C. (eds) Discourse and Pragmatic Markers from Latin to the Romance Languages, pp 211-22. Oxford University Press (2014).
8. Aijmer, K.: The actuality adverbs *in fact*, *actually*, *really* and *indeed* - establishing similarities and differences. In: Edwardes M (ed) Proceedings of the BAAL Conference 2007, pp 111-12. Scitsiugnil Press, London (2008).
9. Ducrot, O. e*t al.* : *D'ailleurs* ou la logique du camelot. In: Ducrot, O. e*t al.* (eds) Les mots du discours, pp 193-232. Les Éditions de Minuit, Paris (1980).
10. Fraser, B.: An account of discourse markers. International Review of Pragmatics 1,1–28 (2009).
11. Lopes, A.C.M. : *Aliás*: contribution à l'étude diachronique d'un marqueur du discours du Portugais. In: Borreguero-Zuloaga, M., Gómez-Jordana Ferary, S. (eds) Marqueurs du discours dans les langues romanes: une approche contrastive, pp 345-354. Lambert-Lucas, Limoges (2014).
12. Modena, S. : L'emploi du connecteur argumentatif *d'ailleurs* dans un discours politique. Le Français moderne 2, 263-271 (2009).
13. Plag, C., Loureiro, A.P., Carapinha, C.: Traduções alemãs do marcador *aliás*: uma análise do corpus Europarl. In: Plag, C., Loureiro, A.P., Carapinha, C. (eds) Marcadores Discursivos E(M) Tradução. Imprensa da Universidade de Coimbra (2017).
14. Ponce de León, R., Duarte, I.M.: *Aliás / alias*: diferencias de empleo en portugués y en español. In: Delbecque, N., Delport, M-F., Michaud Maturana, D. (eds) Du signifiant minimal aux textes : études de linguistique ibéro-romane, pp 137-152. Lambert-Lucas, Limoges (2013).
15. Généreux, M., Hendrickx, I., Mendes, A.: A Large Portuguese Corpus On-Line: Cleaning and Preprocessing. In Caseli, H. et al. (eds.) Computational Processing of the Portuguese Language. Proceedings of the 10th International Conference PROPOR1012, pp. 113-120. Berlin-Heidelberg, Springer-Verlag (2012).

Testing the interoperability of annotation systems for oral DRDs in Spanish language

Elena Pascual Aliaga[1]

[1] Universitat de València, Valencia 46010, Spain
Elena.Pascual@uv.es

Abstract. This paper examines the extent to which two of the annotation systems included in the TextLink COST Action framework can be successfully synthesized in order to enable the analysis of Discourse Relational Devices across different corpora. The paper begins by summarizing previous research on the compatibility of annotation proposals by Crible and Degand and Briz and Pons before outlining the results of a study in which a combined annotation proposal was applied to the analysis of DRDs in Spanish.

Keywords: interoperability, annotation systems, DRD, Spanish conversations

1 Introduction

This study aims to test the interoperability of two annotation systems for Discourse Relational Devices (henceforth DRD) included in the TextLink COST Action framework, namely the proposals set out by Crible and Degand (2017a) and Briz and Pons (2010). Both annotation systems are designed for the analysis of spoken discourse but offer differing approaches for the functional classification of DRD. Until now, both annotation systems have tended to be applied to distinct languages and corpora. For example, the annotation scheme by Crible and Degand (2017a) has been tested in samples of spoken corpora in languages such as French, English, Polish (see Crible and Degand 2017a, 2017b) and Spanish (Broisson in prep.), and a previous version has been applied to the *DisFrEn* corpus (see Crible 2017), a dataset containing several spoken genres in French and English. The annotation approach pursued by Briz and Pons (2010) has been applied mainly to a sample of the *Corpus Val.Es.Co. 2.0* (Cabedo and Pons 2013), which contains spoken Spanish conversations, and has been tested additionally in other languages, such as Italian (Scivoletto in prep.), in other discourse genres, such as humoristic monologues (Ruiz 2013) and semi-formal interviews (Espinosa and García 2017, Pose 2015), and in other areas of study, such as grammaticalization (Salameh, Estellés and Pons in press, Pons in press, Pons 2014, among others) and computer-mediated communication (Romero, in prep.).

Despite the divergences, the interoperability of the systems set out by Crible and Degand and Briz and Pons seems viable (see Pascual and Crible 2017). The two proposals are broadly complementary in nature: one is more fine-grained and focuses on a word-level analysis of DRD (Crible and Degand 2017a), while the other is more

sequential and oriented toward abstract representations of discursive functions and discourse units (Briz and Pons 2010). In Crible and Pascual (2017) a preliminary project for a merged protocol was designed and applied to a 33,000 word sample of English, French and Spanish spontaneous conversations. This common proposal combined the schemes designed by Briz and Pons (2010), Crible (2017) and Pascual (2016) for annotating discourse markers, speech disfluencies and discourse units. This present study is intended to develop further the earlier attempt to bridge the gap between these two models. It focuses exclusively on the annotation of DRD in a specific language and discourse genre, namely Spanish spoken conversations. The ultimate goal is to facilitate the contrastive and cross-linguistic analysis of DRD across different corpora.

The overview of the study presented here begins in section two with a summary of the differences between the annotation schemes by Crible and Degand (2017a) and Briz and Pons (2010). Section three reports on the preliminary conclusions obtained in previous studies by Crible and Pascual (2017) and Pascual and Crible (2017). In the final part of this overview, section four, a brief description of the work carried out in the present study is presented.

2 Two different models for annotating DRD

2.1 The independent domain and function annotation scheme for spoken language by Crible and Degand (2017a)

The proposal by Crible and Degand (2017a) implements the previous annotation model designed by Crible (2017). In Crible's (2017) model, DRD are defined from a functional perspective as a:

> **grammatically heterogeneous, syntactically optional, multifunctional type of pragmatic markers**. Their specificity is to function on a metadiscursive [omitted reference] level as **procedural** cues to constrain the interpretation of the host unit in a co-built representation of ongoing discourse. They do so by either signaling a **discourse relation** between the host unit and its context, expliciting the **structural sequencing** of discourse segments, expressing the speaker's **meta-comment** on their phrasing, or contributing to **the speaker-hearer relationship**. (Crible 2017: 58)

Crible and Degand (2017a) employ a word-level method where some syntactic and pragmatic features of DRD tokens are independently annotated. Regarding the syntactic features, three variables are annotated: the grammatical class (part of speech) of DRD, their co-occurrence (in terms of syntactic contiguity) and their position. This last positional variable is split into three sub-variables according to three different syntactic units:

- the whole dependency syntactic structure, in which case the unit that constitutes the scope of the DRD is identified and its status is defined in relation with the pred-

icate of the main clause (the possibilities are pre-front field, initial field, middle field, end field, post-field, independent or interrupted);
- the minimal syntactic unit in which the DRD is located, generally the clause (initial, medial, final, interrupted and independent);
- the turn (turn-initial, turn-medial, turn final and the whole turn).

Regarding the pragmatic features of DRD, Crible and Degand (2017a) revise the functional taxonomy of the previous annotation model by Crible (2017). In their latest proposal (cf. Crible and Degand 2017b), which is still a work in progress, the authors identify four generic functions or domains[1]: ideational (the relations between real-world events), rhetorical (applies to the relations between speech-act events, to the speaker's subjectivity and to metadiscursive effects), sequential (the structuring of discourse segments) and interpersonal (the interactive management of the speaker-hearer relationship).

The four domains comprise eleven specific functions (see Crible and Degand 2017b): addition, alternative, cause, concession, condition, consequence, contrast, punctuation, temporal and specification. Some of these are derived from taxonomies from the PDTB 2.0 (Prasad et al.2008) and González (2005) and can apply to any of the four domains, as shown in Table 1:

Table 1. Domains and functions (taken from Crible and Degand 2017b)

Ideational	**Rhetorical**	**Sequential**	**Interpersonal**
[addition] [alternative] [cause] [concession] [condition] [consequence] [contrast] [punctuation] [specification] [temporal] [topic]			

A given discourse relational device such as *mais* (but) can perform the contrastive function in each of the four domains: ideational (1), rhetorical (2), sequential (3) and interpersonal (4), as is illustrated in Crible and Degand (2017a):

1. nous sommes animés par le désir de participer à notre échelle au progrès de la connaissance **mais** nos liens avec l'université sont aussi fragiles

   *we are moved by the desire to participate at our own scale to the progress of knowledge **but** our links with the university are fragile too*

2. parce que je vois encore de la poésie en cinquième ce qui peut paraître classique **mais** enfin c'est comme ça que je voulais subdiviser le cours

[1] The domains are inspired by the work of González (2005), Halliday and Hasan (1976), Redeker (1990) and Sweetser (1990).

4

*because I do poetry again in the fifth year which can seem classic **but** well I wanted to divide the class like that*

3. Speaker 1: euh j'aime les néologismes j'aime les régionalismes mais euh je mets le point d'exclamation dessus euh pour dire euh attention
   Speaker 2: **mais** la norme qu'est-ce qu'est-elle pour vous

   Speaker 1: uh I like neologisms I like regionalisms but uh I write an exclamation mark on them uh to say uh careful
   *Speaker 2: **but** the norm what is it to you*

4. Alors cet auditeur vigilant il va vous dire tiens euh encore Jean d'Ormesson **mais** on entend Jean d'Ormesson à chaque automne

   *well this careful listener he will tell you look uh Jean d'Ormesson again **but** we hear Jean d'Ormesson every fall*

2.2 The Val.Es.Co. system of discourse units for spoken conversations (Briz and Pons, 2010)

Briz and Pons (2010) analyse the relationship between discourse markers and discourse units, following the system of conversational units by Briz and Val.Es.Co. group (2014). There are three main syntactic and pragmatic variables that are analysed in order to describe DRD[2] and to systematize the various functions they perform in discourse:

- the type of structural unit that a DRD constitutes, namely an act, subact and part of a subact;
- the position of the DRD, which can be initial, medial, final or independent;
- the unit over which DRD have scope; the Val.Es.Co. model distinguishes a total of eight units (see Briz and Val.Es.Co. group 2014): discourse, dialogue, exchange, adjacency pair, turn, intervention, act and subact.

Regarding the first of the variables, the type of unit, there are two relevant discourse units for the structural description of DRD, the act and the subact. These units can be described as follows (see Briz and Val.Es.Co. group [2014: 36-60] for a more detailed description): The act is a monological discourse unit that expresses an action and an intention. It has two main properties: independence (it conveys an isolable illocutionary force that can stand alone in a speaker's intervention) and identificability

[2] The definition of discourse relational devices given by Briz and Pons (2010) takes as a basis Martín Zorraquino and Portolés's (1999) work and is laid out in previous work by the same authors (Briz 1998, 2006; Pons 1998, 2001) and in the approach pursued in the *Diccionario de Partículas Discursivas del Español* (DPDE), coordinated by Briz, Pons and Portolés (2008).

in a given context (owing to formal linguistic boundaries). In the transcription proposed in the Val.Es.Co. model, acts are enclosed within the number sign #.

Subacts are the monological structural units that make up an act. They are segments of information that can be identified by their semantic and prosodic features (Briz and Val.Es.Co. group 2014: 53). Subacts are enclosed within curly braces {} in the transcription. Subacts are informative segments that can present either propositional or extrapropositional content. In the first case, they are considered Substantive Subacts and in the second case, Adjacent Subacts. This distinction leads to a more specific typology of subacts depending on the kind of propositional content they convey (Briz and Val.Es.Co. group 2014: 57):

- Director Substantive Subacts (DSS) convey primary propositional information (narrative, descriptive, argumentative); since they carry the illocutionary force of the act in which they are found, they are nuclear and constitute independent informative segments;
- Subordinate Substantive Subacts (SSS) and Topicalized Subordinate Substantive Subacts (TopSSS) convey secondary information (cause, condition, consequence, finality, temporal, locative, topicalization) and are dependent on DSS;
- Adjacent Subacts, which constitute the category under which discourse markers fall, convey extrapropositional information; they can, in turn, be of three types: Textual Adjacent Subacts (TAS), when they organize and distribute the flow of speech; Modal Adjacent Subacts (MAS), when they introduce modal information from the perspective of the speaker; and Interpersonal Adjacent Subacts (IAS), when they manage the relationship between speaker and hearer.

The different types of subact posited by the model are illustrated in Figure 1:

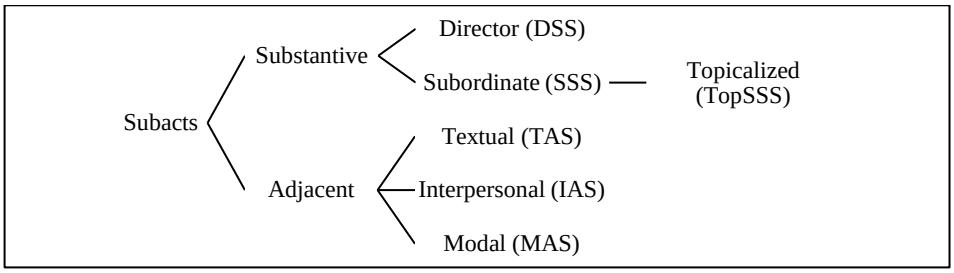


Fig. 1. Types of subact

According to Briz and Pons (2010), a DRD can potentially be: (5) a subact constituting in isolation an act, (6) a subact constituting an act together with other subacts, or (7) part of a subact. The following examples taken from Briz and Pons (2010) illustrate the three different categorical possibilities:

5. A1: # ¿te vienes en mi coche, María? #
   B1: # claro #

A: are you coming in my car, María?
B: sure

6. A: # Déjame el ordenador #
   B: # {No puedo dejártelo} {porque lo necesito} #

   A: Lend me your computer
   B: I can't lend it to you because I need it

7. A: # vas a venir, ¿no? #
   B: #{bueno} {no lo sé} #

   A: you are coming, aren't you?
   B: well I don't know

Briz and Pons (2010) observe that DRD that form part of subacts (*porque* in example 2) correspond mainly to the category of syntactic conjunctions and would be excluded from a pragmatic analysis, since they establish dependence relations that syntax is able to describe. However, DRD that are subacts (*bueno* in example 3) and acts (*claro* in example 1) are considered connectives, modalizers and regulatory elements that are associated with three possible functions: connection and text structuring (related to the organization and distribution of information, in which case they would be catalogued as Textual Adjacent Subacts); modality (related to the expression of the speaker's point of view, performed by the Modal Adjacent Subacts); and interpersonality (related to the speaker-hearer interaction functions, performed by the Interpersonal Adjacent Subact). Figure 2, adapted from Briz and Pons (2010), synthesizes the structural possibilities of DRD as regards the degree to which they constitute a discourse unit from the Val.Es.Co. system:

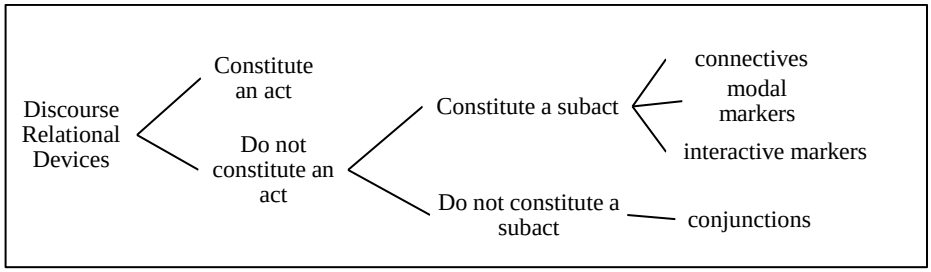


Fig. 2. Relationship between DRD and discourse units

The discourse unit constituted by a DRD taken together with its position and the unit over which it has scope gives a closed set of possible combinations that are highly relevant to predict the functions performed by DRD, giving thus a clearer picture of the multifunctional character of these discursive elements. For example, Table 2 shows a grid where the different functions that *bueno* ('well') can perform in conversational discourse are systematized according to different units and positions:

Table 2. Units of scope and positions of *bueno* (taken from Pons 2015)

Unit / Position	Subact	Act	Intervention		Dialogue	Discourse
			Initiative	Reactive		
Initial	Reformulation	Stress	Ø	Agreement/disagreement	Topic shift	Absolute beginning
Medial	Ø	Formulation	Ø		Ø	
Final	Ø	Stress/hedging	Ø		Ø	
Independent	Ø	Ø	Agreement/disagreement		Ø	

The hypothesis that the discourse unit and the position of a DRD are instrumental in determining the function of a given DRD has been borne out by many studies of Spanish DRD, particularly in diachronic and grammaticalization descriptions of DRD (see, for example, Briz 2006; Briz and Estellés 2009; Montáñez 2015, 2007; Pons in press, 2008; Pons and Salameh 2015; Pons and Estellés 2014; Salameh, Estellés and Pons in press; Salameh 2014).

While the grammatical class of DRD is not the focus of attention in Briz and Pons (2010), the co-occurrence of DRD is highly relevant in their work: the position of a given DRD and the discourse unit it constitutes are two factors that, taken in conjunction, give a clearer picture of the phenomenon of the co-occurrence of DRD. Pons (2008: 157-158) shows that discourse units set boundaries that allow us to distinguish, on one hand, the simple adjacency of two discursive functions carried out by two distinct DRD from, on the other hand, the recurring patterns in which certain DRD are combined and function jointly.

3 Towards a common annotation protocol: previous research by Crible and Pascual (2017) and Pascual and Crible (2017)

The proposals by Crible and Degand (2017a) and Briz and Pons (2010) share a functional perspective for the definition of DRD. Both approaches identify DRD in terms of their general metadiscursive function, that of guiding the interpretation of discourse. Both models take into account syntactic, semantic and pragmatic features for the definition of DRD that are well established in the literature, such as their syntactic marginality and optionality, their fixed form and their procedural meaning.

The main difference underlying these two approaches is the perspective they adopt in the annotation procedure. Whereas Crible and Degand (2017a) employ a word-

level method where DRD tokens are independently annotated, Briz and Pons (2010) adopt a wider and more sequential perspective that takes into account discourse units. Briz and Pons's (2010) proposal takes as a point of departure the annotation of discourse units, following the system of conversational units designed by Briz and Val.Es.Co. group (2014), and DRD are annotated and classified in relation to the discourse units they fulfil and their position and scope over other discourse units.

In the work by Crible and Pascual (2017) and Pascual and Crible (2017) the approaches set out in Crible (2017) and Briz and Pons (2010) were combined with the purpose of analysing DRD, together with other phenomena such as speech disfluencies and discourse units across different corpora. As regards the annotation of DRD specifically, Pascual and Crible (2017) noted that both annotation schemes employ distinct and non-equivalent categories of analysis: first, in terms of identifying DRD, the analysis by Crible (2017) and Crible and Degand (2017a) is performed at a word-level, and the criteria they use to define the DRD category is more inclusive, covering thus more word types. In contrast, the focus of attention in Briz and Pons (2010) is sequential, and DRD that simply establish syntactic relations, such as conjunctions, are excluded from the pragmatic analysis since they do not constitute isolated discourse units (acts or subacts). The second aspect in which the two systems differ substantially relates to the identification of the functions of DRD. The different functions described in the two models cannot be mapped together on a one-to-one basis since the taxonomies of both do not share exact correspondences. Thirdly, the position of DRD is established taking into account different parameters of analysis in each model: the dependency structure, the clause and turns in the case of Crible (2017); and, in the case of Briz and Pons (2010), the conversational units of the Val.Es.Co. model (Briz and Val.Es.Co. group 2014). Moreover, in the work of Briz and Pons (2010) the position of DRD is considered crucial for describing their function.

The solution adopted for the combined annotation in Crible and Pascual (2017) and Pascual and Crible (2017) was the addition of new layers of analysis. A re-annotation procedure was carried out on the corpora based on a set of guidelines and a table of equivalences between the two models, given that a merged protocol in which the divergent functions could be mapped was not feasible. This practical solution, despite being more time-consuming and requiring further training, was more comprehensive in its inclusion of both analyses. The superposition of both systems enabled, at the same time, both a more focused and fine-grained approach, as well as an abstract representation of wider discourse phenomena. The resulting analysis was illuminating not only for the study of disfluencies, but also for the analysis of DRD, as this present study will show.

4 Brief overview of the study

Approximately 1300 DRD were annotated following the guidelines set out in the combined proposal by Crible and Pascual (2017) and Pascual and Crible (2017). The annotation of these DRD incorporated the new functions of the revised taxonomy from Crible and Degand's (2017a) "independent domain and function annotation

scheme for spoken language". The set of data used for the annotation was a sample of Spanish spoken conversations from the *Corpus Val.Es.Co. 2.0* (Cabedo and Pons 2013). The results of this study offer: (1) a quantitative analysis of the different categories of DRD identified by each model and (2) a comparison of the distribution of the different functions of DRD with the aim of highlighting the correspondences and differences between both models. The conclusion will point, furthermore, to the necessity of conserving both models for a contrastive corpus analysis.

This study is intended to illustrate the challenges that arise in cross-linguistic and contrastive studies when attempting to find common ground among multilingual corpora that employ distinct theoretical frameworks for annotating the categories and functions of DRD. At the same time, the study will provide the opportunity to make available annotated Spanish oral data, which until now have not been included in the TextLink Action.

References

Briz, A.: Unidades del discurso, partículas discursivas y atenuantes. El caso de 'no, tienes razón'. In: Falk, J., Gille, J., Wachtmeis, F. (eds.) Discurso, Interacción e identidad. Homenaje a Lars Fant, pp. 13–36. Romanica Stockholmiensia, Stockholm (2006).

Briz, A.: El español coloquial en la conversación. Estudios de pragmagramática. Ariel, Barcelona. (1998).

Briz, A., Estellés, M. (2009): On the relationship between Attenuation, Discourse Particles and Position. Studies in Pragmatics 20, pp. 289–304.

Briz, A., Val.Es.Co. group: Las unidades del discurso oral. La propuesta Val.Es.Co. de segmentación de la conversación (coloquial). Estudios de Lingüística del Español, 35(1), pp. 11–71 (2014).

Briz, A., Pons, S.: Unidades, marcadores discursivos y posicion. In: Loureda O., Acin, E. (eds.) Los Estudios sobre Marcadores del Discurso, pp. 327–358. Acro Libros, Madrid (2010).

Briz, A., Pons, S. and Portolés, J. (coord.): Diccionario de partículas discursivas del español, http://www.dpde.es, last accessed 2018/02/05.

Broisson, Z. Testing independent levels of annotation for the disambiguation and characterization of DMs in a sample of spoken Peninsular Spanish (in preparation).

Cabedo, A., Pons, S. (eds.): *Corpus Val.Es.Co. 2.0*, http://www.valesco.es, last accessed 2018/02/05.

Crible, L.: Discourse Markers and (Dis)fluency across Registers. A Contrastive Usage-Based Study in English and French. PhD Thesis (2017).

Crible, L., Degand, L.: Independent domain and function annotation scheme for spoken language. Corpus Linguistics and Lingustic Theory (2017a). Advanced access: https://doi.org/10.1515/cllt-2016-0046.

Crible, L. and Degand, L.: Testing interdependent annotation levels for sense disambiguation in spoken English, French and Polish. In: 50th Annual Meeting of the Societas Linguistica Europaea, Zurich (2017b).

Crible, L., Pascual, E.: How to be (dis)fluent in English, French and Spanish: discourse markers within repetitions and repairs across languages. In: 15th International Pragmatics Conference, Belfast (2017).

Espinosa, G., García, A.: Conversational structures and discourse genres: A contrastive study of informal conversations, sociolinguistic interviews and broadcast interviews. In: 15th International Pragmatics Conference, Belfast (2017).
González, M.: Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. Discourse Studies 77(1), pp. 53–86.
Halliday, M. and Hasan, R: Cohesion in English. Longman, London (1976).
Martín Zorraquino, M. A. and Portolés, J, Los marcadores del discurso. In: Bosque, I. and Demonte, V. (eds.) Gramática descriptiva de la lengua española, 3, pp. 4051–4213. Espasa Calpe, Madrid (1999).
Montáñez, M.: Marcadores discursivos conversacionales y posición final. Hacia una caracterización discursiva de sus funciones en unidades del habla. PhD Thesis (2015).
Montáñez, M.: Marcadores del discurso y posición final: la forma ¿eh? en la conversación coloquial española. Estudios de Lingüística Universidad de Alicante, pp. 261–280 (2007).
Pascual, E.: Annotating discourse units in spontaneous conversations: The challenge of self-repairs. In: Degand, L., Dér, C., Furkó, P., Webber, B. (eds.): Proceedings of the Second Action Conference and MC Meeting: Cross-Linguistic Discourse Annotation: Towards Shared Resources, Tools and Methods (TextLink: COST Action IS1312), pp. 101–106. Debrecen University Press. Budapest (2016).
Pascual, E., Crible, L.: Discourse markers within (dis)fluent constructions in English, French and Spanish casual conversations: The challenges of contrastive fluency research. In: Fluency & Disfluency Across Languages and Language Varieties, Louvain-la-Neuve (2017).
Pons, S.: Paths of grammaticalization: beyond the LP/RP debate (in press).
Pons, S.: The importance of discourse segmentation for the study of discourse markers (in conversation). In: 4th International Symposium Discourse Markers in Romance Languages: A Contrastive Approach, Heidelberg (2015).
Pons, S.: Paths of grammaticalization in Spanish o sea. In: Ghezzi, P., Molinelli, P. (eds.) Discourse and Pragmatic Markers from Latin to the Romance Languages, pp. 109–136. Oxford University Press, Oxford (2014).
Pons, S.: La combinación de marcadores del discurso en la conversación coloquial: Interacciones entre posición y función. Estudios Lingüísticos / Linguistic Studies, 2, pp. 141–159 (2008).
Pons, S.: Connectives/Discourse markers: an overview. Quaderns de Filologia. Estudis Literaris, 4, pp. 219–243 (2001).
Pons, S.: Conexión y conectores: estudio de su relación en el registro informal de la lengua. PhD Thesis (1998).
Pons, S. and Estellés, M.: Absolute initial position. In: Pons, S. (ed.) Discourse segmentation in Romance languages, pp. 121–155. John Benjamins, Amsterdam/Philadeplhia (2014).
Pons, S. and Salameh, S.: Periferia izquierda, periferia derecha... ¿de qué? Una propuesta desde el sistema de unidades del grupo Val.Es.Co. In: Ferrari, A., Lala, L., Stojmenova, R. (eds.) Testualità. Fondamenti, unità, relazioni / Textualité. Fondements, unités, relations / Textualidad. Fundamentos, unidades, relaciones, pp. 83–99. Cesati, Firenze (2015).
Pose, F.: Actos truncados estratégicos: aspecto formal, hacia el reconocimiento de sus tipos. Oralia 18, pp. 259–280.
Redeker, G.: Ideational and pragmatic markers of discourse structure. Journal of Pragmatics, 14 (3), pp. 367–381 (1990).
Romero, A. E. La comunicación mediada electrónicamente. Análisis de la interacción en las redes sociales. PhD Thesis (in preparation).

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B.: The Penn Discourse TreeBank 2.0. In: Proceedings of the 6th Language Resources and Evaluation Conference, pp. 2961–2968 (2008).

Ruiz, L.: El monólogo humorístico como tipo de discurso. El dinamismo de los rasgos primaries. Cuadernos AISPI 2, pp. 195–218 (2013).

Salameh, S.: Aproximaciones al estudio de subjetividad e intersubjetividad en marcadores discursivos: relación entre periferias, unidades y posiciones. BA Thesis (2014).

Salameh, S., Estellés, M., Pons, S.: Beyond the notion of periphery: An account of polyfunctional discourse markers within the Val.Es.Co. model of discourse segmentation. In: Beeching, K., Ghezzi, P., Molinelli, P. (eds.) positioning the self and others: Linguistic traces. John Benjamins, Amsterdam/Philadelphia (in press).

Scivoletto, G. Marcatori del discorso nel repertorio italiano-dialetto in Sicilia: mutamento, contatto, variazione. PhD Thesis (in preparation).

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

Using annotation to identify connective meanings in a multilingual environment. Romanian and English contrast markers in a parallel corpus

Sorina Postolea[1][0000-0001-9689-1909]

[1] Alexandru Ioan Cuza University, 11 Carol I Blvd., 700506, Iași, Romania
sorinapostolea@gmail.com

Abstract. This paper presents the results of an annotation project that used a parallel corpus to investigate the senses of four Romanian contrast conjunctions – DAR, ÎNSĂ, CI, and IAR – and their translations into English. The spans of discourse linked by these connectives were annotated using the PDTB 2.0 Annotation Scheme. The findings substantiate the status of DAR as the most general Romanian adversative connective. ÎNSĂ is shown to differ from DAR by its preferential use in the case of contra-expectation relations. Sense annotation confirms CI as a dedicated marker for correction. Because it is used to signal non-contrast relations in 76% of its occurrences, IAR is shown to be closer to additive markers than to adversative ones. The English connective BUT covers most of the senses signalled by DAR and CI, HOWEVER is the preferred translation of ÎNSĂ, whereas AND is the most frequent translation of IAR.

Keywords: Contrast Connectives, Romanian Adversative Conjunctions, PDTB, Discourse annotation.

1 Background and Methodology

Several studies and projects have already demonstrated the advantages of using annotated resources in order to map and inventory coherence relations and their markers in either monolingual, e.g. [1–3], or multilingual environments, e.g. [4, 5]. However, such contributions are currently lacking for Romanian, the main representative of the Eastern block and one of the major Romance languages spoken today.

Indeed, due to various historical and geographical factors, such as the influence of neighbouring Slavic dialects or of modern Greek[6], Romanian has unique features that distinguish it from Western Latin-based languages. One of these peculiarities is its system of adversative coordinating conjunctions.

Adversative discourse relations have received many definitions and have been classified in various ways [7, 8]. In most traditional views, the notion of adversity is seen as encompassing the subtype of contrast relations [9, 10]. However, in more recent approaches, contrast is defined as the relation which signals that "in the

speaker's opinion two propositions A and B are valid simultaneously and proposition B marks a contrast to the information given in proposition A" [7], encompassing subcategories such as *adversity/denial of expectation*, *semantic contrast/opposition* or *correction* [7, 8].

It is in the signalling of these subcategories that Romanian differs from the main Western Romance languages and from English. Unlike French, Italian and Portuguese or English, which have a PA-system – with a main marker (i.e. MAIS, MA, and MAS) signalling both adversity/contrast and correction, Romanian is similar to Spanish and is an SN-type of language [8], having a dedicated marker for corrective relations, CI [8, 11–13]. Moreover, according to the literature, Romanian has not one, but two general adversative conjunctions, DAR and ÎNSĂ, usually described as quasi-synonyms [10, 12], and a fourth marker, IAR, with no clear equivalent in the other Romance languages or in English. Among other things, IAR has been described as a marker of "unoriented semantic contrast" [12] or as a "pragmatic discourse organizer" [14]. In English, all of the subtypes of contrast relations described so far may be signalled by BUT, its "primary contrast marker" [7, 15].

Starting from these premises, this paper presents and discusses the results of a pilot annotation project which had a twofold aim:

(1) to shed more light on the way in which the four Romanian adversative conjunctions – DAR, ÎNSĂ, CI, and IAR – may be mapped onto the main subcategories of adversity/contrast relations described in the literature;
(2) to establish a connection between these connectives and the markers used to signal the same relations in English.

The project used a corpus of 200 EUROPARL statements translated from Romanian into English (~40,000 tokens per language). The spans of discourse linked by the four Romanian connectives and their translations into English were extracted from the corpus and annotated using the senses described in the PDTB 2.0 Annotation Manual [3]. As shown in the literature, sense annotation and the *translation spotting technique* may produce reliable data not only on the finer-grained features of the source language but also, from a contrastive perspective, on those of the target language as well [16].

2 Sense Annotation Results

The 200-statement parallel corpus extracted from the EUROPARL database was divided into two subcorpora, one for each language. The text spans linked by the four connectives of interest were retrieved from the corpus, inventoried with their corresponding translations, and checked so as to remove non-clausal uses and correlatives (e.g. NOT ONLY... BUT ALSO), which were beyond the scope of this paper.

2.1 Frequency Distribution

The process described above resulted in a sample of 129 text spans in both languages, which were then manually annotated using the PDTB 2.0 sense classes, types and subtypes [3]. Table 1 shows the frequencies of the four Romanian

conjunctions as well as the distribution of the English connectives used to translate them in the corpus.

Table 1. Frequency distribution of Romanian conjunctions and their translation equivalents

	DAR	ÎNSĂ	CI	IAR	Total
and				26	**26**
and+*			1	1	**2**
as a result				1	**1**
but	21	3	9		**33**
but+			3		**3**
even though	1				**1**
however	8	22		1	**31**
in fact				1	**1**
indeed				5	**5**
nor				1	**1**
omitted	3	1	1	12	***17***
on the other hand				1	**1**
while	1	1		4	**6**
while+				1	**1**
Total	**34**	**27**	**14**	**54**	129

* the "+" sign identifies connectives used in pair with other markers, e.g. BUT INSTEAD

Even if it is described in the literature as the "most general" marker in the Romanian adversative system [11], contrary to expectations, with 34 occurrences, DAR is not the most frequent of the four conjunctions, being clearly surpassed by IAR (54). Moreover, IAR has the largest number of different translation solutions in the corpus (11), and it is the most prone to be omitted (implicitated) – 12 cases (22.2% of its occurrences). On the other hand, BUT is the most frequent translation solution for the four Romanian connectives (33 single uses and 3 paired with INSTEAD, RATHER and AS WELL), being closely followed by HOWEVER (31). Whereas BUT covers most of the occurrences of DAR and CI, HOWEVER is mostly used as an equivalent of ÎNSĂ.

2.2 PDTB 2.0 Annotation

The Penn Discourse Treebank (PDTB) is a widely known project which annotated the argument structure, senses and attribution of discourse connectives and their arguments in a corpus of Wall Street Journal texts. It used a hierarchical sense classification, divided into four large *classes* – TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION – and several *types* and *subtypes* [3]. These senses mapped as follows onto the four Romanian adversative connectives under analysis.

DAR. Described as the "strongest" Romanian adversative connective [12] used to signal mainly denial of expectation relations [8, 11] or "oriented thematic contrast" [12], DAR covers in the corpus almost all the relations included in the PDTB comparison class, e.g. (1) and (2):

– COMPARISON:Contrast:"opposition"

(1) Comunismul este o filosofie perfidă. *În teorie menţionează bunăstare, egalitate, respectarea drepturilor omului*, dar **în practică a însemnat minciună, discriminare, ură şi chiar crimă**.

Communism is a deceitful philosophy. *In theory, it talks about well-being, equality and respect for human rights*, while **in practice, it has meant lies, discrimination, hatred and even crime**.

– COMPARISON:Concession:contra-expectation

(2) *Rezoluţia pe care am votat-o astăzi transmite un semnal politic puternic la Chişinău*, dar **acest semnal trebuie dublat în mod clar de acţiuni concrete ale Comisiei şi Consiliului**.

The resolution which we voted on today sends a powerful political signal to Chişinău, but **this signal must be clearly backed up by specific actions from the Commission and the Council**.

In the corpus DAR was also used to signal two cases of *Pragmatic contrast* and *Pragmatic concession*. The subtype Concession:expectation was the only COMPARISON sense not present in our sample.

ÎNSĂ. Considered an equivalent of DAR, in the corpus, ÎNSĂ signals mostly (74%) COMPARISON:Concession:contra-expectation relations, as in (3).

– COMPARISON:Concession:contra-expectation

(3) *El a fost condamnat la 12 ani de închisoare în urma unui gest simbolic de protest faţă de modalitatea în care a fost trasată graniţa cu Vietnamul*. **Motivul real al condamnării pare a fi** însă **înlăturarea opoziţiei din cursa electorală pentru alegerile parlamentare din 2013**.

He has been sentenced to 12 years in prison following a symbolic gesture of protest which he made against the way in which the border with Vietnam has been marked. However, **the real reason for his conviction seems to be to remove the opposition from the parliamentary election to be held in 2013**.

ÎNSĂ is also used to mark 7 cases of COMPARISON:Contrast, and seems to be preferred in subjective (epistemic) contexts.

CI. PDTB 2.0 does not include an explicit COMPARISON class sense for the meaning "correction", which is the dedicated function that CI has acquired in Romanian [8, 12, 13]. The closest equivalent is the relation labelled as "EXPANSION:Alternative:chosen alternative", said to apply when "two alternatives are evoked in the discourse but only one is taken, as with the connective *instead*" [3]. Out of the four connectives analysed, CI was the easiest to annotate, since all of its 14 occurrences matched this subtype, as in (4) below:

(4) *Presiunea asupra consumului, îngheţarea lui, nu ne-a scos din criză*, ci **a generat o presiune socială fără precedent**.

The pressure exerted on consumption by restricting it has not brought us out of the crisis, but **has actually created unprecedented social pressure**.

IAR. With a still unidentified origin – Philippide [6] claims it comes from Lat. *vero*, but the latest research mark its etymology as uncertain [12] – IAR has received many definitions in the literature, being described as a copulative marker (Bîtea in [12, 14]), as a boundary marker between the additive and the adversative domains specialised in signalling "unoriented semantic contrast" [10, 12] or as a contrast connective subject to a "double contrastiveness constraint" [17].

Sense annotation revealed that in actual discourse IAR signals COMPARISON:Contrast relations in only a limited number of cases (24%), in most of its occurrences the idea of comparison/contrast being completely absent, as in (5):

(5) *Pledarea în favoarea independenţei mandatului deputatului în Parlamentul European este responsabilitatea Parlamentului*, iar **această independenţă nu poate fi pusă în pericol**.

Advocating the independence of an MEP's mandate is the responsibility of Parliament, and **that independence cannot be jeopardised**.

The absence of any identifiable difference between the shared predicates/properties or, indeed, of a shared property, justified the inclusion of the relations signalled by IAR either in the EXPANSION:Specification or in the EXPANSION:Conjunction class. Moreover, in 7 cases, the relation signalled by IAR was judged to be causal, as in (6):

(6) *Populaţia Uniunii îmbătrâneşte*, iar **ponderea persoanelor active din totalul populaţiei scade**.

The EU's population is ageing, ___ **with the proportion of working people among the total population falling**.

Summary. Table 1 below summarizes the PDTB 2.0 senses annotated for the four connectives under analysis.

Table 2. Synoptic distribution of connective senses

CONNECTIVE / CLASS:TYPE	DAR	ÎNSĂ	CI	IAR
COMPARISON:Contrast	35%	26%		24%
COMPARISON:Concession	58%	74%		
COMPARISON:*Pragmatic contrast*	3.5%			
COMPARISON:*Pragmatic concession*	3.5%			
EXPANSION:Alternative			100%	
EXPANSION:Restatement:specification				28%
EXPANSION:Conjunction				39%
CONTINGENCY:Cause:result				9%

3 Conclusions

The results of this annotation project confirm and expand the senses of the four Romanian conjunctions hitherto described in the literature.

DAR seems to be, indeed, the most general contrast marker in Romanian, covering all the relations in the COMPARISON class, with the exception of "Concession:expectation", which in Romanian is signalled by subordinating conjunctions. Even if they are described as quasy-synonyms, DAR and ÎNSĂ have different distributions, the latter being used to signal concessive contra-expectation relations in almost 75% of its occurrences. The different nature of ÎNSĂ when compared to DAR is also highlighted by its being translated by HOWEVER, a more "restrictive" marker [18], in 81% of cases. Annotation confirms CI as a dedicated marker for correction relations.

The most interesting findings refer to IAR. Annotations show that this connective signals contrast relations per se ("opposition" and "juxtaposition") in only some of its occurrences (24%), covering a wide array of senses that may be actually included in the additive EXPANSION class, from "Restatement:specification" to simple "Conjunction". These findings suggest that the inclusion of IAR in the series of adversative connectives should be reconsidered along the lines prefigured by Mauri [13], who described it as a connective signalling atemporal/simultaneous additive relations and semantic contrast, and by Vasilescu [14], who sees it as a "pragmatic discourse organiser" used to introduce new arguments into discourse. The fact that IAR is more akin to additive connectives than to adversative ones is also supported by its capacity to signal "CONTINGENCY:Cause:result" relations – which correspond to the "and-so" meaning of AND discussed by Sweetser [19] – and by the fact that its most frequent translation in the corpus is AND (48%).

The analysis thus suggests a functional equivalence between the English BUT and the Romanian DAR and CI, between ÎNSĂ and HOWEVER, and between IAR and AND.

Larger scale annotation projects are needed to further these findings.

References

1. Carlson, L., Marcu, D.: Discourse Tagging Reference Manual, https://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf, (2001).
2. Péry-Woodley, M.-P., Afantenos, S., Lydia-Mai, H.-D., Asher, N.: La ressource ANNODIS, un corpus enrichi d'annotations discursives. TAL. 52, 71–101 (2011).
3. Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., Webber, B.: The Penn Discourse Treebank 2.0 Annotation Manual, https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf, (2007).
4. Hoek, J., Zufferey, S., Evers-Vermeul, J., Sanders, T.J..: Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. Journal of Pragmatics. 121, 113–131 (2017).
5. Zufferey, S., Degand, L.: Annotating the meaning of discourse connectives in multilingual corpora. Corpus Linguistics and Linguistic Theory. 1–24 (2013).
6. Philippide, A.: Istoria limbii române. Polirom, Iaşi (2011).
7. Rudolph, E.: Contrast: Adversative and Concessive Expressions on Sentence and Text Level. Walter de Gruyter, Berlin (1996).
8. Isutzu, M.N.: Contrast, concessive, and corrective: Toward a comprehensive study of opposition relations. Journal of Pragmatics. 40, 646–675 (2008).

9. Halliday, M.A.K., Hasan, R.: Cohesion in English. Longman, London (1976).
10. Academia Română [Romanian Academy]: Gramatica limbii române [Grammar of the Romanian language]. Editura Academiei Române, Bucureşti (2008).
11. Ragea, O.A.: Conectori pragmatici de contrast, (2009).
12. Zafiu, R.: Conjuncţiile adversative în limba română: tipologie şi niveluri de incidenţă. In: Pană Dindelegan, G. (ed.) Limba română – structură şi funcţionare. pp. 243–258. Editura Universităţii din Bucureşti, Bucureşti (2005).
13. Mauri, C.: Coordination relations in the languages of Europe and beyond. Mouton de Gruyter, Berlin (2008).
14. Vasilescu, A.: Iar – Operator pragmatic. In: Zafiu, R., Dragomirescu, A., and Nicolae, A. (eds.) Limba română: controverse, delimitări, noi ipoteze. Actele celui de-al nouălea Colocviu al Catedrei de Limba Română. pp. 341–355. Editura Universităţii din Bucureşti, Bucureşti (2010).
15. Fraser, B.: On the universality of discourse markers. In: Aijmer, K. and Simon-Vandenbergen, A.-M. (eds.) Pragmatic Markers in Contrast. pp. 73–92. Elsevier, Amsterdam (2006).
16. Cartoni, B., Zufferey, S., Meyer, T.: Annotating the meaning of discourse connectives by looking at their translation: The translation spotting technique. Dialogue and Discourse. 65–86 (2013).
17. Bîlbîie, G., Winterstein, G.: Expressing contrast in Romanian. The conjunction IAR. In: Berns, J., Jacobs, H., and Scheer, T. (eds.) Romance Languages and Linguistic Theory 2009. Selected papers from "Going Romance" Nice 2009. pp. 1–18. John Benjamins, Amsterdam (2011).
18. Blakemore, D.: Relevance and linguistic meaning. The semantics and pragmatics of discourse markers. Cambridge University Press, Cambridge (2002).
19. Sweetser, E.: From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure. Cambridge University Press, Cambridge (1990).

Choosing among alternatives: Conjunction variability comes from both inference and the semantics of discourse adverbials

Hannah Rohde
University of Edinburgh
Hannah.Rohde@ed.ac.uk

Alexander Johnson
University of Edinburgh
ajohnso5@exseed.ed.ac.uk

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Bonnie Webber
University of Edinburgh
Bonnie.Webber@ed.ac.uk

Discourse coherence relations serve to link clause-level semantics and discourse-level semantics. The typical assumption is that they are signalled either explicitly, by conjunctions (BECAUSE, SO, OR) or discourse adverbials (*therefore*, *however*), or else implicitly, through inference, **but not simultaneously via explicit and implicit signals**.

Recent findings challenge this simple explicit vs. implicit dichotomy in two ways: (i) A discourse relation may be inferred even when a discourse adverbial is present (Rohde et al. 2015, 2016, 2017b; see also Asr and Demberg 2013; Tatiana Scheffler, p.c.); and (ii) the available evidence may license more than one inferred relation (Prasad et al., 2014).

Our findings come from conjunction-completion experiments in which naive participants were asked to read passages such as (1), made up of two text segments joined by a gap followed by a discourse adverbial, and then asked to fill in the gap with a *conjunction* that best expressed how they took the segments to be related. Participants endorsed conjunctions whose sense differed from the discourse adverbial and which usually signal different coherence relations (Rohde et al., 2017b). For example, for passage (1) with the discourse adverbial *in other words*, participants frequently and systematically chose to insert OR as well as SO. This SO∼OR substitutability is unexpected because, even if one takes their semantics to be "weak", the two conjunctions appear neither synonymous nor representative of the same relation.

(1) Unfortunately, nearly 75,000 acres of tropical forest are converted or deforested every day _____ in other words an area the size of Central Park disappears every 16 minutes. [SO∼OR]

Rohde et al. (2017b) note other improbable substitutability pairs (e.g., BECAUSE∼BUT, BUT∼OR, and BECAUSE∼OR) that emerged systematically across participants and across passages for particular adverbials, but they did not provide empirical evidence for what motivates these possible substitutions. Here we do so for three adverbials with related lexical semantics — *in other words*, *otherwise* and *instead* — all of which convey that the clause in which they appear provides a (disjunctive) *alternative*. Similar lexical semantics could be realised by the conjunction OR.

The passages used in the current study are simplified variants of the naturally occurring passages used in our previous studies. As well as simplifying the passages, we manipulated them systematically, in ways that alter how available different coherence relations were to the participants. The goal is to understand how properties of the passage drive preferences for the establishment of particular (sometimes co-occurring) coherence relations.

Here, we first present results for *in other words*. While its lexical semantics of disjunctive alternative, plus consequence (for its sense of entailed reformulation) can be realised with the conjunctions OR and SO, our results show that manipulating the immediately preceding segment can shift participants' preference from relations associated with OR and SO to relations of contrast or concession. We take this as evidence that adjacency affects what coherence relations participants take to be available.

We then present results for *otherwise*. Again, different properties of the passage yield preferences among the set of available coherence relations. The lexical semantics of *otherwise*, as an indicator of an alternative, permits it to appear in passages which cohere via inferences of causal reasoning, or enumeration, or contrast between a generalization and an exception. Passages that instantiate each of these inferences yield different preferences in participants' conjunction choices, showing how manipulating semantic properties of the passage can alter the availability of particular coherence relations.

We close with results for *instead*. These will show that manipulating even a single property of the segments in a passage can alter the perceived availability of different coherence relations, as evident in participants' choice of conjunction. In this case, the lexical semantics of *instead* is not realized in participants' choice of conjunction since they rarely select a marker of disjunction; rather the conjunctions reflect relations of contrast and causality that are inferrable links between the segments.

All the results we present involve explicit discourse adverbials, from a task where we ask participants to fill in the conjunction(s) that best express how the two segments in a passage link together. The reason for this use of explicit adverbials and the conjunction-completion task is that these discourse adverbials are *anaphoric* (Webber et al., 2000, 2001) and are thus not constrained by structure as to what they establish discourse relations with. The same doesn't hold of conjunctions such as AND, BECAUSE, BUT, OR and SO. So a conjunction-completion task can be used to assess links between the segments.

1 *In other words*: Inference and Adjacency

We first noticed an OR∼SO split for *in other words* in the crowd-sourced conjunction-completion experiment reported in (Rohde et al., 2016). In this experiment, participants identified only their top choice of conjunction to fill in the gap. While SO dominated participants' choice in all cases, only one case lacks OR as one of the choices (Figure 1). For this and other figures in the paper, each vertical bar represents a passage with the responses from each of our participants color-coded by conjunction.

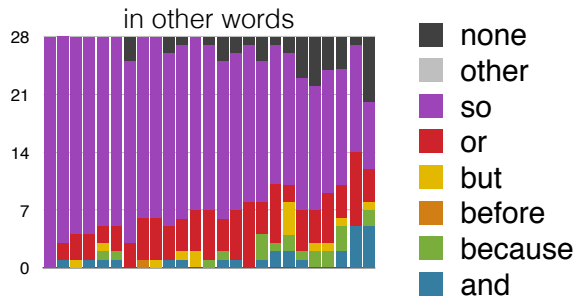


Figure 1: Stacked bar chart for participants' (N=28) conjunction completions in passages with *in other words* (Rohde et al., 2016)

The current study considered OR∼SO splits associated with participants' identifying OR or SO or both as their top choice of conjunction. One possibility is that this split, as in passage (1), arises from two simultaneous sources: the lexical semantics of *in other words* and an inference of causal consequence. The latter derives from the segments themselves, whereby the second (reformulation) segment (i.e., the disappearance of an area the size of Central Park) is entailed by the first segment (the deforestation of 75,000 acres). One might therefore speculate that *in other words* would always license OR via its lexical semantics and SO via the entailment relationship. But this is not always the case,

(2) Unfortunately, nearly 75,000 acres of tropical forest are converted or deforested every day. *I don't know where I heard that* _____ in other words an area the size of Central Park disappears every 16 minutes.

Here BUT has become more available. That is, the substitutability of SO∼OR in (1) appears to depend on the two segments being immediately adjacent.

Starting with 16 passages containing *in other words*, we created minimal pairs which varied in the presence/absence of a meta-linguistic comment, as in the pair (1)–(2) and the pair in (3)–(4).

(3) Typically, a cast-iron wood-burning stove is 60 percent efficient _____ in other words 40 percent of the wood ends up as ash, smoke or lost heat.

(4) Typically, a cast-iron wood-burning stove is 60 percent efficient. *How this is measured is unclear* _____ in other words 40 percent of the wood ends up as ash, smoke or lost heat.

For each passage, our 28 participants selected their preferred conjunctions, half seeing the passage variant with no intervening comment, and half seeing it with a comment.

Results Figure 2 shows the results. As predicted, participants selected SO/OR in the no-intervening-content condition, whereas with intervening content, the selection of OR decreases, while BUT (and occasionally BECAUSE) increase.[1] In the figure, the pair (1)–(2) corresponds to passage C and the pair in (3)–(4) to passage O. The latter shows no selection of OR, and a sharp drop in the selection of SO. We posit that increases in BUT associated with the intervening content indicate either an interruption of the meta-linguistic tangent or an intention to signal a contrast with the negative affect of the tangent itself (e.g., "I don't know where…", "frustrating way of putting it", "How this is measured is unclear"). We speculate that an increase in BECAUSE in the with-intervening-content condition may arise when the intervening material implies that the situation is somehow surprising, which in turn merits explanation (e.g., "it's an UNUSUAL role for her", "Their ability to actually work sensitively is perhaps QUESTIONABLE", "It's STRANGE to think of a planet being born"). These hypotheses will themselves need to be tested.

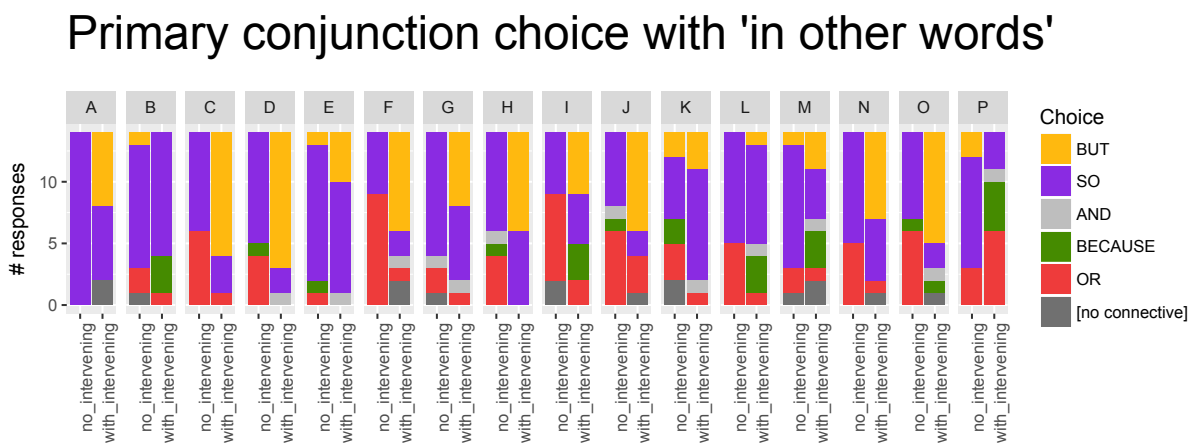


Figure 2: Distribution of first choice for participants' conjunction completions in passages with *in other words*. Each participant saw only one variant.

2 *Otherwise*: Inference from semantic features of segments

We first noticed unexpected BECAUSE∼BUT∼OR splits for *otherwise* in the same crowd-sourced conjunction-completion experiment as with *in other words* (Rohde et al., 2016). (See Figure 3.)

Although in this earlier study, participants only identified their top choice of conjunction, our goal here was to test the hypothesis that such splits arose from a combination of the lexical semantics of *otherwise* and inference from the segments themselves of either causal reason or contrast.

[1] Passage P in Figure 2 is an outlier. We speculate that participants took the *in other words* clause to link to the intervening material itself.

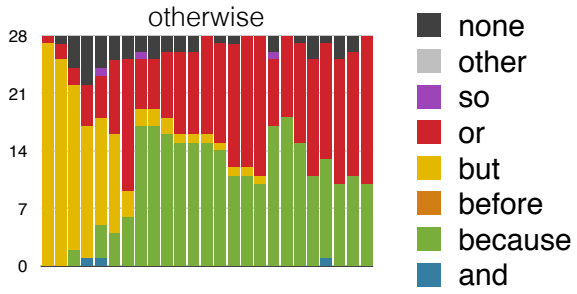


Figure 3: Stacked bar chart for conjunction completion passages involving *otherwise*, from (Rohde et al., 2016)

These inferences are linked to different uses of *otherwise*: in ARGUMENTATION, to provide a reason for a given claim, as in (5); in ENUMERATION, when the speaker first gives some preferred or more salient options, with *otherwise* introducing some alternative options, as in (6); and in expressing an EXCEPTION to a generalization that covers all but the specified disjunctive alternative(s), as in (7).

(5) Proper placement of the testing device is an important issue _____ otherwise the test results will be inaccurate.

(6) A baked potato, plonked on a side plate with sour cream flecked with chives, is the perfect accompaniment _____ otherwise you could serve a green salad and some good country bread.

(7) Mr. Lurie and Mr. Jarmusch actually catch a shark, a thrashing 10-footer _____ otherwise the action is light.

Rohde et al. (2017b) showed that passages like these permit the establishment of disjunction alongside another relation. In (5), *otherwise* delivers the disjunctive alternative "if not placed properly", alongside the inferred relation that the second segment conveys a reason for the first. Here Rohde et al. showed participant judgments of OR and BECAUSE, but not BUT.

In (6), *otherwise* delivers a disjunctive alternative that is another element of the enumeration, but stands in contrast with it (as less preferred or salient). Here Rohde et al. showed pairings of OR and BUT, but not BECAUSE.

In (7), *otherwise* delivers an alternative situation – an incident in which John Lurie and Jim Jarmusch catch a shark. On infers that, except for this incident, the right segment (that the action in the film is light) is an appropriate generalization. Here Rohde et al. showed only BUT (and the less specific AND) convey that this generalization contrasts with the first segment's exception. There is neither causal reasoning nor a disjunction between alternatives since the scenarios described in both segments hold simultaneously.

Note that because of several overlaps in conjunction choice, some conjunctions cannot be unambiguously associated with one use of *otherwise*: While BECAUSE may unambiguously signal that a participant has inferred ARGUMENTATION, OR might indicate inference of either ARGUMENTATION or ENUMERATION.

We chose 16 passages for each use of *otherwise*. (While this was based on our judgment, we also elicited participant judgment through paraphrase selection, not discussed here.) For each passage, we asked participants to select the conjunction that best expressed how its two segments were related, and then any other connectives that they took to express the same thing.

Results On aggregate, our assigned use type correlate strongly with the connectives chosen by the participants – specifically, of the 448 judgments on ARGUMENTATION passages (28 participants $\times$ 16 passages), 411 were BECAUSE or OR or both ($\approx$92%). On EXCEPTION passages, 364 of the 448 judgments were BUT, AND or both BUT and AND ($\approx$81%). On ENUMERATION passages, 426 of the 448 judgments were BUT, AND or OR, or some subset thereof ($\approx$95%).

Turning to individual passages, participant choices are shown in Figures 4-6. For ARGUMENTATION (Figure 4), the effect is uniformly strong, with all passages showing BECAUSE or OR as participants' top choice, with OR or BECAUSE chosen as equivalent (shown in the columns labelled "second"). For EXCEPTION (Figure 5), BUT is consistently the participants' top choice.

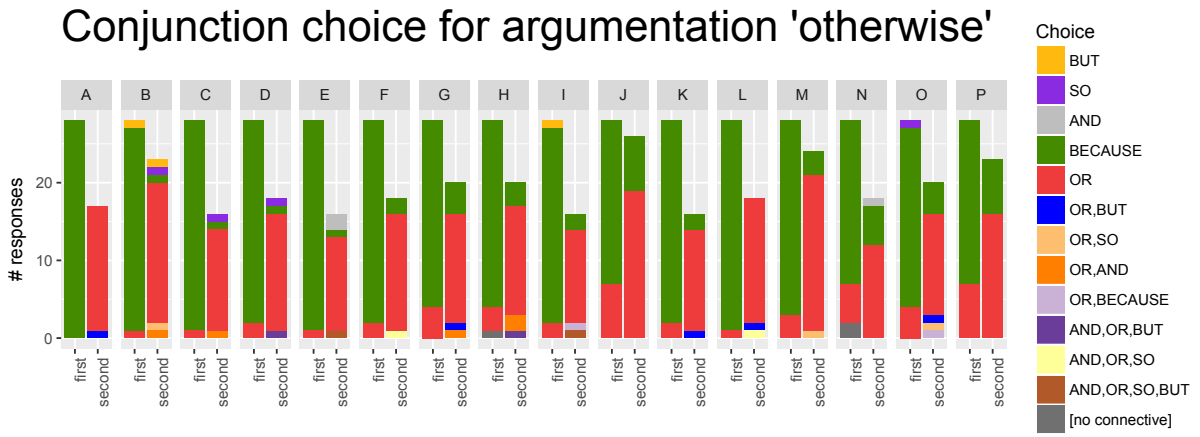


Figure 4: Distribution of first and second choice conjunctions for 'argumentation' *otherwise*. Labels in the legend such as "OR,BUT" are for multiple second choices.

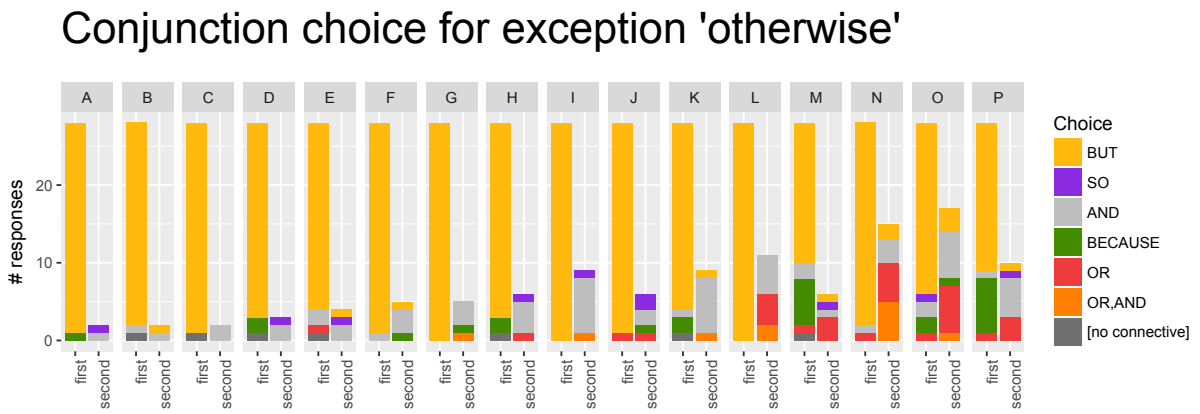


Figure 5: Distribution of first and second choice conjunctions for 'exception' *otherwise*. The label "OR,AND" in the legend implies both as second choices.

There are a few deviations (passages L through P in Figure 5) from this near uniform endorsement of BUT for EXCEPTION. But they would require too much space to discuss, and in any case, suggest further experimentation. Just for example, in passage M (see (8)) and P (see (9)), participants rarely identified any conjunction as conveying the same sense as BUT. However, when they selected BECAUSE as their top choice, they also selected OR as conveying the same sense. As noted above, BECAUSE and OR predominate with *otherwise* used in ARGUMENTATION. This raises the question as to what in passages M and P leads some participants to infer ARGUMENTATION and others, either EXCEPTION or ENUMERATION.

(8) Democrats insist that the poor should be the priority, and that tax relief should be directed at them _____ otherwise they lack a cogent vision of the needs of a new economy. (Passage M)

(9) He said that the proposed bill would give states more flexibility in deciding whether they wanted

to use the Federal money for outright grants to municipalities or to set up loan programs _____ otherwise it left last fall's Congressional legislation unchanged. (Passage P)

Finally, though the pattern for ENUMERATION (Figure 6) is harder to see, combinations of BUT, OR and AND predominate throughout participants' top choice, with a few tokens of BECAUSE and SO, but too few to analyse as anything but noise.

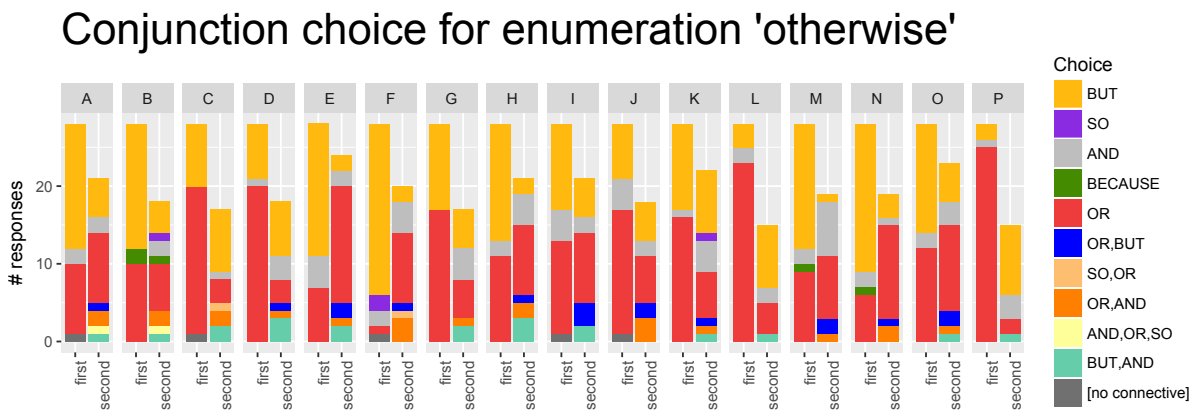


Figure 6: Distribution of first and second choice conjunctions for 'enumeration' *otherwise*. Labels in the legend such as "SO,OR" are for multiple second choices.

We conclude from the part of our experiment involving *otherwise* that our hypothesis is correct, that variability in participants' choice of conjunctions follows from both the lexical semantics of *otherwise* itself and the relation that participants infer between the segments in the passage.

3 *Instead*: Inference from a single manipulated property

Figure 7 shows participant choices in the conjunction-completion passages involving *instead* from (Rohde et al., 2017a). They range from passages on the left in which participants uniformly chose BUT, to one on the right where they uniformly chose SO. In the middle are many more in which some participants chose BUT and some chose SO. Even more surprising were passages like (10) from a subsequent experiment (Rohde et al., 2017b) where some participants selected both BUT and SO as equally expressing how the segments in the passage were related.

(10) There may not be a flight scheduled to Loja today _____ instead we can go to Cuenca. [BUT∼SO]

Figure 7: Stacked bar chart for participants' (N=28) conjunction completions in passages with *instead* (Rohde et al., 2017a)

These various BUT∼SO splits cannot follow from *instead* itself, which simply conveys that what follows is an alternative to an unrealised situation in the context (Prasad et al., 2008; Webber, 2013). So the current experiment tested the hypothesis that the BUT∼SO split is a consequence (as with *otherwise*) of inference from properties of the segments themselves.

Here we took 16 passages with *instead* and created one variant that emphasized the information structural parallelism between the clauses as in (11) and another variant as in (12) that de-emphasized that parallelism in favor of a causal link implied by a downward-entailing construction such as *too X* (Webber, 2013). We used the same conjunction-completion task as above. However, we report results for only 15 passages due to an error in how the 16th was presented to the participants.

(11) There was no flight scheduled to Loja yesterday _____ instead there were several to Cuenca.

(12) There were too few flights scheduled to Loja yesterday _____ instead we went to Cuenca.

Results On aggregate, participants responded very differently to the parallel and causal variants.

participant top choice	parallel	causal
BUT	169	6
SO	12	205
AND	19	13
BECAUSE	6	–
OR	1	–

Considering the individual passages, Figure 8 shows that in all cases, the parallel variant yielded more BUT responses, whereas the causal variant yields more causal SO.

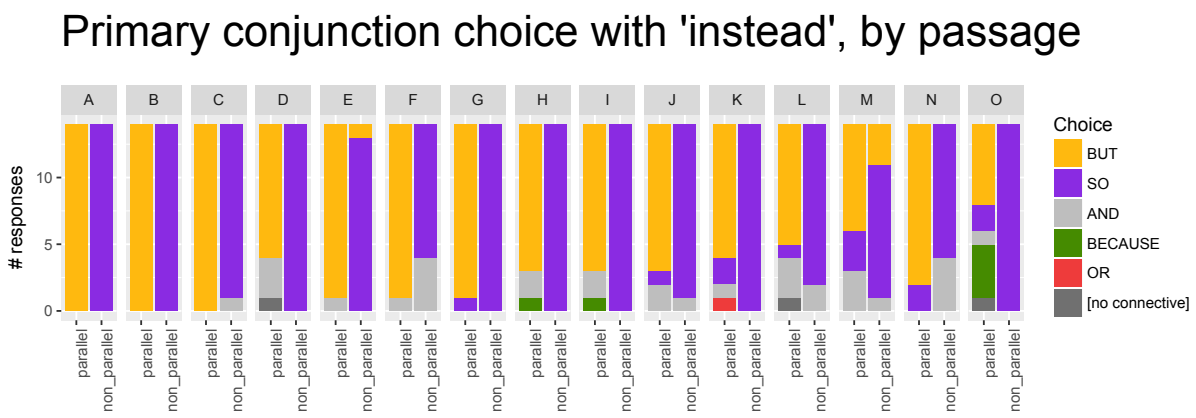


Figure 8: *Instead* passages, pairing a parallel variant and a causal variant. Each column shows the distribution of participants' first choice in the conjunction-completion task. Each participant saw only one variant.

There is a question though as to why inference yields such clean results for both parallel and causal variants of (13), corresponding to Passage A, while yielding much noisier results for the parallel variant of (14a), corresponding to Passage O.

(13) a. Despite the change in government, Miss Bohley could have kept her seat in the German Parliament _____ instead she decided to retire from public view.

b. With the change in government, Miss Bohley would have had a difficult battle for her seat in the German parliament _____ instead she decided to retire from public view.

(14) a. Smugglers nowadays don't use overland passages _____ instead they use the seas to transport their goods.

b. Smugglers' overland passages nowadays are too visible _____ instead they use the seas to transport their goods.

The answer simply seems to be that the negative claim in the first segment of (14a) could be explained by the positive claim in the second segment (BECAUSE), or contrasted with it (BUT), or a result of it (SO). That is, parallel constructions don't guarantee contrast, by virtue of their parallelism alone.

4 Conclusion

The analysis presented here explains conjunction substitutability in terms of both the lexical semantics of discourse adverbials and properties of the passages that contain them. This conjunction substitutability is additional evidence for believing in the simultaneous availability of multiple coherence relations and for believing that they arise from both explicit and implicit signals.

Acknowledgments

This work was supported by a grant from the Nuance Foundation.

References

Asr, F. T. and V. Demberg (2013). On the information conveyed by discourse markers. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, Sofia, Bulgaria.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber (2008). The Penn Discourse TreeBank 2.0. In *Proceedings, 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, pp. 2961–2968.

Prasad, R., B. Webber, and A. Joshi (2014). Reflections on the Penn Discourse TreeBank, comparable corpora and complementary annotation. *Computational Linguistics 40*(4), 921–950.

Rohde, H., A. Dickinson, C. Clark, A. Louis, and B. Webber (2015). Recovering discourse relations: Varying influence of discourse adverbials. In *Proceedings, First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, Lisbon, Portugal, pp. 22–31.

Rohde, H., A. Dickinson, N. Schneider, C. Clark, A. Louis, and B. Webber (2016). Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the Tenth Linguistic Annotation Workshop (LAW-X)*, Berlin, pp. 49–58.

Rohde, H., A. Dickinson, N. Schneider, A. Louis, and B. Webber (2017a). ConnText: Recognizing concurrent discourse relations. Second annual report to the Nuance Foundation.

Rohde, H., A. Dickinson, N. Schneider, A. Louis, and B. Webber (2017b). Exploring substitutability through discourse adverbials and multiple judgments. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*, Montpellier.

Webber, B. (2013). What excludes an alternative in coherence relations? In *Proceedings, 10th International Conference on Computational Semantics*, Potsdam, Germany.

Webber, B., A. Joshi, and A. Knott (2000). The anaphoric nature of certain discourse connectives. In *Making Sense: From Lexeme to Discourse*, Groningen, The Netherlands.

Webber, B., A. Knott, and A. Joshi (2001). Multiple discourse connectives in a lexicalized grammar for discourse. In H. Bunt, R. Muskens, and E. Thijsse (Eds.), *Computing Meaning (Volume 2)*, pp. 229–249. Kluwer.

Discourse Connectives and Reference

Kateřina Rysová and Magdaléna Rysová

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

Abstract. In the present paper, we examine discourse connectives from the perspective of reference (i.e. a presence of an anaphoric element). We introduce a division of connectives into: i) connectives without an inherent (internal) reference (e.g. *and*, *but*, *or*, *if*, *however*, *so*), and ii) connectives with an inherent (internal) reference that is either optional (e.g. *as a result* vs. *as a result of this*), or obligatory – cf. already grammaticalized connectives (e.g. *thereafter*, *therefore* or *thereby*) vs. not yet grammaticalized connectives (e.g. *because of this* or *for this reason*). We apply this general division on Czech and German connectives and conduct a contrastive study on the parallel data of the corpus InterCorp 10. Specifically, we focus on the group of Czech connectives in the form of prepositional phrases with an obligatory inherent reference that do not have any fully grammaticalized form in Czech (like *kromě toho*, lit. "except this", 'moreover') and we search for their most frequent semantic counterparts in German. The results of our research demonstrate that the German counterparts of the selected connectives in Czech are mostly (in 72%) grammaticalized connectives containing a referential morpheme (e.g. *außerdem*, *deswegen*, *stattdessen*, *dagegen*, *demgegenüber*, *daneben*, *infolgedessen*).

Keywords: Discourse Connectives, Reference, Anaphoric Connectives.

1 Introduction

Semantic discourse relations as well as coreference relations substantially participate in creating a coherent text. Both of them belong to the basic cohesive relations with an ability to form cohesive ties and chains (see Halliday and Hasan, 1976). Semantic discourse relations may be signaled explicitly by discourse connectives or they may be only implicit (details on borderlines between explicit and implicit discourse relations are given in Taboada, 2009). Coreference relations are realized very often through demonstrative and personal pronouns – a detailed description of coreference and anaphoric realizations in Czech is presented in Nedoluzhko (2011) or more recently in Zikánová et al. (2015).

It is interesting that many discourse connectives also contain an anaphoric element (like *therefore*, *thereby* etc.). In this way, semantic and coreference relations are mutually interconnected and investigation of their relationship is essential for text coher-

ence in general (the need for studying coherence through interplays is addressed e.g. by Hajičová, 2011 or Nedoluzhko and Hajičová, 2015).

In the present paper, we aim to examine the discourse connectives containing a referential (anaphoric) component. Specifically, we divide discourse connectives into several groups according to their ability to express anaphora in the surface structure and we present a contrastive analysis of Czech and German connectives containing an explicit anaphoric element like *kromě toho* – *außerdem* lit. "except this", 'moreover', or *kvůli tomu* – *deswegen* "because of this".

2 Discourse Connectives: Description and Delimitation

Generally, a discourse connective is defined as a predicate of a binary relation opening two positions for two text spans as its arguments and signaling a semantic or pragmatic relation between them (Prasad et al., 2008). Discourse connectives may be further divided into primary and secondary (Rysová and Rysová, 2014 and 2015), the groups differing in the degree of grammaticalization – cf. the grammaticalized primary connectives (e.g. *and, but, however, therefore*) and not yet fully grammaticalized secondary ones (e.g. *for this reason, on condition that*).

From the perspective of anaphora and discourse structure, a description of connectives is given in Webber et al. (2003) who distinguish between anaphoric connectives (mostly certain adverbials; picking up their external argument by means of anaphora resolution) and structural connectives (taking arguments qua the syntactic configuration they appear in). Anaphoric connectives in German were studied by Stede and Grishina (2016) who focused on the description of a group of German connectives containing a morpheme overtly referring backward (e.g. *demzufolge*). Anaphoric connectives in Czech are rather an unexplored topic – the first probe was carried out by Poláková et al. (2012) exploring a subgroup of these expressions in the form of a preposition and a demonstrative pronoun.

3 Discourse Connectives and Reference: General Overview

As mentioned above, a general property of discourse connectives is to connect two text units. Thus, if a discourse connective appears in a text, we assume that it somehow refers to the previous context, i.e. the presence of a connective implies the presence of the first discourse argument (see Halliday and Hasan, 1976). In this respect, discourse connectives and (co)reference relations are strongly inter-related – all discourse connectives may be viewed as implicitly referential (e.g. connectives like *but, and, or* do not contain any anaphoric element but still they signal a presence of the first discourse argument).

At the same time, within the discourse connectives, there is a narrower set of expressions containing a referential (anaphoric) element explicitly, cf. examples of secondary connectives like *because of this, after this, as a result of this, this is the reason why, under these conditions, for this reason* etc. However, also these expressions differ from each other, as some of them (e.g. *because of this*) contain the anaphoric

element obligatorily while some of them only optionally (cf. *as a result* vs. *as a result of this*). Concerning the presence of a referential (anaphoric) element, connectives may be thus divided into the following groups:

1) connectives **without an inherent (internal) reference** (e.g. *and*, *but*, *or*, *if*, *however*, *so*);

2) connectives **with an inherent (internal) reference** that is:

2a) **optional** (e.g. *as a result* vs. *as a result of this*);

2b) **obligatory**
- already grammaticalized connectives (e.g. *thereafter*, *therefore*, *thereby*);
- not yet grammaticalized connectives (e.g. *because of this*, *for this reason*).

The way of expressing reference in connectives may differ across languages, which is noticeable especially on semantic equivalents (cf. e.g. Czech *místo toho* vs. English *instead* vs. German *stattdessen*). For example, we cannot use prepositions without reference as discourse adverbs in Czech, which is possible in English, see Examples (1a) and (1b) from the parallel corpus InterCorp 10 (Rosen and Vavřín, 2017).

In the Czech example (1a), the discourse relation is expressed by the connective *místo toho* (lit. "instead of this") that cannot be used without the anaphoric part *toho* "this". Example (1a) without *toho* is ungrammatical (**Místo dál pochodovala…*). On the contrary, such usage of *instead* in the English version (1b) is fully functional.

(1a) Czech: *Ale i když měla pokušení koupit si dlouhé černé šaty, které viděla viset v butiku Betsey Johnsonové, nevešla ani dovnitř.* ***Místo toho*** *dál pochodovala jednou z uliček tam a druhou zase zpátky.*

(1b) English: *But even though she's tempted by a long black dress she sees hanging on the far wall in Betsey Johnson, she doesn't go inside.* ***Instead,*** *she continues trance-like up one street and down another*.

(1c) German: *Selbst als sie in einem Betsey-Johnson-Shop ein langes schwarzes Kleid hängen sieht, das sie reizen könnte, betritt sie den Laden nicht.* ***Stattdessen*** *schreitet sie wie in Trance eine Straße nach der anderen ab.*

Examples (1a) and (1c) illustrate that referential connectives may differ also in the degree of grammaticalization – whereas Czech *místo toho* (lit. "instead of this") is not yet fully grammaticalized, its German anaphoric counterpart *stattdessen* is fully lexicalized as a one-word connective.

4 Referential Connectives in Czech and German

In our study, we focus on the group of Czech secondary connectives in the form of prepositional phrases with an obligatory inherent reference (representing the most frequent set of expressions in the group 2b in the scheme above) that do not have any fully grammaticalized form in Czech (like *kromě toho*, lit. "except this", 'moreover'). We select 10 most typical representatives of these referential connectives in Czech (listed in Table 1)[1] and we search for their most frequent semantic counterparts in German based on the parallel data of the corpus InterCorp 10 (using the Treq tool, see http://treq.korpus.cz/). The German counterparts were firstly found automatically in InterCorp 10 and then sorted out manually.

In the first step, we examine how often (in total numbers) these non-grammaticalized referential connectives in Czech are expressed as grammaticalized referential connectives in German (see Table 1), i.e. how many corpus occurrences correspond to the relation between Czech and German connectives demonstrated on Examples (1a) and (1c).

Table 1. Percentage of German counterparts of Czech connectives like *kromě toho* (lit. "except this", 'moreover') in InterCorp.

Czech non-grammaticalized referential connectives	**German grammaticalized referential equivalents**		**Other German equivalents**	
	Occurrences in InterCorp	**%**	**Occurrences in InterCorp**	**%**
kromě toho "except this" 'moreover'	5,671	71%	2,328	29%
naproti tomu "in contrast to this"	924	74%	324	26%
místo toho "instead of this"	708	84%	139	16%
kvůli tomu "because of this"	576	74%	205	26%
navzdory tomu "in contrast to this"	141	57%	106	43%
díky tomu "thanks to this"	156	71%	64	29%
vedle toho "besides this"	103	84%	19	16%
oproti tomu "in contrast to this"	109	94%	7	6%
vzhledem k tomu "with regard to this"	39	49%	40	51%
na rozdíl od toho "in contrast to this"	0	0%	9	100%
In Total	**8,427**	**72%**	**3,241**	**28%**

[1] The selection was based on the language material of the Prague Discourse Treebank 2.0 (PDiT 2.0; Rysová et al., 2016), a corpus containing manual discourse annotation of both primary and secondary connectives.

Table 1 demonstrates that non-grammaticalized referential connectives in Czech (like *naproti tomu*, *vedle toho*) are expressed as grammaticalized referential connectives in German (like *dagegen*, *daneben*) in 72%, i.e. German grammaticalized variants are the preferable ones in these cases.

In the next step, we analyse the individual German equivalents for the selected connectives in Czech in more detail, see Table 2.

The results of our research demonstrate that the selected connectives in Czech have diverse counterparts in German. In most cases (in 72%), these German counterparts are grammaticalized primary connectives containing a referential morpheme (e.g. *außerdem*, *deswegen*, *stattdessen*, *dagegen*, *demgegenüber*, *daneben*, *infolgedessen*). However, in some cases, a primary connective without an inherent reference is also used (e.g. *auch*, *obwohl*, *doch*). Some German counterparts in InterCorp are also non-grammaticalized secondary connectives, very often containing an explicit reference (cf. *abgesehen davon*, *hinzu kommt*, *ergänzend dazu*, *dessen ungeachtet*, *im Gegensatz dazu*, *angesichts dessen*).

Table 2. List of German counterparts of Czech connectives like *kromě toho* (lit. "except this", 'moreover') in InterCorp.

Czech connectives	**German equivalents (occurrences in InterCorp)**
kromě toho lit. "except this" 'moreover'	*außerdem* (3,114), *darüber hinaus* (1,413), *auch* (1,059), *zudem* (595), *ferner* (589), *zusätzlich* (174), *überdies* (169), *des Weiteren* (149), *dazu* (149), *im Übrigen* (111), *weiterhin* (96), *abgesehen davon* (78), *daneben* (75), *hinzu kommt* (63), *ebenso* (53), *außer* + NP (49), *übrigens* (26), *nebenbei* (19), *desgleichen* (7), *weiters* (7), *ergänzend dazu* (4)
naproti tomu "in contrast to this"	*dagegen* (549), *hingegen* (300), *im Gegensatz dazu/hierzu* (238), *andererseits* (75), *demgegenüber* (69), *im Gegenteil* (11), *trotzdem* (6)
místo toho "instead of this"	*stattdessen* (689), *vielmehr* (98), *anstatt* + NP (41), *dagegen* (19)
kvůli tomu "because of this"	*wegen* + NP (199), *deswegen* (191), *deshalb* (142), *dafür* (119), *darüber* (64), *dazu* (33), *darum* (27), *aufgrund* + NP (6)
navzdory tomu "in contrast to this"	*trotzdem* (79), *dennoch* (62), *trotz + NP* (53), *doch* (19), *obwohl* (10), *trotz allem* (11), *allerdings* (9), *dessen ungeachtet* (4)
díky tomu "thanks to this"	*damit* (50), *dadurch* (47), *durch* + NP (38), *deshalb* (22), *dank* + NP (20), *infolgedessen* (17), *somit* (13), *deswegen* (7), *aufgrund* + NP (6)
vedle toho "besides this"	*daneben* (67), *außerdem* (25), *zudem* (11), *zusätzlich* (7), *nebenbei* (7), *andererseits* (5)
oproti tomu "in contrast to this"	*dagegen* (72), *hingegen* (28), *demgegenüber* (9), *im Gegensatz dazu/hierzu* (7)
vzhledem k tomu "with regard to this"	*daher* (39), *angesichts dessen* (34), *im Hinblick darauf* (4), *infolgedessen* (2)
na rozdíl od toho "in contrast to this"	*im Gegensatz dazu/hierzu* (9)

5 Conclusion

In our paper, we focused on the interaction of discourse connectives and (co)reference. We divided connectives into several general groups according to whether they contain an inherent reference (*and* vs. *therefore*), whether the reference is optional or obligatory (*as a result (of this)* vs. *because of this*) and whether the connectives with the obligatory reference are already grammaticalized (*therefore* vs. *for this reason*). This general description works for connectives across languages but languages differ in preferences of the individual groups. These differences are especially noticeable if they concern semantic equivalents.

In our study, we further focused on referential connectives in Czech and German (and slightly in English) in parallel data of the corpus InterCorp 10. We demonstrated that there is a group of semantic equivalents of connectives with a similar structure in Czech, German and English that differ right in this referential aspect, cf. Czech *místo toho* belonging to the group of non-grammaticalized connectives with an obligatory reference, German anaphoric *stattdessen* that is already grammaticalized and English *instead* that is typically used without an explicit reference.

Based on the Czech-German analysis, we conclude that the German counterparts of Czech non-grammaticalized referential connectives (like *naproti tomu*, *vedle toho*) are mostly (in 72%) grammaticalized referential connectives (like *dagegen*, *daneben*). From this point of view, there is a stronger tendency to grammaticalization of referential connectives in German than in Czech.

Acknowledgements

We acknowledge support from the Czech Science Foundation project no. GA17-06123S (*Anaphoricity in Connectives: Lexical Description and Bilingual Corpus Analysis*). This study has utilized the language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

1. Hajičová, E.: On interplay of information structure, anaphoric links and discourse relations. *Societas linguistica europaea, SLE 2011, 44th Annual Meeting*, Book of Abstracts, pp. 139–140. Copyright © Universidad de la Rioja, Center for Research in the Applications of Language, Logrono, Spain (2011).
2. Halliday, M. A. K., Hasan, R.: *Cohesion in English*. London: Longman (1976).
3. Nedoluzhko, A.: *Rozšířená textová koreference a asociační anafora (Koncepce anotace českých dat v Pražském závislostním korpusu)*. Copyright © Ústav formální a aplikované lingvistiky, Praha, Česká republika, ISBN 978-80-904571-2-6, 268 pp. (2011).
4. Nedoluzhko, A., Hajičová, E.: *Information Structure and Anaphoric Links – A Case Study and Probe.* Contributed talk, Corpus Linguistics 2015 (CL2015), Lancaster University, UK, Lancaster, UK (2015).

5. Poláková, L., Jínová, P., Mírovský, J.: Interplay of Coreference and Discourse Relations: Discourse Connectives with a Referential Component. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 146-153. Copyright © European Language Resources Association, İstanbul, Turkey, ISBN 978-2-9517408-7-7 (2012).
6. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber B.: The Penn Discourse Treebank 2.0. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08)*, pp. 2961–2968. Marrakech, Morocco, ISBN 2-9517408-4-0 (2008).
WWW: <http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf>.
7. Rosen, A., Vavřín, M.: *Czech National Corpus – InterCorp*, Institute of the Czech National Corpus, Prague. 5. 2. 2018. WWW: <http://www.korpus.cz>.
8. Rysová, M., Rysová, K.: Secondary Connectives in the Prague Dependency Treebank. In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 291–299. Copyright © Uppsala University, Uppsala, Sweden, ISBN 978-91-637-8965-6 (2015).
9. Rysová, M., Rysová, K.: The Centre and Periphery of Discourse Connectives. In: *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pp. 452–459. Copyright © Department of Linguistics, Faculty of Arts, Chulalongkorn University, Bangkok, Thailand, ISBN 978-616-551-887-1 (2014).
10. Rysová, M., Synková, P., Mírovský, J., Hajičová, E., Nedoluzhko, A., Ocelák, R., Pergler, J., Poláková, L., Pavlíková, V., Zdeňková, J., Zikánová, Š.: *Prague Discourse Treebank 2.0*. Data/software, ÚFAL MFF UK, Prague, Czech Republic, http://hdl.handle.net/11234/1-1905 (2016).
11. Stede, M., Grishina, Y.: Anaphoricity in Connectives: A Case Study on German. In: *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016), co-located with NAACL 2016*, pp. 41–46. San Diego, California, (2016). WWW: <http://www.aclweb.org/anthology/W16-0706>.
12. Taboada, M.: Implicit and explicit coherence relations. In Renkema, J. (ed.): *Discourse, of course*, pp. 125–138. Amsterdam: John Benjamins (2009).
13. *Treq – Translation Equivalent Database*, Institute of the Czech National Corpus, Prague. 5. 2. 2018. WWW: <http://treq.korpus.cz/>.
14. Webber, B., Stone, M., Joshi, A., Knott, A.: Anaphora and discourse structure. *Computational Linguistics* 29(4), 545–587 (2003).
WWW: <https://dl.acm.org/citation.cfm?id=1105705>.
15. Zikánová, Š., Hajičová, E., Hladká, B., Jínová, P., Mírovský, J., Nedoluzhko, A., Poláková, L., Rysová, K., Rysová, M., Václ, J.: *Discourse and Coherence. From the Sentence Structure to Relations in Text.* Copyright © ÚFAL, Praha, Czechia, ISBN 978-80-904571-8-8, 274 pp. (2015).

Describing CzeDLex – a Lexicon of Czech Discourse Connectives

Magdaléna Rysová, Lucie Poláková, Jiří Mírovský, Pavlína Synková

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

Abstract. In the present contribution, we introduce a pilot version of CzeDLex, a Lexicon of Czech Discourse Connectives. Currently, CzeDLex contains 205 lemmas of connectives coming from the annotation of the Prague Discourse Treebank 2.0 (PDiT). CzeDLex reflects division of connectives into primary (e.g. *když* [if]) and secondary (e.g. *za této podmínky* [under this condition]). Altogether, 134 lemmas in CzeDLex are primary connectives and 71 are lexical cores of secondary connectives (i.e. words like *podmínka* [condition]). All 205 lemmas are manually annotated with basic linguistic information; the full annotation is now in progress. At this stage, 19 lemmas have been fully manually processed, which covers more than two thirds of all discourse relations in the PDiT. In the present paper, we describe the process of building CzeDLex, we give a list of connective properties annotated in lexicon entries of both primary and secondary connectives and we present the way of their nesting. The technical solution of CzeDLex is based on the (XML-based) Prague Markup Language that allows for an efficient incorporation of the lexicon into the family of Prague treebanks and also for interconnecting CzeDLex with existing lexicons in other languages.

Keywords: CzeDLex, Discourse Connectives, Lexicon.

1 Introduction

In the present contribution, we introduce a pilot version of CzeDLex (a Lexicon of Czech Discourse Connectives, developed within the COST-cz project TextLink-cz). The lexicon is a result of a long-term investigation of Czech discourse relations in both theoretical and practical aspects (see e.g. the monograph by Zikánová et al., 2015; summarizing research of coherence with focus on discourse relations in Czech) and logical follow-up of Prague annotation projects like Prague Discourse Treebank 1.0 (PDiT, see Poláková et al., 2012) and 2.0 (Rysová et al., 2016) – a large corpus annotated with discourse relations and discourse connectives. CzeDLex is thus based on an extensive linguistic research of discourse in Czech.

CzeDLex contains connectives partially automatically extracted from the PDiT 2.0. The lexicon entries are being manually checked and supplemented by additional lin-

guistic information, starting with the most frequent connectives. The current development version of the lexicon is available online (http://ufal.mff.cuni.cz/czedlex/) and was published as a pilot version (version 0.5, Mírovský et al., 2017) in the Lindat/Clarin repository.

The data format and the data structure of the lexicon are based on a study of similar existing resources, especially on DiMLex – a lexicon of German discourse markers first introduced by Stede and Umbach (1998) and Stede (2002) and recently updated by Scheffler and Stede (2016). The main principle adopted for nesting entries in CzeDLex is a semantic type of discourse relations expressed by the given connective word, which enables us to deal with a broad formal variability of connectives. The technical solution of CzeDLex is based on the (XML-based) Prague Markup Language that allows for an efficient incorporation of the lexicon into the family of Prague treebanks – it can be directly opened and edited in the tree editor TrEd (see Pajas and Štěpánek, 2008), processed from the command line in btred, interlinked with its source corpus and queried in the PML-Tree Query engine (details on PML-TQ are given in Štěpánek and Pajas, 2010) – and also for interconnecting CzeDLex with existing lexicons in other languages.

In this presentation, we first discuss theoretical linguistic aspects underlying the division and the description of Czech connectives adopted in CzeDLex, we present a list of connective properties annotated in the lexicon and finally, we provide an example of a lexicon entry (the connective *proto* [therefore]).

2 Theoretical Linguistic Aspects behind CzeDLex – Division of Connectives

CzeDLex reflects a division of discourse connectives into primary and secondary (the terms and definitions introduced by Rysová and Rysová, 2014) which differ especially in the degree of grammaticalization. Primary connectives are rather short and grammaticalized expressions belonging to certain parts of speech (mostly conjunctions, particles and some types of adverbs), such as English *but*, *or*, *when*, *thus*. On the other hand, secondary connectives are especially multiword phrases like *under these conditions*, *this means*, *because of this* etc. that are not yet fully grammaticalized. At the same time, secondary connectives contain the so-called core words, cf. e.g. the word *condition* in structures like *under this condition* or *on condition that* etc. (see also Rysová and Rysová, 2015).[1] Since the PDiT 2.0 contains a detailed annotation of both primary and secondary connectives, both of these types are included also into CzeDLex.

Discourse connectives in CzeDLex are further divided into the following categories: complex vs. single and modified vs. non-modified (Rysová, 2015). Complex connectives consist of two or more connective words all participating in expressing the given discourse relation type. Complex connectives occur in a single argument (*a*

[1] The annotation and description of primary connectives in the PDiT is given in Poláková (2015) and of secondary connectives in Rysová (2015).

proto [and therefore]) or they may form correlative pairs (*buď_nebo* [either_or]). Modified connectives contain an expression (often of evaluative or modal nature) that further specifies/modifies the discourse relation, without changing its semantic type (*hlavně protože* [mainly because]).

3 List of Connective Properties in CzeDLex

3.1 Level-One and Level-Two Entries

The entries in CzeDLex are structured according to a two-level nesting principle. On the first level, entries are nested according to the lemma of a connective and contain the following linguistic information:

• type of the connective (primary vs. secondary),
• structure of the connective (single vs. complex),
• variants of the connective (e.g. stylistic or orthographic),
• connective usages – a list of level-two entries representing semantico-pragmatic relations the connective expresses and their properties,
• non-connective usages – another list of level two entries, representing contexts where the lemma does not function as a discourse connective (e.g. *young and beautiful*).

Level two for primary connectives reflects the discourse-semantic types (usages) and contains the following pieces of information:

• semantic type of the discourse relation (condition, opposition etc.),
• gloss (an explanatory Czech synonym),
• English translation,
• part of speech of the connective,
• argument semantics (for asymmetric relations like reason–result, e.g. *protože* [because] expresses reason while *proto* [therefore] expresses result),
• ordering, i.e. position of the argument syntactically associated with the connective in relation to the other (external) argument,
• integration, i.e. placement of a connective in an argument,[2]
• list of connective modifications,
• list of complex connectives containing the given connective,
• examples from the PDiT (i.e. a context for the given discourse relation) and their English translations,
• is rare (set to '1' for rare usages),
• register (formal, neutral, informal).

[2] The names of the elements ordering and integration are taken from DiMLex (Scheffler and Stede, 2016).

An entry for a secondary connective contains several modifications. On level one of the lexicon structure, entries are nested according to the lemma of the core word for a secondary connective (see above). A level-two entry then contains the following additional properties (details on them are given in Rysová, 2015):

• syntactic characteristics of the structure (e.g. *za této podmínky* [under this condition] is a prepositional phrase),
• dependency scheme (general pattern) for each structure (e.g. *za této podmínky* [under this condition] = "*za* ((anaph. Atr) *podmínka*.2)", i.e. a preposition *za* [under] plus an anaphoric attribute and the word *podmínka* [condition] in genitive),
• realizations of the dependency scheme (e.g. *za této podmínky* [under this condition], *za dané podmínky* [under the given condition] etc.).

Details on building and designing of CzeDLex are given in detail in Mírovský et al. (2016), Synková et al. (2017) and Mírovský et al. (2017).

3.2 Frequencies from the PDiT 2.0

The lexicon entries are also enriched by frequencies of the individual connectives in the PDiT 2.0. Numbers of occurrences in the corpus are added to all connective variants, complex forms, modifications and realizations, as well as to connective and non-connective usages and the whole lemmas.

The numbers reflect the total occurrences as well as intra-sentential (as opposed to inter-sentential) occurrences using the whole PDiT 2.0 data.

3.3 Example of a Lexicon Entry

The following is a shortened entry for a connective *proto* [therefore] (e.g. we shortened or deleted too long context examples and their English translations for better readability of the entry).

We may read the following information from the entry. E.g. 99% of all of its tokens in the PDiT are in a connective usage (i.e. its non-connective usage is very rare – cf. an example from the PDiT where *proto* [therefore] does not connect two discourse arguments but only two sentence elements: *Ještě ne na světové úrovni, a právě proto tak rozkošně živoucí.* [Not yet on the world level and exactly therefore so adorably lively.]). 28% within all of its connective usages is intra-sentential, which demonstrates the preference of this connective in inter-sentential discourse relations.

We may see that most preferably (in 98%), the connective signals a relation of reason-result (semantically, it expresses result). It appears in the second discourse argument and concerning its integration, the connective is not strictly bound to any position in the sentence (it may be used e.g. in the first as well as in the second position). 19% within the reason-result relation is formed by complex forms like *a proto* [and therefore] or *proto také* [therefore also] and 1% by modified forms like *právě proto* [exactly therefore].

Concerning other semantic types of discourse relations, the connective *proto* [therefore] expresses also pragmatic reason-result or equivalence; however, these relations are rather rare in this case.

Fig 1. A shortened lexicon entry for the connective *proto* [therefore] in CzeDLex.

4 Conclusion

CzeDLex in its present version contains 205 lemmas (i.e. basic lemmas of primary and core words of secondary connectives) – all of them are manually annotated for modifications, complex forms, and variants. An additional manual annotation is provided to the most frequent ones, currently for primary connectives with at least 300 occurrences in the source corpus, and several most frequent secondary connectives.

Altogether, 19 lemmas have been fully manually processed, which covers more than two thirds of all discourse relations in the source corpus. Although the annotation of the rest of lemmas in CzeDLex is still in progress, we demonstrated that its first version offers valuable linguistic information already in its current form.

Acknowledgements

We acknowledge support from the Czech Science Foundation project no. GA17-06123S (*Anaphoricity in Connectives: Lexical Description and Bilingual Corpus Analysis*). This study has utilized the language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

1. Mírovský, J., Synková, P., Rysová, M., Poláková, L.: CzeDLex – A Lexicon of Czech Discourse Connectives. *The Prague Bulletin of Mathematical Linguistics*, No. 109, Copyright © Univerzita Karlova v Praze, Prague, Czech Republic, ISSN 0032-6585, 61–91 (2017).
2. Mírovský, J., Synková, P., Rysová, M., Poláková, L.: *CzeDLex 0.5*. Data/software, Charles University, Prague, Czech Republic, http://hdl.handle.net/11234/1-2538 (2017).
3. Mírovský, J., Synková, P., Rysová, M., Poláková, L.: Designing CzeDLex – A Lexicon of Czech Discourse Connectives. In: *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation*, pp. 449–457. Copyright © Kyung Hee University, Seoul, Korea, ISBN 978-89-6817-428-5 (2016).
4. Pajas, P., Štěpánek, J.: Recent Advances in a Feature-Rich Framework for Treebank Annotation. In: *The 22nd International Conference on Computational Linguistics – Proceedings of the Conference*, pp. 673–680. Copyright © The Coling 2008 Organizing Committee, Manchester, UK, ISBN 978-1-905593-45-3 (2008).
5. Poláková, L., Jínová, P., Zikánová, Š., Hajičová, E., Mírovský, J., Nedoluzhko, A., Rysová, M., Pavlíková, V., Zdeňková, J., Pergler, J., Ocelák, R.: *Prague Discourse Treebank 1.0*. Data/software, ÚFAL MFF UK, Prague, Czech Republic, http://ufal.mff.cuni.cz/pdit/ (2012).
6. Poláková, L.: *Discourse Relations in Czech*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic, 197 pp. (2015).
7. Rysová, M., Rysová, K.: Secondary Connectives in the Prague Dependency Treebank. In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 291–299. Copyright © Uppsala University, Uppsala, Sweden, ISBN 978-91-637-8965-6 (2015).
8. Rysová, M., Rysová, K.: The Centre and Periphery of Discourse Connectives. In: *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pp. 452–459. Copyright © Department of Linguistics, Faculty of Arts, Chulalongkorn University, Bangkok, Thailand, ISBN 978-616-551-887-1 (2014).
9. Rysová, M., Synková, P., Mírovský, J., Hajičová, E., Nedoluzhko, A., Ocelák, R., Pergler, J., Poláková, L., Pavlíková, V., Zdeňková, J., Zikánová, Š.: *Prague Discourse Treebank 2.0*. Data/software, ÚFAL MFF UK, Prague, Czech Republic, http://hdl.handle.net/11234/1-1905 (2016).
10. Rysová, M.: *Diskurzní konektory v češtině (Od centra k periferii)* [Discourse connectives in Czech (From Centre to Periphery)]. Ph.D. thesis, Charles University in Prague, Prague, Czechia, 268 pp. (2015).

11. Scheffler, T., Stede, M.: Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In: *Proceedings of LREC 2016*, pp. 1008–1013. European Language Resources Association, Paris, France (2016).
12. Stede, M., Umbach, C.: DiMLex: A Lexicon of Discourse Markers for Text Generation and Understanding. In: *Proceedings of Coling 1998*, pp. 1238–1242. Association for Computational Linguistics (1998).
13. Stede, M.: DiMLex: A lexical approach to discourse markers. In: *Exploring the Lexicon – Theory and Computation*. Alessandria (Italy): Edizioni dell'Orso (2002).
14. Synková, P., Rysová, M., Poláková, L., Mírovský, J.: Extracting a Lexicon of Discourse Connectives in Czech from an Annotated Corpus. Accepted for publication in: *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pp. 1–9. Copyright © Computing Society of the Philippines, Cebu, Philippines (2017).
15. Štěpánek, J., Pajas, P.: Querying Diverse Treebanks in a Uniform Way. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 1828–1835. Copyright © European Language Resources Association, Valletta, Malta, ISBN 2-9517408-6-7 (2010).
16. Zikánová, Š., Hajičová, E., Hladká, B., Jínová, P., Mírovský, J., Nedoluzhko, A., Poláková, L., Rysová, K., Rysová, M., Václ, J.: *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Copyright © ÚFAL, Praha, Czechia, ISBN 978-80-904571-8-8, 274 pp. (2015).

Annotation proposal for *Sp*. DRDs and their interaction with other units in written texts: an analysis from the Val.Es.Co. discourse segmentation model

Discourse segmentation is a big research field highly influenced by different theoretical approaches: "macro-syntax (Van Dijk 1977), Conversation Analysis (Sacks et al. 1974) or Discourse Analysis (Sinclair and Coulthard 1992)", among others (Pons Bordería 2014a: 1). Works on discourse segmentation attempt to analyze a broad spectrum of discourses and genera by applying a set of segmentation units and sub-units. Each discourse segmentation model in Romance Languages (*Basel Model, Geneva Model, Freiburg Model, Val.Es.Co. Model, Co-Enunciation Model, Prominence-Demarcation Model, Basic Discourse Units Model*) comprises different units and sub-units depending on the type of discourse that they can address (conversations, interviews, journals, chats, specialized texts, etc.) (De Cesare/ Borreguero 2014: 56). Some of these units are:

- Paragraphs, Textual Movements, Nucleus, Background Frame, Background Appendix (Ferrari et al. 2008; Pons Bordería 2014: 9)
- Incursions, Exchanges, Move, Act (Roulet 1991; Roulet et al. 2001; Pons Bordería 2014: 11)
- Morphemes, Clauses, Enunciation, Periods (Groupe de Friburg 2012; Pons Bordería 2014: 12)
- Rheme, Preambule, Framework, Disjointed lexical support (Pons Bordería 2014: 16)
- BDU, Congruent, Intonation-bound, Syntax-bound, Regulatory (Degand/ Simon 2009; Pons Bordería 2014: 19)
- Pitch, Accent, Intensity, Syllable length (Lombardi Vallauri 2014; Pons Bordería 2014: 16)

Discourse analyses based on discourse segmentation methods allow more accurate explanations of: (i) general patterns at organizing contents in different types of discourse; and (ii) pragma-discursive phenomena that cannot be addressed by traditional grammatical and syntactic units (Narbona 1988). Some general interests in discourse segmentation are:

- Organization of information and degrees of informativity;
- Organization of topics -digressions, change of topic- and topicalization items;
- Formulation/ enunciation procedures and formulation items -reformulations, paraphrases, corrections-;
- Hedging, stressing, intonation marks;
- The role played by different linguistic items within the structure of texts and conversations (e.g. DRDs, constructions, pronouns, etc.);
- Distribution of contents and phenomena over various hierarchical levels.

This presentation shows the applicability of the Val.Es.Co. model of discourse segmentation units (Briz et al. 2003; Grupo Val.Es.Co. 2014) at describing and annotating DRDs in Spanish textual discourses. Some previous studies (Pons Bordería 2014b) have applied the Val.Es.Co. model to diachronic analyses whose main data-base are textual discourses. These studies have been a challenge for the Val.Es.Co. model, conceived as

a tool for the analysis of colloquial conversations. Since results obtained in these diachronic analyses based on Val.Es.Co. have been clear and systematic, one of the last aims within the group is to extend its units and sub-units to textual discourses and analyze them from a synchronic perspective.

A set of texts have been segmented and annotated. We specially have focused our attention on the different DRDs employed and their relationship with the rest of contents within the discourse structure. The set of texts analyzed have been mainly retrieved from journals and written corpus, such as the Real Academia Española CREA and CORPES, as well as from the Diccionario de partículas discursivas del español -DPDE, Briz, Pons and Portolés 2008). Last, the labels for the annotation are based on the different Val.Es.Co. discourse units (Subact, Act, Intervention, Exchange, Turn, Turn-exchange, Dialogue, Discourse), positions (Initial, Medial, Final, Independent) and levels (Monological, Dialogical).

Some of the issues addressed by the application of the Val.Es.Co. model are:

- The position and unit occupied by different DRDs within different levels in textual discourses, and their interaction with previous and subsequent contents. Informativity, new topics, argumentative moves, etc., and their DRDs can be analyzed within the Val.Es.Co. model;
- The status of some *Sp.* adverbs and connectives (also treated as DRDs) whose boundaries are not clear (*pues*, *entonces*, *y luego*, *justamente*): despite being used in textual contexts, they are located between the dictus and the modus, closer to some discourse markers;
- Discursive subordinations produced through DRDs (reformulations and neighbor functions -paraphrase, correction, explanation, summary, etc.-). These subordinations are not lineal and sometimes can contain further subordinations (digressions). Some of these *Sp.* DRDs are: *o sea*, *bueno, quiero decir, es decir*, etc.)

REFERENCES

Briz, Antonio and Grupo Val.Es.Co. 2003. "Un sistema de unidades para el estudio del lenguaje coloquial." *Oralia 6*, 7-61.

Briz, Antonio; Pons, Salvador and José Portolés. 2008. *Diccionario de partículas discusivas del español*. Available online: http://www.dpde.es

Degand, Liesbeth, and Anne Catherine Simon. 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.

De Cesare, Anna-Maria and Margarita Borreguero. 2014. "The contribution of the Basel model to the description of polyfunctional discourse markers". In Pons Bordería, S. (ed.): *Models of Discourse Segmentation. Explorations across Romance Languages*, 55-95.

Ferrari, Angela, Luca Cignetti, and Anna-Maria De Cesare et al. 2008. *L'interfaccia linguatesto. Natura e funzioni dell'articolazione informativa dell'enunciato*. Alessandria: Edizioni dell'Orso.

Grupo Val.Es.Co. 2014. "Las unidades del discurso oral. La propuesta Val.Es.Co. de segmentación de la conversación (coloquial)." In *Estudios de Lingüística del Español 35*, ed. by Luis Cortés, 13-73.

Groupe de Fribourg (A. Berrendonner, dir.) 2012. *Grammaire de la période.* Berne: Peter Lang.

Lombardi Vallauri, Eduardo. 2014. "The topologic hypothesis on prominence as a cue to information structure in Italian". In Pons Bordería, S. (ed.): *Models of Discourse Segmentation. Explorations across Romance Languages*, pp. 219-243.

Narbona Jiménez, Antonio. 1988. Sintaxis coloquial: problemas y métodos. *LEA* X (1): 81-106

Pascual, Elena. 2014. *Aproximación a la segmentación del subacto en la conversación coloquial española.* Master dissertation. Valencia: University of Valencia.

Pons Bordería, Salvador, ed. 2014a. *Models of Discourse Segmentation. Explorations*

across Romance Languages. Amsterdam: John Benjamins

Pons Bordería, Salvador. 2014b. Paths of grammaticalization in Spanish *o sea*. En Ch.

Ghezzi y P. Molinelli, eds. *Pragmatic Markers from Latin to Romance Languages*.

Oxford, OUP, pp. 108-135

Roulet, Eddy. 1991. Vers une approche modulaire de l'analyse du discours". *Cahiers de Linguistique Française* 12: 53-81

Roulet, Eddy; Fillietaz, Laurent; Grobet, Anne. 2001. *Un modéle et un instrument d'analyse de l'organisation du discours*. Berna: Peter Lang

Sacks, Harvey; Schegloff, Emanuel; Jefferson, Gail. 1974. A Symplest Systematics for

the Organization of Turn-Taking for Conversation, *Language* 50/4: 696-735

Sinclair, John; Coulthard, Malcolm. 1992. Towards an analysis of discourse. En M. Coluthard, ed. *Advances in spoken discourse analysis*. London/New York:

Routledge, pp. 1-35

Van Dijk, Teun A. 1977. *Text and context: Explorations in the semantics and pragmatics of discourse.* London: Addison-Wesley Longman Limited.

Unifying dimensions in coherence relations

How various annotation schemes are related

Ted J.M. Sanders,[1] Vera Demberg,[2] Jet Hoek,[1] Merel C.J. Scholman,[2]

Sandrine Zufferey,[3] and Jacqueline Evers-Vermeul[1]

[1] Utrecht University
[2] Saarland University
[3] University of Bern

In recent decades, linguistics has seen major developments in the area of corpus linguistics. Large corpora allow us to obtain qualitative and quantitative observations about language use. For a long time, corpus annotation was limited to annotation at the morphological, syntactic or semantic level, but over the last fifteen years the annotation of corpora at the discourse level has been realized in large annotation efforts. Leading examples of discourse annotation frameworks include the Penn Discourse Treebank (Prasad et al. 2008) the Rhetorical Structure Theory Discourse Treebank (RST-DT, Carlson and Marcu 2001), and the Segmented Discourse Representation Theory (SDRT; Asher and Lascarides 2003).

This development enables us to take the study of coherence relations an important step forward. Corpora can now be searched for coherence relations, whether they remain implicit, or are linguistically marked by cue phrases or connectives. For instance, looking at all annotated occurrences of a connective like English since allows us to determine how often and under which circumstances since expresses a TEMPORAL relation, as in (1) or a CLAIM- ARGUMENT relation, as in (2). In addition, we can search corpora for cases in which alternative connectives are used to express the same relation, as in (3), which also expresses a claim-argument relation, or cases in which the same relation is conveyed implicitly as in (4). Analyzing such cases in a qualitative and quantitative way provides us with important insights into a connective's distribution over coherence relations.

(1) Since Cruijff[1] played on the team, they never lost a game.
(2) It was impossible they would lose the game, since Cruijff played on the team.
(3) It was impossible they would lose the game, because Cruijff played on the team.
(4) It was impossible they would lose the game. Cruijff played on the team.

From annotations in the Penn Discourse Treebank, we know that more than half of all coherence relations are not explicitly marked by a connective or cue phrase. This observation raises several important issues for discourse annotation. One question is whether some relation types are more often conveyed implicitly or by the use of alternative signals than others, and if so, what the causes of these differences are (Asr and Demberg 2012; Das and Taboada 2017; Hoek, Zufferey et al. 2017). To address these issues, the development of extensive and comparable sets of annotated data with coherence relations across several languages and genres will represent a major step ahead.

In other words, the existence of discourse-annotated corpora is crucial to the field of discourse studies and language use. Therefore, it would be worthwhile to make the various annotated corpora accessible to and comparable for all researchers in the field.[2] At present this is not yet possible. While there is a large consensus regarding the usefulness of discourse-annotated data, there are many alternative ways of annotating coherence relations, and discourse annotation schemes differ strongly in the type of coherence relations that are distinguished, varying from sets of approximately 20 relations (such as the original RST), others of only two relations (Grosz and Sidner 1986). 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

[1] Johan Cruijff (1947) was the best Dutch soccer player ever, and one of the best in the world; he passed away on March 24, 2016.
[2] This is a central goal of the EU-COST TextLink project.

distinguishes 78 relations that can be partitioned into 16 classes (Carlson and Marcu 2001).

The annotation schemes do not only differ in granularity, but also in their choice of labels: different labels are used for the same conceptual relations, and the same labels are used for different relation sense definitions. This makes it extremely difficult to make comparisons across corpora that are annotated according to different frameworks.

We propose a way to "translate" annotation tags from one framework to the terminology of other frameworks, so that the different annotation systems can "talk to each other". Our concrete goal is to develop an interface that will allow researchers to find identical or at least closely related relations within a set of annotated corpora, even if these relations carry different names in the respective frameworks in which they were annotated. To make this goal more concrete: imagine a discourse researcher who uses the PDTB framework, and who is interested in REASON relations in English. She might want to know how often these relations are made explicit with connectives like because or since, and how often and under which circumstances they remain implicit. She knows the labels provided in the PDTB, but in order to benefit from other annotated corpora, she also needs to know what labels to search for in RST-DT or SDRT in order to retrieve similar relations. Our interface will allow her to do exactly that: start from the tag REASON, and find similar or closely related relations in other frameworks (for example, relations labeled as RESULT in SDRT for this particular case), so that her research corpus is larger. Being able to use several discourse-annotated corpora at the same time, instead of just one, multiplies the amount of available data and unlocks a whole new set of research possibilities for the whole community.

We see several advantages of such an interface. It will allow researchers in the field of discourse to answer research questions like the ones mentioned above, making use of all annotated corpora, from all frameworks. Furthermore, the mapping will be useful for researchers and engineers working on automatic coherence relation labeling. Many natural language processing tasks, such as information retrieval and question-

answering systems, text summarization systems, and machine translation systems would improve from increased performance in automated coherence relation classification. Current state-of-the-art coherence relation classification systems (see Xue et al. 2015 for an overview) make use of human-annotated coherence relations in corpora for training, especially the large resources PDTB and RST-DT. The performance of these tools, and generalizability from one text type to another would likely improve if more training data could be used. The mapping proposed here would enable researchers to train their models on all of the annotated resources, and not just those corresponding to a specific framework. Finally, we believe that a mapping between frameworks might help us extend current theories of coherence relations, because it will improve our understanding of the features defining different types of coherence relations, and pinpoint the exact differences and similarities between existing frameworks.

In this paper, we show how three often used and seemingly different discourse annotation frameworks – PDTB, RST and SDRT – can be related by using a set of unifying dimensions. These dimensions are taken from the Cognitive approach to Coherence Relations (Sanders et al. 1992, 1993; Scholman et al, 2016), and combined with more fine-grained additional features from the frameworks themselves to yield a posited set of dimensions that can successfully map three frameworks. The resulting interface will allow researchers to find identical or at least closely related relations within sets of annotated corpora, even if they are annotated within different frameworks. Furthermore, we tested our unified dimension (UniDim) approach by comparing PDTB- and RST-annotations of identical newspaper texts and converting their original end-label annotations of relations into the accompanying values per dimension. Subsequently, rates of overlap in the attributed values per dimension were analyzed. Results indicate that the proposed UniDim-dimensions indeed create an interface that makes existing annotation systems "talk to each other".

References

1. Asher, Nicholas & Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.
2. Asr, Fatemeh Torabi & Vera Demberg. 2012. Implicitness of discourse relations. *Proceedings of COLING*. Mumbai, India.
3. Das, Depodam & Maite Taboada. 2017. RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources & Evaluation*. 1-36.
4. Carlson, Lynn & Daniel Marcu. 2001. Discourse tagging reference manual. Grosz, Barbara & Candace Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics* 12(3). 175-204.
5. Hoek, Jet, Sandrine Zufferey, Jacqueline Evers-Vermeul & Ted Sanders. 2017. Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics* 121. 113-131.
6. Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi & Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. *Proceedings of the 6th International Conference of Language Resources and Evaluation (LREC 2008)*. Marrakech: Morocco.
7. Sanders, Ted, Wilbert Spooren & Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15. 1-35.
8. Sanders, Ted, Wilbert Spooren & Leo Noordman. 1993. Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics* 4(2). 93-133.
9. Scholman, Merel, Jacqueline Evers-Vermeul & Ted Sanders. 2016. Categories of coherence relations in discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue and Discourse* 7(2). 1-28.
10. Xue, Nianwen, Hwee Ng, Sameer Pradhan, S., Rashmi Prasad, Christopher Bryant & Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. *Proceedings of the 19th Conference on Computational Natural Language Learning: Shared Task*, 1-16. Beijing, China.

A multilingual database of connectives: connective-lex.info

Tatjana Scheffler, Manfred Stede, Peter Bourgonje, and Felix Dombek

Computational Linguistics
UFS Cognitive Science
University of Potsdam, Germany
firstname.lastname@uni-potsdam.de

Despite some progress made very recently, human- and machine-readable resources providing information about the syntactic, semantic, and pragmatic behavior of connectives are still scarce. In order to encourage research on languages for which a connective lexicon is not yet available, we have produced a web-based database that provides access to existing resources, and is built in such a way that new ones can be straightforwardly added to it. The required "common denominator" for the individual language lexicons, to be described in more detail below, is a relatively simple technical format (in XML), and more importantly, the presence of compatible information in the lexicons. This enables multilingual queries across all integrated lexicons, using a set of syntactic categories (subordinating and coordinating conjunctions; adverbials; prepositions), as well as semantic/pragmatic relations (the PDTB 3.0 sense hierarchy (Webber et al., 2016)). In the Fall of 2017, we made the database available at http://connective-lex.info, and it has already generated interest from other researchers who are in the process of adding two more languages to it (viz. Arabic and Dutch).

Currently, discourse connective lexicons for the following languages (in alphabetical order) are made accessible through the multilingual interface:

English: We extracted the connectives and their syntactic and semantic properties as annotated in the Penn Discourse Treebank corpus (Prasad et al., 2008).

French: We built a sense-mapping so that LexConn (Roze et al., 2012), which uses SDRT relations, could be integrated.

German: DiMLex (Stede, 2002) was the original basis for the system and recently underwent substantial extension and addition of PDTB3 senses (Scheffler & Stede, 2016).

Italian: LiCo (Feltracco et al., 2016) has been modelled along the lines of DiMLex, and thus was easy to integrate.

Portuguese: The format of LDM-PT (Mendes & Lejeune, 2016) was also inspired by DiMLex, and we could map it to the common schema.

In the following, we describe the database interface, the underlying lexicon schema, and the process of manual or corpus-based creation of discourse connective lexicons for new languages in detail.

1 Lexicon Schema

The lexicons are represented in a format based on the German DiMLex connective lexicon XML schema. For the multilingual version, it was important to define a common core of connective information, so that the effort for building and integrating new lexicons is relatively low. The schema requires the lexicons to specify information on the spelling variants of each connective item, its syntactic categories, and its possible semantic senses. As is reflected by the XML structure, the semantic senses are subordinate to the syntactic category, meaning that different senses can be assigned for different syntactic categories of a connective. Examples can be provided optionally for each connective sense. If the lexicons contain additional language-specific information which conforms to DiMLex 2.0's schema, this data is also displayed with the results (but currently not searchable). The German lexicon for example provides ordering constraints for adverbials and subordinating conjunctions, as well as an indication of the possible use of a connective as a focus particle.

```
<entry id="25" word="in addition">
    <orths>
       <orth canonical="0" orth_id="25o1" type="cont">
          <part type="phrasal">In addition</part>
       </orth>
       <orth canonical="1" orth_id="25o2" type="cont">
          <part type="phrasal">in addition</part>
       </orth>
    </orths>
    <syn>
       <cat>PP</cat>
       <sem>
          <pdtb2_relation anno_N="165" freq="165"
                          sense="Expansion.Conjunction" />
       </sem>
    </syn>
</entry>
```

Fig. 1. Example entry on the underlying connective lexicons.

Figure 1 shows an example of an underlying lexicon entry from the English lexicon. Each discourse connective entry contains information about the orthography of the item, including whether a phrasal item has to occur continuously or not. The schema allows for several alternative spellings of each item (`<orth/>`), in addition to the canonical one. These variants can be optionally displayed in the web interface (see below). Each entry specifies the syntactic categories (`<syn/>`) available for the connective. In turn, each syntactic subentry of a connective lists the semantic relations (`<sem/>`) available for this syntactic category. Semantic

types are associated with specific syntactic incarnations of a connective item because the available semantic readings depend on the syntactic category of a connective in some languages.

The current database distinguishes the semantic senses proposed in the updated PDTB 3.0 (Webber et al., 2016) sense hierarchy and the following basic types of syntactic categories:

- coordinating conjunction (cco)
- subordinating conjunction (csu)
- adverb (adv)
- preposition (prep)
- other

However, for each underlying connective lexicon, the original data is retained and mapped to the database categories on-the-fly based on provided mapping tables. The setup is therefore flexible to changes or updates of the categories, as well as in the mapping, if needed. Adding a syntactic category (for example, for a subtype of adverbs) would merely require an updated mapping table for the individual lexicons' syntactic annotations to this new kind of category.

2 Extraction of a Lexicon from a Corpus

It is possible to extract a compatible discourse connective lexicon from a corpus annotated with explicit connectives (e.g. in PDTB style). For this work, we have done this in order to obtain the English lexicon. Since no machine readable English discourse connective lexicon was publicly available, we have extracted the lexical information from the PDTB2 annotations (Prasad et al., 2008). The lexicon is available at `https://github.com/TScheffler/Connectives`. For each explicit connective token in the corpus, we extracted the connective head and possible modifiers[1], its syntactic category (the part of speech or phrasal category covering the explicit connective), and its semantic sense (according to the PDTB2 annotation). For syntactic and semantic annotations, quantitative data are also retained, recording how often each category/sense was annotated for each connective, since this may provide useful information for disambiguation.[2] For example, the entry from the English lexicon shown in Figure 1 specifies that the Expansion.Conjunction relation was seen in 165 out of 165 cases for "[Ii]n addition". The PDTB2 senses found in the corpus are mapped to PDTB3 senses on-the-fly when using the lexicon in the web-based database. We provide the mapping table in Appendix A. Where senses cannot be matched exactly, we back off to the broader level 2 senses, following the idea that a lexicon should list the possible meanings of a connective as comprehensively as possible (and leaving disambiguation of individual cases aside).

[1] Only the heads are currently included in the lexicon.

[2] For connectives annotated with two senses, both primary and secondary sense are counted.

3 Web Interface

The web interface to the multilingual database is pictured in Figure 2. Initially, it shows the list of available lexicons and several options for searching them. For each lexicon, metadata including the lexicon's authors, license, release date, and reference publications are included. Users must select one or more lexicons to include in the search. The interface allows the user to search for specific words, for connective entries which have particular features, i.e.: a syntactic class, a PDTB3 sense, or combinations thereof. Currently, features and search terms can only be combined intersectively. For illustration, one can obtain:

- all information on a specific connective, such as the French "parce que"
- all subordinating conjunctions in one language
- all connectives that can signal a particular relation (e.g., CONCESSION) in various languages

Fig. 2. Example view of the web interface `connective-lex.info`

Figure 2 shows the results of a search for coordinating conjunctions expressing CONTRAST relations in the English, French, and German lexicons. The results pane on the right-hand side shows a summary of the matches in each language (in this case, 5 German, 2 French, and 5 English coordinating conjunctions were found), followed by the list of matching connectives, sorted by language and alphabetically.

4 Technical Details

The web application consists of a frontend which runs in the browser and a backend which runs on the server. It is implemented using modern web app technologies: HTML5, CSS3, and JavaScript (with AJAX). The PHP backend mainly hosts the lexicon data. The frontend handles all queries autonomously in the browser, loads the lexicons from the backend, and displays the interface. Frontend and backend communicate primarily via JSON. For this reason, all the lexicons have been converted into a space-conserving JSON format.

The web application is a display interface only, and the results of searches cannot be exported. The main purpose of the application is to enable browsing and explorations of the included lexicons. For further analyses, all the lexicons are independently available from their respective authors (with links provided in `connective-lex.info`) in XML or similar format.

5 Adding New Lexicons

The database and web app can be easily extended by adding new lexicons (for existing or new languages). Lexicons that follow (a subset of) the DiMLex format (Scheffler & Stede, 2016) can be added to the backend's lexicon directory by the administrator. In addition, a metadata file must be created with authorship and license information. In order to map an individual lexicon's syntactic and semantic annotation to a common interface, the application uses a syntactic and a semantic mapping table. If the new lexicon uses a new syntactic tagset or a new sense inventory, these mappings need to be provided as well.

At present, we are beginning to work on integrating an Arabic lexicon provided by Keskesa et al. (2014), and we started building a lexicon for Dutch in collaboration with colleagues in the Netherlands.

6 Summary

We have introduced the `connective-lex.info` web application, which allows interactive queries in multilingual connective lexicons. The application currently supports simultaneous search in lexicons in five languages, but is easily extensible with other machine-readable lexicons. Users of the application can compare connectives across different languages with respect to their basic syntactic and semantic properties.

Bibliography

Feltracco, Anna, Elisabetta Jezek, Bernardo Magnini, & Manfred Stede 2016. LICO: A Lexicon of Italian Connectives. In Proceedings of the 3rd Italian Conference on Computational Linguistics (CLiC-it), Napoli, Italy.

Keskesa, Iskandar, Farah Benamara Zitoune, & Lamia Hadrich Belguith 2014. Learning explicit and implicit Arabic discourse relations. Journal of King Saud University - Computer and Information Sciences, 26(4):398–416.

Mendes, Amália, & Pierre Lejeune 2016. LDM-PT. A Portuguese Lexicon of Discourse Markers. In Conference Handbook of TextLink – Structuring Discourse in Multilingual Europe Second Action Conference, Budapest, Hungary.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, & Bonnie Webber 2008. The Penn Discourse Treebank 2.0. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrackech, Morocco.

Roze, Charlotte, Laurence Danlos, & Philippe Muller 2012. LEXCONN: A French Lexicon of Discourse Connectives. Discours [En ligne], 10. http://discours.revues.org/8645.

Scheffler, Tatjana, & Manfred Stede 2016. Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), Portoroz, Slovenia.

Stede, Manfred 2002. DiMLex: A Lexical Approach to Discourse Markers. In Exploring the Lexicon - Theory and Computation. Edizioni dell'Orso, Alessandria.

Webber, Bonnie, Rashmi Prasad, Alan Lee, & Aravind Joshi 2016. A Discourse Annotated Corpus of Conjoined VPs. In Proceedings of the 10th Linguistic Annotation Workshop (LAW-X), Berlin, Germany.

A Mapping from PDTB 2.0 senses to PDTB 3.0

PDTB 2.0 sense	PDTB 3.0 sense
Comparison	COMPARISON
Comparison.Concession	COMPARISON:Concession
Comparison.Concession.Contra-expectation	COMPARISON:Concession
Comparison.Concession.Expectation	COMPARISON:Concession
Comparison.Contrast	COMPARISON:Contrast
Comparison.Contrast.Juxtaposition	COMPARISON:Contrast
Comparison.Contrast.Opposition	COMPARISON:Contrast
Comparison.Pragmatic concession	COMPARISON:Concession+SpeechAct
Comparison.Pragmatic contrast	COMPARISON:Similarity
Contingency	CONTINGENCY
Contingency.Cause.Reason	CONTINGENCY:Cause:Reason
Contingency.Cause.Result	CONTINGENCY:Cause:Result
Contingency.Condition	CONTINGENCY:Condition
Contingency.Condition.Factual past	CONTINGENCY:Condition
Contingency.Condition.Factual present	CONTINGENCY:Condition
Contingency.Condition.General	CONTINGENCY:Condition
Contingency.Condition.Hypothetical	CONTINGENCY:Condition
Contingency.Condition.Unreal past	CONTINGENCY:Condition
Contingency.Condition.Unreal present	CONTINGENCY:Condition
Contingency.Pragmatic cause.Justification	CONTINGENCY:Cause+belief
Contingency.Pragmatic condition.Implicit assertion	CONTINGENCY:Condition+SpeechAct
Contingency.Pragmatic condition.Relevance	CONTINGENCY:Condition+SpeechAct
Expansion	EXPANSION
Expansion.Alternative	EXPANSION:Disjunction
Expansion.Alternative.Chosen alternative	EXPANSION:Disjunction
Expansion.Alternative.Conjunctive	EXPANSION:Disjunction
Expansion.Alternative.Disjunctive	EXPANSION:Disjunction
Expansion.Conjunction	EXPANSION:Conjunction
Expansion.Exception	EXPANSION:Exception
Expansion.Instantiation	EXPANSION:Instantiation
Expansion.List	EXPANSION:Conjunction
Expansion.Restatement	EXPANSION:Level-of-detail
Expansion.Restatement.Equivalence	EXPANSION:Equivalence
Expansion.Restatement.Generalization	EXPANSION:Level-of-detail:Arg1-as-detail
Expansion.Restatement.Specification	EXPANSION:Level-of-detail:Arg2-as-detail
Temporal	TEMPORAL
Temporal.Asynchronous	TEMPORAL:Asynchronous
Temporal.Asynchronous.Precedence	TEMPORAL:Asynchronous:Precedence
Temporal.Asynchronous.Succession	TEMPORAL:Asynchronous:Succession
Temporal.Synchrony	TEMPORAL:Synchronous

For example, specifically, or because; Individual differences in coherence relation interpretation biases?

M.C.J. Scholman,[1] Vera Demberg,[1] and Ted J.M. Sanders[2]

[1] Saarland University
[2] Utrecht University

In order to comprehend a text, readers must construct a coherent representation of the discourse segments and the *coherence relations* that connect these segments (cf. Hobbs, 1979, Mann & Thompson, 1988; Sanders, Spooren & Noordman, 1992). Coherence relations are semantic-pragmatic links between two (or more) discourse units. They can be explicitly signaled by connectives such as *because* or *for example*. However, many relations are implicit, that is, they are not marked by a connective, as in Example 1.

(1) Packaging has some drawbacks. The additional technology, personnel training and promotional effort can be expensive.
wsj_0085

To understand this sentence pair properly, the reader would have to infer the coherence relation between these two sentences, but multiple relation senses can be inferred for this example: the second argument can be interpreted as providing examples of the drawbacks (marked by *for example*), specifying what exactly the drawbacks are (*specifically*), and/or giving an argument for the claim that there are drawbacks (*because*).[1] In the current contribution, we investigate whether readers systematically differ in how they interpret relations that can have multiple readings.

[1] This example was in fact annotated as INSTANTIATION in the PDTB.
[2] This is not to say that a single connective cannot mark multiple types of relations; connectives are in fact known to be ambiguous and multifunctional (see Asr & Demberg, 2013; Degand, 1998, among many others). However, the method is based on the assumption that readers choose the connective that best matches the

Many different types of relations can co-occur together, but in the current experiment we focus on two particular relation types: SPECIFICATIONS and INSTANTIATIONS. In both of these relation types, one segment further specifies a set or situation described in the other segment (Halliday, 1994). SPECIFICATIONS and INSTANTIATIONS are elaborative relations: the second argument of the relation (Arg2) elaborates on the first argument (Arg1) by specifying or instantiating something mentioned in Arg1. However, certain SPECIFICATIONS and INSTANTIATIONS can have an additional interpretation: Arg2 can also be interpreted as an argument for a claim proposed in Arg1 (the argumentative function of SPECIFICATIONS and INSTANTIATIONS). This double function was brought up by Carston (1997, p. 164), who noted that "exemplification is a common way of providing evidence to support a claim, or, equivalently, of giving a reason for believing something." Building on this, Blakemore (1997) argues that SPECIFICATIONS and INSTANTIATIONS can have different functions in a text, and that classifying them as only elaborative or argumentative does not do justice to the way these relations are interpreted. In a crowdsourcing connective insertion experiment investigating how readers interpret these relations, Scholman and Demberg (2017) found that readers do interpret certain SPECIFICATION and INSTANTIATION items as argumentative as well (not only elaborative), but they did not find evidence that readers interpret both functions simultaneously for a single item.

In the current study, we investigate whether readers show biases when interpreting such multi-interpretable relations, and whether readers differ from each other in their biases (i.e., whether there is individual variability). We asked participants to insert a connective from a predefined list between the two arguments of implicit INSTANTIATIONS and SPECIFICATIONS. This allows us to tap into their interpretations of these relations. This study was exploratory and focused on one group of readers: highly educated native English speakers.

Method

Participants – 92 native English speakers completed the experiment. All participants had an educational level higher than undergraduate.

Items – The material consisted of 24 SPECIFICATION and INSTANTIATION relations (all originally implicit) from the Penn Discourse Treebank (Prasad et al., 2008). These items were also included in a previous connective insertion experiment (Scholman and Demberg, 2017), and were chosen based on the requirement that both the elaborative and the argumentative function were inferred in that experiment (i.e., for every item, certain participants inferred the elaborative reading, and other the argumentative reading). Fillers consisted of 40 causal, additive, contrastive and concessive relations.

Connective list – Participants were presented with a list of connectives that typically mark our target relations.[2] The list was constructed based on a classification from Knott and Dale (1994) and consisted of: *as an illustration* (indicating an INSTANTIATION relation), *more specifically* (SPECIFICATION), *because*, *as a result* (both CAUSE), *by contrast* (CONTRAST), *even though*, *nevertheless* (both CONCESSION), and *in addition* (CONJUNCTION).

Procedure – The items were divided over four batches, with 6 experimental and 8 filler items per batch. Participants were recruited via the crowdsourcing platform Prolific. They completed all batches over a period of four months. This allows us to examine how readers interpret these relations, and to compare an individual's distribution of insertions between different batches.

The order of the batches was randomized. For every batch, item order was randomized, and for every trial, connective order was randomized as well. Participants were

[2] This is not to say that a single connective cannot mark multiple types of relations; connectives are in fact known to be ambiguous and multifunctional (see Asr & Demberg, 2013; Degand, 1998, among many others). However, the method is based on the assumption that readers choose the connective that best matches the strongest reading that they infer.

instructed to drag and drop the connecting phrase that best expressed the meaning of the two relational segments. They could choose multiple connectives, or choose none of the connectives, in which case they were prompted to provide another connective (not from the list). Figure 1 shows an example of the experiment interface.

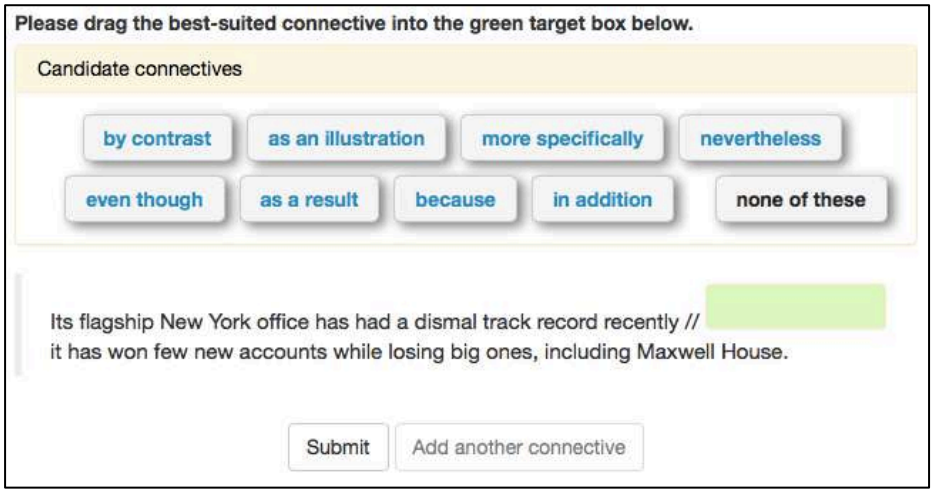


Figure 1. Experiment interface.

Results

We only consider insertions in experimental items in our analyses. Before going into the statistics, we will consider some graphical illustrations of the data. First, we classified participants in three groups: participants who interpreted items mainly as argumentative in at least three of four batches were classified as having an argumentative bias (22 participants); participants who interpreted items as mainly elaborative in at least three batches were classified as having an elaborative bias (37 participants); and the remaining participants were classified as having no bias (33 participants). The insertions of participants in each group were grouped together, as shown in Figure 2.

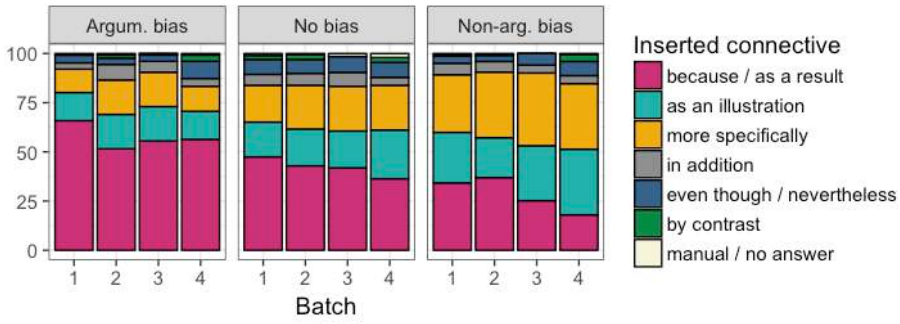


Figure 2. Insertions per batch for the three groups of participants.

Next, we ordered the participants' insertions in the fourth batch according to their "argumentative" bias displayed in the previous three batches: their insertions in the first three batches are coded as argumentative vs. elaborative, and the average of argumentative insertions represents a participant's argumentative bias. In the next two figures, every bar represents the insertions of one participant for the six items in the fourth batch; every color represents a connective type. Figure 3a displays the insertions for all participants; Figure 3b displays the insertions of participants with an argumentative or elaborative bias (59 participants, excluding those who did not display a bias).

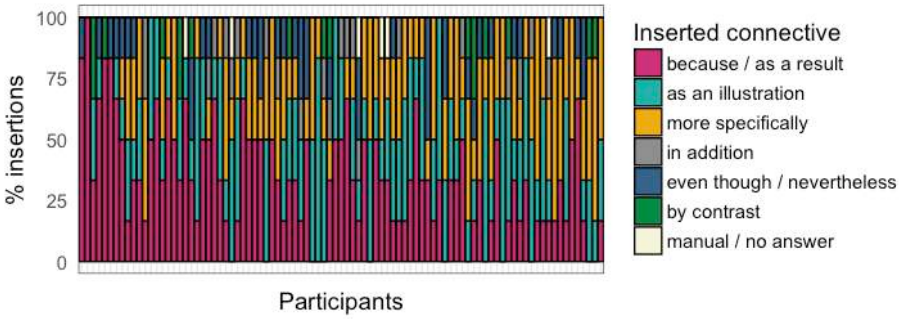


Figure 3a. Insertions in the fourth batch per participant. Participants are ordered according to their prior bias.

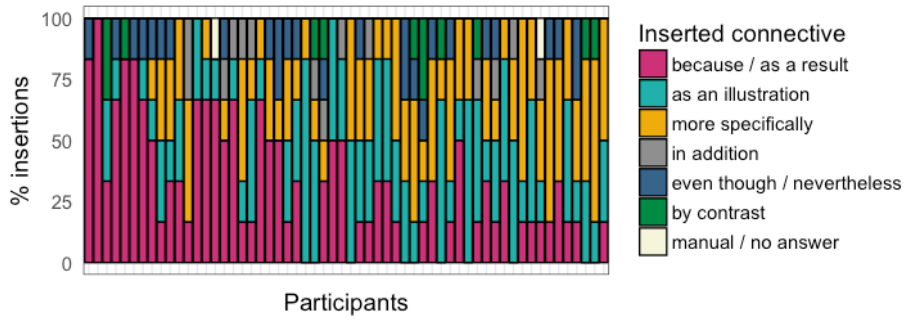
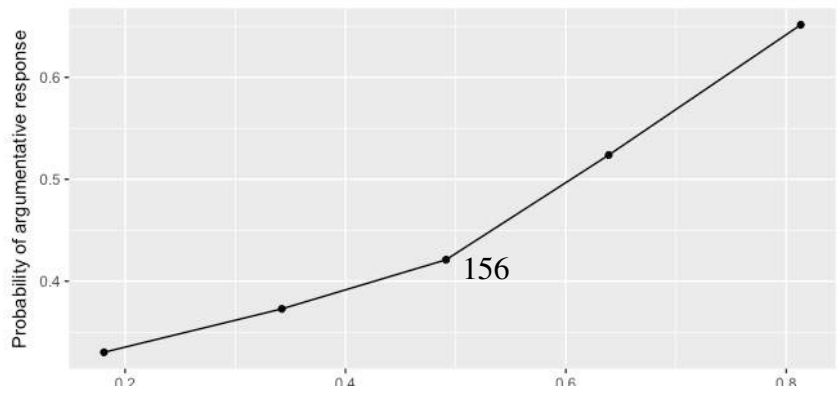


Figure 3b. Insertions in the fourth batch per participant for those who showed an argumentative or non-argumentative bias. Participants are ordered according to their prior bias.

In Figure 3a, there seem to be more participants that interpreted many items as argumentative on the left compared to the right, but this is not the case for all participants: some participants that displayed no bias in the previous three iterations (i.e., in the middle of this graph) inserted no or barely any causal connectives in this iteration. In Figure 3b, the tendency of the number of argumentative interpretations decreasing from left to right is more visible. The results displayed in these graphs indicate that (at least some) participants do in fact have different interpretation biases.

In order to test whether these differences in biases are statistically significant, we ran binomial logistic regression models. Our analysis is similar to exhaustive, leave-one-out cross-validation: we tested the significance of every possible combination of three batches as a set for estimating bias vs. one batch for observing whether that bias was stable (i.e., 1 2 3 vs. 4; 1 2 4 vs. 3; etc.). We first recoded the participants' responses into a binary variable representing the type of responses (‘argumentative’ versus ‘non-argumentative’). In order to be able to interpret the coefficient, we standardized the ratio. We then modeled the results using a binomial GLMER model in R. The results show that the participant’s prior bias is predictive of insertion ($\beta = 1.78$, SE = 0.17, $z = 10.6$, $p < .001$). This is visualized in the figure below, which shows the



probability of an argumentative response as a function of the prior bias.

Figure 3. Probability of argumentative response as function of the prior bias binned into five bins.

Discussion and future directions

Coherence relations can often be interpreted in different ways or convey multiple relation senses (see also Rohde et al., 2016; Sanders, 1997; Webber, 2013). Few studies have investigated how readers process multi-interpretable coherence relations: do readers have a common systematic bias to interpret such relations in a certain manner? Or do they show individual differences in their interpretation preferences? The results from the current study suggest that readers differ in how they interpret relations: some readers are more prone to interpret relations as argumentative, whereas others are more prone to interpret them as elaborative.

This bias could be a characteristic inherent to all readers. However, another possible explanation could be that the bias is caused by differences in processing; i.e., the depth of processing could affect readers' interpretations. Several studies suggest that comprehenders' reading processes are affected by shallow versus deep processing (e.g., Aaronson & Ferres, 1986; Noordman, Vonk & Kempf, 1992). The shallow/deep processing account is based on the hypothesis that when people read a text with a particular goal in mind, they process sentences more thoroughly and engage in inference processes. In order to investigate whether the shallow/deep processing account can provide an explanation for individual biases in coherence relation interpretation, we conducted a follow-up experiment with the same participants, inviting them to a new task where they will summarize the relation first, and then be asked to provide a connective. The results of this experiment are currently being processed.

The crowdsourcing method used in the current study allows us to tap into the interpretations of relations by naïve, untrained readers. However, it does have its limits. First, the method currently does not provide clear results regarding multiple interpretations of a single item. It's possible that participants inferred two readings for a particular

item but only inserted one connective. Our results are not conclusive regarding this issue, since only few participants inserted multiple connectives. However, we asked participants to choose the connective that "best expresses" the meaning of the relation, in which case the chosen connective should represent the strongest reading that was inferred. Second, the frequency of connectives is currently not controlled for in the design. The argumentative connective *because* is more frequent than *as an illustration* or *more specifically*. It is possible that this played a role in the participants' choice. In future experiments, we aim to test whether the frequency of connectives influences the results by including less frequent argumentative connectives.

Regarding the implications for the annotation of discourse relations, we take these results to indicate that manual annotation should be done by more than two annotators; an idea that has been proposed by Krippendorf (2004). Collecting a large number of annotations for single items allows researchers to obtain a distribution of relation senses. This distribution can give researchers more insight into the multiple (concurrent or alternative) readings of ambiguous relations, and into how dominant each sense is for a particular relation.

References

1. Asr, F. T., & Demberg, V. (2013). On the information conveyed by discourse markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics* (pp. 84–93).
2. Aaronson, D., & Ferres, S. (1986). Reading strategies for children and adults: A quantitative model. *Psychological Review*, *93*(1), 89.
3. Blakemore, D. (1997). Restatement and exemplification: A relevance theoretic reassessment of elaboration. *Pragmatics & Cognition*, *5*, 1–19.
4. Carston, R. (1993). Conjunction, explanation and relevance. *Lingua*, *90*, 27–48.
5. Das, D. and M. Taboada (2017) Signalling of coherence relations in discourse. *Discourse Processes,* 1-29.
6. Degand, L. (1998). On classifying connectives and coherence relations. In *Proceedings of the 1998 ACL Workshop on Discourse Relations and Discourse Markers* (pp. 29–35).
7. Halliday, M. A. (1994). *An introduction to functional grammar.* London: Edward Arnold.
8. Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, *3*, 67–90.
9. Knott, A., & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, *18*, 35–62.
10. Krippendorff, K. (2004). Reliability in content analysis. *Human communication research*, *30*(3), 411-433.

11. Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, *8*, 243–281.
12. Noordman, L. G., Vonk, W., & Kempff, H. J. (1992). Causal inferences during the reading of expository texts. *Journal of Memory and Language*, *31*(5), 573-590.
13. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Citeseer.
14. Rohde, H., Dickinson, A., Schneider, N., Clark, C. N., Louis, A., & Webber, B. (2016). Filling in the Blanks in Understanding Discourse Adverbials: Consistency, Conflict, and Context-Dependence in a Crowdsourced Elicitation Task. In *Proceedings of Linguistic Annotation Workshop X* (pp. 49-58).
15. Sanders, T. J.M., Spooren, W. P., & Noordman, L. G. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, *15*, 1–35.
16. Sanders, T. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse processes*, *24*(1), 119-147.
17. Scholman, M. C. J., & Demberg, V. (2017). Examples and specifications that prove a point: Identifying elaborative and argumentative discourse relations. *Dialogue & Discourse*, *8*(2), 56-83.
18. Webber, B. L. (2013). What excludes an Alternative in Coherence Relations?. In *Proceedings of the International Conference on Computational Semantics* (pp. 276-287).

The automatic analysis of subjectivity and causal coherence in text

Wilbert Spooren[1][0000-0002-2982-3970] and Ted Sanders[2]

[1] Centre for Language Studies, Radboud University Nijmegen
[2] UIL-OTS, Utrecht University
w.spooren@let.ru.nl

Understanding a text means making a coherent representation of the information in that text. Causal coherence place an important role in making that representation. Dutch has a rich repertoire of causal connectives to express such causal links, the so-called causal DRDs. Previous research has shown that causal DRDs have their own profile: Dutch omdat expresses mostly relatively objective relations, whereas want tends to express more subjective relations. The following examples demonstrate the point.

1. D De velden zijn nat omdat het veel geregend heeft deze week.
   E The fields are wet OMDAT it much rained has this week
   'The fields are wet because it has rained a lot last week.'
2. D De voetbalwedstrijden worden vast afgelast, want het heeft dezeweek erg veel geregend.
   E The soccer games become surely cancelled, WANT it has this week very much rained
   'Surely the soccer games will be cancelled, because it has rained a lot this week.'
3. D Jan kwam terug omdat hij van haar hield.
   E Jan came back OMDAT he from her loved.
   'Jan came back because he loved her.'
4. D Jan hield van haar, want hij kwam terug.
   E Jan loved from her, WANT he came back.
   'Jan loved her, because he came back.'

5. D Wat doe jij vanavond want er draait een goede film.
   E What do you tonight WANT there turns a good movie.
   'What are you doing tonight, because there's a good movie on.'

The differences between examples (1-5) have been described in terms of subjectivity [1]. Subjectivity can be defined as the degree to which the interpretation of an utterance requires that there is an active Subject of Consciousness who is responsible for the truth of the utterance. An utterance is subjective because there is some thinking entity in the discourse who evaluates. For example, the truth of an utterance such as *The height of the Eiffel Tower is 330 meters* can be evaluated directly in reality, and hence it is not subjective. By contrast, an utterance like *The Eiffel Tower is the greatest achievement of modern day architecture* requires the assumption that there is a Subject of Consciousness who is responsible for its truth.

Relations like (1), which Sweetser has termed content relations [2], can be described as objective: they report real world causality and do not assume the presence of a Subject of Consciousness. So-called epistemic relations like (2) and (4) are subjective because they present the outcome of an active reasoning process from the speaker or writer of the utterance. Similarly, speech act relations like (5) are subjective, because the Subject of Consciousness is motivating his or her performance of the speech act. Reason relations like (3) are in-between, because they do require the assumption of a Subject of Consciousness, but that is typically a character that is quoted in the text, whose reason for performing an action is reported.

Dutch has a preference to use *omdat* for more objective relations and *want* for more subjective relations, as in the examples above. The frequency with which *want* and *omdat* occur is very much genre-dependent: *want* is much more frequent in spontaneous conversations whereas *omdat* occurs more often in written newsreports and opinion pieces. At the same time, the subjectivity profile seems to be independent of genre: the difference in subjectivity between *want* and *omdat* is constant for each of the three genres that were investigated by Sanders and Spooren [1].

This type of findings is typically based on manual analyses of relatively small corpora. Such studies generally use a research design in which subsets of 100 instances of *omdat* and of *want* are compared in different genres (see for example, [3] for an analysis of forward causal DRDs in Dutch, and [4] for causality in Mandarin Chinese).

In this paper, we present a tool that makes use of state-of-the-art language technology to carry out such analyses automatically. The tool is the output of an ongoing project ACAD (Automatic Coherence Analysis of Dutch). For details on the project see https://www.clariah.nl/projecten/research-pilots/acad. The project aims at reaching three goals: (i) carry out these analyses automatically, thus preventing intercoder reliability issues; (ii) scale up the analyses by looking at many more instances and many more causal DRDs than is possible in manual analyses; (iii) look at many different genres.

The present study links to work done by Bestgen et al. [5], who used so-called thematic text analysis: the difference in subjectivity between, for example, *want* and *omdat* leads to the prediction that there are more subjective adjectives and adverbs in the segments that are connected by *want*, and more objective adjectives and adverbs in the segments connected by *omdat*. For our list of subjective and objective adjectives and adverbs we made use of the gold1000 list determined by De Smedt and Daelemans [6], who had participants rate the subjectivity of 1012 adjectives on a scale from 0 to 1. We identified those adjectives as subjective that had a score of 0.7 or higher for each of its meanings (650 adjectives, examples: *overweldigend* (‘overwhelming’), *afschuwelijk* (‘horrible’)), whereas objective adjectives had a score of 0.2 or lower (171 adjectives; examples: *visueel* (‘visual’), *zwart* (‘black’)).

The analysis goes through a number of steps: (i) identification of the relevant cases of causal DRDs; (ii) establishing the scope of the segments S_1 and S_2 that are connected by the DRDs; (iii) determining the direction of the causal link (backward, where the first segment expresses the consequent in the causal relation, as in examples (1-5), or forward); (iv) counting the number of subjective and objective adjectives and adverbs in the two segments; and (v) statistically testing the subjectivity hypothesis.

The current study made use of the corpora available in the Clariah environment: the SONAR corpus (a 500M words corpus containing 25 genres varying from newspaper texts, to wiki-pages, chat and texting, cf. [7]); the VU-DNC corpus (a 2M word corpus containing texts from

newspapers from the 1950s and from 2002; [8]); the Corpus of Spoken Dutch (CGN, [9]); and two newly added corpora: WhatsApp messages obtained in a recent study on the relationship between new media use by adolescents and young adults ([10]), and news texts from a Dutch quality newspaper published both on paper and online, matched for topics and genre (the NRC corpus; 1M words).

First results show indeed that the instrument is sensitive enough to detect the expected differences in the subjectivity of the environment of want and omdat. Thereotical implications and urgent next steps will be discussed. The discussion will be related to the the corpus build in DiscAn [11].

References

1. Sanders, T., Spooren, W.: Causality and subjectivity in discourse: The meaning and use of causal connectives in spontaneous conversation, chat inter- actions and written text. Linguistics 53(1):53–92 (2015).
2. Sweetser, E.: From etymology to pragmatics. Cambridge, Cambridge University Press (1990).
3. Sanders, T.J.M., Stukker, N.:. Causal connectives in discourse: A cross-linguistic perspective. Journal of Pragmatics 44(2), 131-137 (2012).
4. Li, F., Sanders, T.J.M., Evers-Vermeul, J.: On the subjectivity of Mandarin reason connectives - Robust profiles or genre-sensitivity?. In Stukker, N., Spooren, W., Steen, G. (eds.) GENRE IN LANGUAGE, DISCOURSE AND COGNITION, pp. 15-49. De Gruyter Mouton, Berlin/New York (2016).
5. Bestgen, Y., Degand, L., Spooren, W.: Toward automatic determination of the semantics of connectives in large newspaper corpora. Discourse Processes 41(2), 175–193 (2006).
6. Smedt, T. D., Daelemans, W.: 'Vreselijk mooi!'(terribly beautiful): A subjectivity lexicon for Dutch adjectives. In PROCEEDINGS OF THE EIGHTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2012), pp. 3568-3572 (2012).
7. Oostdijk, N., Reynaert, M., Hoste, V., Schuurman, I.: The construction of a 500-million-word reference corpus of contemporary written Dutch. In ESSENTIAL SPEECH AND LANGUAGE TECHNOLOGY FOR DUTCH, pp. 219–247. Springer (2013).
8. Vis, K. (2011). Documentation of the VU Diachronic Newspaper texts Corpus. Unpublished ms., retrieved on Feb. 10, 2018 from http://tst-centrale.org/images/stories/producten/documentatie/vu-dnc_doc3-documentation-of-corpus.pdf
9. Oostdijk, N.: The Spoken Dutch Corpus Project. The ELRA Newsletter 5(2), 4–8 (2000).
10. Verheijen, L., Spooren, W., Kemenade, A. van, (submitted). The relationship between Dutch youths' social media use and school writing.
11. Sanders, T., Vis, K., Broeder, D.: Project notes of Clarin project DISCAN: Towards a discourse annotation system for Dutch language corpora. In EIGHTH JOINT ACL - ISO

WORKSHOP ON INTEROPERABLE SEMANTIC ANNOTATION, pp. 61–65. Pisa, Italy (2012).

Annotating the Meaning of Discourse Marker *so* in TED Talks

Valūnaitė Oleškevičienė, Giedrė[1], Burkšaitienė, Nijolė[2], Rackevičienė, Sigita[3] and Mockienė, Liudmila[4]

[1] Mykolas Romeris University, Ateities st. 20, LT-08303 Vilnius, Lithuania
gvalunaite@mruni.eu
[2] Mykolas Romeris University, Ateities st. 20, LT-08303 Vilnius, Lithuania
n.burksaitiene@mruni.eu
[3] Mykolas Romeris University, Ateities st. 20, LT-08303 Vilnius, Lithuania
sigita.rackeviciene@mruni.eu
[4] Mykolas Romeris University, Ateities st. 20, LT-08303 Vilnius, Lithuania
liudmila@mruni.eu

Abstract. The purpose of this paper is to report on the results of cross-linguistic annotation of the English spoken discourse marker *so* and its counterparts in Lithuanian using the multilingual open translation project TED Talks as a resource of data. The purpose is achieved by investigating the domains and functions of the discourse marker *so* in English and Lithuanian as well as analysing its translations into Lithuanian.

The present study was conducted in two stages. First, the meanings of the discourse marker *so* were annotated and compared to the meanings of their counterparts in Lithuanian. Later, translations of the discourse marker *so* into Lithuanian were analysed. In the present investigation, the taxonomy of functional annotation for spoken discourse (Crible and Degand 2017) and manual translation spotting were employed.

The findings showed that the discourse marker *so* and its counterparts in Lithuanian in most cases express rhetorical consequence, followed by rhetorical structuring. It was also established that the most frequent variants of translation of the discourse marker *so* were those provided by bilingual English-Lithuanian dictionaries, whereas the least frequent translations were expressed by a particle or a verb. On the other hand, translation by omission was also frequently used.

Keywords: Discourse Marker, Cross-linguistic Discourse Annotation, Consequence, Pragmatics, Translation

1 Introduction

Technological advancement enables linguists to apply linguistic corpora annotation for language analysis while enhancing applied language use for diverse purposes. Computer-mediated language use reveals pragmatic text relations. Discourse markers working on the pragmatic level ensure text coherence and clear relations between sentences. The problems related to discourse markers become a particular challenge for translators who have to adapt them to a new language and culture, in which textual

strategies involving their use are often different from those of the source text (Zufferey and Degand 2017). Hence, analysis of discourse markers plays a relevant role in the field of cross-cultural communication and translation.

The present research focuses on the functions of the discourse marker *so* by comparing the use of the discourse marker *so* in annotated TED Talks in English and Lithuanian. The investigation was conducted in two stages, including annotating the domain and functions of the discourse marker *so* in English and its counterparts in Lithuanian, followed by the analysis of the translations of this discourse marker into Lithuanian. In this paper, first, the concept of discourse markers will be defined, which is followed by the description of the state-of-the-art methods for the annotation of discourse markers. The research methodology and results will be then reported.

The limitation of this investigation is that it focuses on the annotation of only one discourse marker of spoken English and its Lithuanian counterparts using TED Talks, which calls for the analysis of other spoken discourse markers. The annotation of the domains and functions of the English discourse marker *so* and its counterparts in Lithuanian as well as the analysis of translation variants enables translators to choose the translation equivalent which is the closest to the source language.

The present study contributes to the field of research of discourse annotation of spoken data conducted cross-linguistically using multilingual corpora. While English spoken discourse markers and their counterparts in other languages have been widely investigated, little known research has focused on cross-linguistic discourse annotation involving their counterparts in Lithuanian.

2 Theoretical Background

Discourse markers have been analysed by a number of researchers using different approaches, which has resulted in different definitions. In the present investigation, the discourse marker definition provided by Crible (2014) is used. According to the author, *discourse markers* are "grammatically heterogeneous, multifunctional type of pragmatic markers" which signal "a discourse relation between the host unit and its context <…>, expliciting the structural sequencing of discourse segments, expressing the speaker's meta-comment on his phrasing, or contributing to interpersonal collaboration" (Crible 2014:3-4).

Annotating the meaning of discourse markers is one of the major tasks of discourse analysis as it discloses the principles of coherence of spoken and written discourse, facilitates the process of collecting linguistic data that are important to the specialists of language acquisition and translators. Research in the field of discourse marker annotation has been extensive and was conducted using different methods, each of which has its advantages and limitations. The state-of-the-art methods used for the annotation of discourse markers include the classical sense annotation, translation spotting and functional annotation (Cartoni et al. 2013; Crible 2014; Crible and Degand 2017). The present study employs Crible and Degand's (2017) taxonomy of domains and functions of discourse markers, which is specifically designed for annotating discourse markers used in spoken discourse and consists of four main domains.

The first one is the ideational domain. It is linked to "states of affairs in the world, semantic relations between real events". The second domain is rhetorical and is linked to "the speaker's meta-discursive work on the ongoing speech". The sequential domain is related to "the structuring of discourse segments, both at macro- and micro-level, whereas the interpersonal domain refers to "the interactive management of the exchange, in other words to the speaker-hearer relationship" (Crible 2014:18)

In the original version of this taxonomy (Crible 2014), certain functions were ascribed to each domain, e.g. the ideational domain covered the functions of cause, consequence, concession, contrast, alternative, condition, temporal, and exception. Besides, the domains and functions were "inter-dependent", e.g. a "cause" always belonged to the ideational domain. In the revised version of this taxonomy, the inter-dependency between domains and functions is not used, thus, any domain can apply to any function and any function can apply to any domain. The second major modification was the reduction of the number of function-labels, which was achieved by merging similar pairs of discourse marker functions. According to Crible and Degand (2017), using the revised taxonomy, annotators "can choose to start at domain-level or function-level, to annotate both levels simultaneously or independently, and could even decide to stop at one level if a particular domain DM token is under-specified for the other level" (2017:20). Also, the authors believe that this system can substantially improve inter-annotator agreement, which was supported by the results of the annotation experiment.

3 Research Methodology

The methodological choices made in this study are related to the choice of the corpus and the annotation method. These choices are determined by the aim of the research. That is, to annotate the English spoken discourse marker *so*, to compare its meanings with the counterparts in Lithuanian as well as to analyse the translations of *so* into Lithuanian, the multilingual TED Talks were chosen. This choice was predetermined by the fact that parallel texts are considered to be ideal for optimal comparability between languages as they provide more flexible and accurate ways to compare discourse markers (Zufferey and Degand 2017).

The choice of the functional approach to be used for this investigation was due to the specific nature of discourse markers, which covers some specific features, e.g. even though most languages possess discourse markers, they have a high degree of contextual variation (Crible and Degand 2017). Moreover, discourse markers are often multifunctional, i.e. they can convey several discourse relations (Cartoni et al. 2013). Therefore, to annotate the domains and functions of the English discourse marker *so* and compare them with Lithuanian counterparts, the revised taxonomy for spoken discourse relational devices (Crible and Degand 2017) was employed.

The empirical research consisted of two stages. Initially, the discourse marker *so* was compared to its Lithuanian counterparts by applying Crible and Degand's taxonomy of domains and functions of discourse markers. Then, the translations of *so* into Lithuanian, identified in the annotated sample, were analysed.

4 Research Findings

The functional taxonomy (Crible and Degand 2017) which was used for the annotation in the present study describes discourse markers as functioning in four domains, i.e. in the ideational domain (related to real-world events), the rhetorical domain (related to the speaker's expressed subjectivity and meta-discursive effects), the sequential domain (concerns the structuring of local and global units of discourse), and the interpersonal domain (related to managing the speaker-hearer relationship). The four domains correspond to the overall discourse intentions or entities, which depend on what the speaker is targeting: content (the ideational domain), illocutionary value (the rhetorical domain), discourse structure (the sequential domain) or intersubjective inferences (the interpersonal domain) (Crible 2017).

The results of the present study illustrate that in most cases the discourse marker *so* and its Lithuanian counterparts function in the rhetorical domain and express rhetorical consequence and rhetorical specification, followed by the sequential domain and the ideational domain (Figure 1).

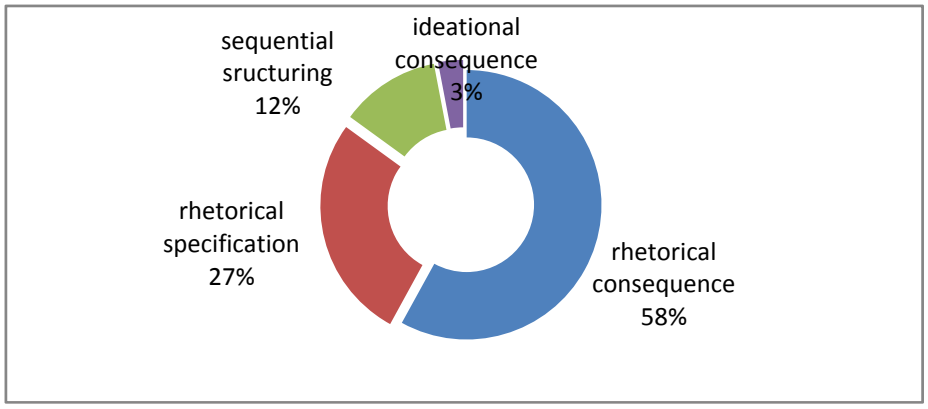


Figure 1. The annotated values of the discourse marker *so*

To be more exact, in the annotated sample, 58% of the occurrences in both languages convey rhetorical consequence and 27% of occurrences express rhetorical specification. These findings mean that the occurrences covey the speaker's subjective perception and produce the effect of subjective discourse management, which can be illustrated by the example of rhetorical consequence, see (1):

(1) *[So] I would have thought that perhaps the most successful relationships were ones where there was a really high negativity threshold.*

(1) *[Taigi] būčiau pagalvojusi, kad patys sėkmingiausi santykiai yra tie, kur negatyvumo slenkstis yra labai aukštas.*

It was also established that 12% of the occurrences in the sample were used in the sequential domain, where the discourse marker *so* and its Lithuanian counterparts perform the function of sequential structuring, see (2):

(2) *[So] those are my top three tips of how maths can help you with love and relationships.*
(2) *[Taigi], štai ir mano top trys patarimai kaip matematika gali jums padėti meilėje ir santykiuose.*

On the other hand, the findings revealed that the cases of *so* expressing the meaning of ideational consequence are scarce (3%).

An interesting finding of this investigation was related to the counterpart of the discourse marker *so* in Lithuanian. It was established that in 65% of cases, *so* was used in rhetorical domain in the source language, however, it did not have counterparts in Lithuanian. This might be explained by the translation requirements of TED Talks, i.e. they are, in fact, subtitles. Thus, translators use special techniques to reduce the text to meet specific requirements regarding the number of characters allowed to be on subtitles (http://translations.ted.org/wiki/How_to_Compress_Subtitles).

The results of the second stage of the research showed several translations for the discourse marker *so* (Figure 2).

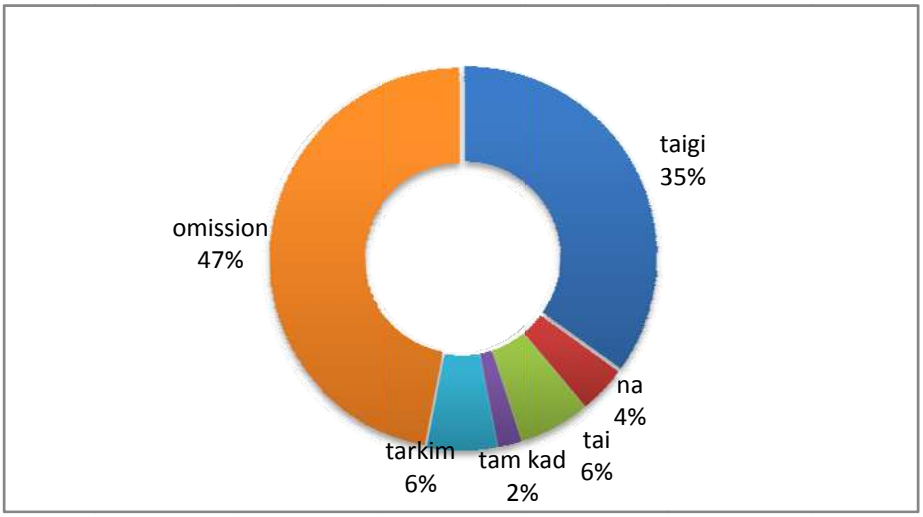


Figure 2. Translation values of the discourse marker *so*

It was established that the most frequent variants of translating *so* into Lithuanian were *taigi* (35%) and *tam kad* (2%), which are the variants provided by bilingual English-Lithuanian dictionaries. In the rest cases, *so* was translated by particles *tai* (6%) and *na* (4%) and the verb *tarkim* (6%). The latter cases occurred when the trans-

lator chose to use the translation strategy of transposition, i.e. grammatical forms in the source and target languages differ, see (3):

(3) [So] if you take someone like Portia de Rossi, for example, everybody agrees that Portia de Rossi is a very beautiful woman.
(3) [Tarkim], pasiimkime Portia de Rossi kaip pavyzdį, visi sutiks, kad Portia de Rossi yra graži moteris.

Example (3) demonstrates that the discourse marker *so* was rendered by a parenthetical verb which performs the function of rhetorical specification in Lithuanian translation.

The investigation also disclosed that the translator used the translation strategy of omission in 47% of cases in which *so* in the source language was used in rhetorical domain or for sequential structuring, which is characteristic of spoken discourse, see (4):

(4) [So] let's imagine then, that you start dating when you're 15 and ideally, you'd like to be married by the time that you're 35.
(4) Įsivaizduokite, kad pradedate susitikinėti kai jums 15, ir idealiu atveju norėtumėt susituokti kai jums 35.

These findings lead to the assumption that in such cases the translator's choice was predetermined by the requirements of translating subtitles, synchronizing them and making them concise.

5 Conclusions

The present investigation revealed that in the annotated sample the discourse marker *so* and its Lithuanian counterparts convey consequential meaning and mainly function in the rhetorical domain of discourse management, which is associated with expressing the speaker's subjectivity. It was also established that the discourse marker *so* was used for structuring discourse, which was supported by a number of occurrences of sequential structuring which was used for opening, resuming or closing the topic. On the other hand, the results also showed that the occurrences of the discourse marker *so* expressing ideational consequence are scarce.

The results regarding the translations of the discourse marker *so* into Lithuanian illustrated that the most frequent translation variant of *so* was *taigi*, which is the translation variant provided by bilingual English-Lithuanian dictionaries. On the other hand, the translator chose to translate *so* by particles and by a verb, which was a suitable choice for conveying its meaning into Lithuanian. Finally, a big number of omissions was observed, which may be due to the specificity of TED Talks translations.

The present investigation has some limitations. First, it focuses on spoken-like texts of TED Talks, the annotation of which in Lithuanian has started only recently.

Second, the annotated corpus of TED Talks in English and Lithuanian used for this research was not big. In the future, more texts will be annotated. Also, the present study focuses on the annotation of only one discourse marker of spoken English and its Lithuanian counterparts using TED Talks, which calls for the analysis of other spoken discourse markers in both languages.

The comparative research of discourse markers provides specific information and knowledge both for language learners and translators. Knowledge of pragmatic functions and semantic meanings provides easily identifiable advice on how discourse markers could be used and translated.

The object of the research is comparatively new in Lithuania and adds to the research field related to discourse relations studies. The present investigation has been conducted within the framework of TextLink COST action IS1312.

References

1. Cartoni, B., Zufferey, S., Meyer, T.: Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation Spotting Technique. In: Dialogue and Discourse 4(2), 65–86 (2013).
2. Crible, L.: Identifying and describing discourse markers in spoken corpora. Université Catholique de Louvain, Louvain-la-Neuve (2014).
3. Crible, L.: Discourse markers, (dis)fluency and the non-linear structure of speech: a contrastive usage-based study in English and French. Université Catholique de Louvain, Louvain-la-Neuve (2017).
4. Crible, L., Degand, L.: Reliability vs. Granularity in discourse annotation: What is the trade-off? In: Corpus Linguistics and Linguistic Theory 14(2), 1–29 (2017).
5. How to Compress Subtitles? Available at http://translations.ted.org/wiki/How_to_Compress_Subtitles [accessed 3 December, 2017].
6. Zufferey, S., Degand, L.: Annotating the meaning of discourse connectives in multilingual corpora. In: Corpus Linguistics and Linguistic Theory 13(2), 1–24 (2017).

A FrameNet lexicon and annotated corpus as DRD resource: Causality in the Asfalda French FrameNet

Laure Vieu

IRIT, CNRS & Université de Toulouse

1 Introduction

A FrameNet for French has been developed within the Asfalda project [5]. This new freely available resource[1] consists of a set of frames updated with respect to the original FrameNet for English [4] with new, merged or semantically redefined frames, a lexicon and an annotated corpus of written text. The project did not aim at full coverage, so the resource has been developed using a domain-by-domain methodology around 4 notional domains [8]. The causality domain includes 11 frames associated to 332 French lexical units (simple or complex, with POS) giving rise to 3,895 annotated occurrences[2] in a corpus of French treebanks of 624,187 tokens [13]. I argue here that this resource, while not designed for this purpose, is of interest as a Discourse Relational Device (DRD) resource, at least for this causality domain.

Freely available French corpora of written texts annotated with discourse relations are few. The first-ever resource built is Annodis [10], on which the Explicadis resource dedicated to causality has been built [2, 3]. LexConn, an inventory of 328 discourse connectives [11], is another important DRD resource for French. LexConn serves as a basis for annotating the French Discourse Treebank (FDTB) [6], the only other such corpus I am aware of, in which annotation is still in progress. With such few resources, any addition is worth considering, especially with a corpus already POS-tagged and parsed, something Annodis lacks.

2 Frames and discourse relations, and their associated lexicons

Frames in FrameNet and Asfalda are descriptions of prototypical situations, semantically characterized by their participants (called frame elements) and how these are related. The set of frames is structured by frame-to-frame relations such as inheritance. Frames are associated with triggering lexical units (called frame-evoking elements in FrameNet), and the annotation of such lexical units with a frame requires the annotation of its frame elements occurring in the sentence.

Of course, frames are not discourse relations (DRs), and triggering lexical units cover all sorts of parts of speech. Nevertheless, 6 out of 11 frames of the causality domain in Asfalda (Causation, Evidence, FR-Reason[3], FR-Cause_Enunciation, Explaining_the_facts, FR-Contingency-Objective_Influence) are semantically close to DRs and associated to a significant number of DRDs or discourse markers. These 6 frames are also used to describe the semantics of propositional contents and associated to nouns, verbs and adjectives as triggers. Still, lexical units that operate as DRDs can be simply selected through their POS in the lexicon: adverbials, prepositions or conjunctions. A few additional expressions used as DRDs, such as "*suite à*" (*due to*), "*résultant de*" (*resulting from*) or "*résultat*" (*as a result*) were not tagged as prepositions or

[1] `https://sites.google.com/site/anrasfalda/`

[2] Not all occurrences of the 332 lexical units have been annotated. The annotation of frequent lexical units is limited to 100 occurrences.

[3] Frames whose name starts with "FR-" are new or significantly modified frames for Asfalda.

adverbials in the treebank, but the corresponding annotations of nouns and verbs can be manually selected without much effort. This extraction process yields a sub-resource of 6 frames and 81 lexical units with 1,215 annotated occurrences. This is significant, as the Explicadis resource contains 8 DRs with 319 annotated occurrences in which 53 lexical clues (not all of them being discourse markers) appear, and the causal part of LexConn contains 98 lexical units associated with 5 causal DRs.

Two main features of Asfalda (and FrameNet) certainly are weaknesses and would require further annotation efforts to make the extracted part of Asfalda a full DRD resource. First, frames are annotated only through the occurrence of a triggering lexical unit, while it is well-known that many DRs are unmarked in texts, a phenomenon estimated in Explicadis at around 39%. Second, the annotation of "frame elements" or thematic roles, among which we find the two arguments of the DR corresponding to the frame, is done within the sentence in which the trigger appears only. Since DRDs may relate discourse units appearing in different sentences, in many cases, one of the two discourse units is not annotated.[4]

In Asfalda (and FrameNet), there is no distinction in the (rhetorical) order of presentation. This means that each frame, e.g., Causation, is used to annotate the occurrences of discourse markers that would be annotated with two different relations, e.g., Explanation and Result in Explicadis. However, this rhetorical distinction can be directly computed from the corpus annotations, since no causal discourse marker is ambiguous in this respect.

Section 3 will address the semantic specificities of the set of causal frames in Asfalda. Without entering in those details yet, a semantic correspondance can be established between the 6 Asfalda frames selected and Explicadis DRs. Explicadis's 8 DRs are SDRT's Explanation and Result [1] plus 6 additional DRs: epistemic Explanation_ep and Result_ep, inferential Explanation_inf and Result_inf, and pragmatic (or speech-act) Explanation_prag and Result_prag.[5] The annotation of discourse markers with the frame Causation corresponds to either an Explanation or a Result; the frame Evidence corresponds to Explanation_ep or Result_ep; and the frame FR-Cause_Enunciation to Explanation_prag or Result_prag. The frames FR-Reason, FR-Cause_Enunciation, Explaining_the_facts, FR-Contingency-Objective_Influence have no exact counterpart in Explicadis (nor in SDRT and LexConn), but their occurrences on discourse markers may be considered as cases of Explanation or Result. On the other hand, Explicadis's distinction of Explanation_inf and Result_inf has not been adopted in Asfalda; such cases would be annotated with the frame Evidence, reflecting the fact that inferential DRs are sub-relations of epistemic ones. LexConn uses SDRT's Explanation, Result, Explanation* and Result*, plus an Evidence relation. Explanation* and Result* can be considered as corresponding to the frames Evidence or FR-Cause_Enunciation, and the DR Evidence to the frame Evidence.

There is a large overlap between the three lexicons; overall, they contain 146 lexical units. Asfalda contains 15 new (with respect to Explicadis and LexConn) lexical units with 115 occurrences, e.g., "*sous l'effet de*" (*as a result of*, *under the influence of*) and "*au vu de*" (*given*), and 4 more without occurrences. Some differences are accounted by the facts that Explicadis contains some lexical clues that are not discourse markers, and only those appearing in its corpus, and that LexConn considers also lexico-syntactic patterns such as "à + Vinf". Table 1 shows the merged lexicon, with its distribution and its association with frames or DRs in the three resources. It reveals the polysemy of these lexical units and the variable scope of the frames and DRs in the three resources. Bold is used to signal a lexical unit not already present in Explicadis

[4] When a frame element is not filled within the sentence, a typology of "null instantiations" was used to flag the frame element, especially "Definite null instantiation" if the frame element is expressed elsewhere in the text. Identifying such elements is required to check the semantics while annotating.

[5] These 6 new DRs have been introduced and characterized in Explicadis to clarify the confusing uses of SDRT's Explanation* and Result* in Annodis [2, 3].

and LexConn, or a new meaning on the basis of the correspondances described above. Frame-X means the marker is associated to that Frame in Asfalda albeit with no annotated occurrence (Frame-? when annotation for that lexical unit has not been done yet). O stands for the rhetorical order of the lexical unit: E for Explanation-like, R for Result-like.

Lexical unit	O	Frames – Asfalda	Drs - Explicadis	DRs - Lexconn
à + Vinf	E			Explanation
à cause de	E	Causation	Explanation	
à ce point que	R	Causation		Result
à ce rythme	R		Result_inf	
à défaut de	E	Causation-X		Explanation
à en + Vinf	R			Result
à force de	E	Causation		Explanation
à force	R	Causation-X		Result
à l'origine de	R	**Causation**		
à la suite de quoi	R	**Causation-X**		
à la suite de	E	Causation	Explanation	
à présent que	E			Explanation*
à preuve	E	Evidence		Explanation*
à tel point que	R	Causation		Result
à telle enseigne que	R			Result
ainsi	R	Causation, Evidence-X	Result, Result_ep, Result_inf	Result
alors	R	Causation, Evidence	Result	Result, Result*
après tout	E			Explanation*
attendu que	E			Explanation
au motif que	E	**FR_Reason**		
au point de	R	Causation		Result
au point que	R	Causation	Result	Result
au prix de	R		Result_ep	
au vu de	E	**Evidence, FR_Reason**		
aussi	R	Causation-X, Evidence	Result_ep	Result
aussitôt	R			Result
aussitôt que	E			Explanation
autant dire que	R			Result
autrement dit	R			Result*
avec	E		Explanation	
avec pour conséquence	R	Causation (conséquence.n)	Result	
bref	R			Result*
c'est à dire que	E			Explanation*
c'est dire que	R	**Evidence-X**		
c'est pourquoi	R	Causation, Evidence	Result, Result_ep	Result
c'est que	E	Evidence-X, Explaining_TF	cf. si … c'est que	
car	E	Causation, Evidence, FR_Cause_en	Explanation, Explanation_ep, Explanation_inf	Explanation*
cette fois que	E			Explanation*
comme quoi	R			Result*
comme	E	Causation, Evidence	Explanation	Explanation, Explanation*
conclusion (adv)	R	Causation-X		
conduisant à	R	Causation, FR_Reason (conduire.v)	Result	
conséquemment	R	Causation-X		Result
conséquence de	E	Causation (conséquence.n)	Explanation	
conséquence (adv)	R	**Causation**		
considérant que	E			Explanation*
considéré que	E			Explanation
d'abord (...ensuite)	E			Explanation
d'après	E	**Evidence**		
d'autant moins que	E	**Causation-X, Evidence-X**		
d'autant plus que	E	Causation-?, **Evidence-?**	Explanation	Explanation
d'autant que	E	Causation-?, Evidence-?	Explanation_ep	Explanation
d'où que	R			Result
d'où	R	Causation, Evidence	Result_inf	Result
d'un côté (...d'un autre côté)	E			Explanation
d'une part (...d'autre part)	E			Explanation
d'ailleurs	E			Evidence
dans la mesure où	E	**Causation**, Evidence	Explanation_ep	Explanation*
dans le coup	R			Result
dans le sens où	E			Explanation*
dans le sens que	E			Explanation*
de	E		Explanation	
de ce fait	R	Causation, **Evidence**		Result
de façon que	R	Causation-X		Result
de fait	E	Evidence		Explanation*, Evidence
de sorte que	R	Causation	Result, Result_inf	Result
de telle façon que	R			Result
de telle manière que	R	Causation-X		Result
décidément	R			Result*
déjà	E			Explanation
depuis	R			Result
dès lors que	E	Evidence		Explanation*
dès lors	R	Causation, Evidence, FR_Cause_en	Result	Result*
dès que	E		Explanation	Explanation
des suites de	E	Causation (suite.n)	Explanation	

Lexical unit	O	Frames - Asfalda	Drs - Explicadis	Drs - Lexconn
devant	E		Explanation	
donc	R	Causation, Evidence	Result, Result_ep, Result_inf	Result, Result*
du coup	R	Causation, **Evidence**		Result
du fait de	E	Causation, **Evidence**	Explanation	
du fait que	E	Causation, Evidence-X		Explanation, Explanation*
du reste	E			Evidence
effectivement	E	Evidence		Evidence
en + V-ant	E			Explanation
en ce sens que	E			Explanation*
en conséquence	R	Causation, **Evidence**		Result
en effet	E	Causation, Evidence, FR_Cause_en	Explanation, Explanation_ep, Explanation_inf	Explanation*
en raison de	E	Causation, **Evidence**	Explanation	
en témoignage de	E		Explanation	
en vertu de	E	**FR_Reason**		
et	R		Result, Result_ep	
et pour cause	E	**Causation**		
étant donné que	E	Evidence	Explanation_ep	Explanation
étant donné	E	Evidence	Explanation_ep, Explanation_inf	
faute de	E	Causation	Explanation	Explanation
grâce à	E	Causation	Explanation	
instantanément	R			Result
jusqu'à ce que	R	Causation	Result	Result
jusqu'à	R	Causation-X		Result
la preuve	E	Evidence-X (preuve.n)		Evidence
le fait est que	E			Explanation*
le temps de	E		Explanation	
lorsque	E			Explanation
maintenant que	E			Explanation*
même que	E			Evidence
par	E		Explanation	
par conséquent	R	Causation		Result
par contrecoup	R	**Causation**		
par exemple	E			Explanation*
par la faute de	E	**Causation**		
par le fait que	E			Explanation
par suite	R	Causation		Result
par voie de conséquence	R	**Causation**		
parce que	E	Causation, Evidence	Explanation, Explanation_ep, Explanation_inf	Explanation, Explanation*
pendant que	E			Explanation*
pour	E		Explanation	
pour cause de	E	**Causation**		
pour commencer	E			Explanation
pour conclure	R			Result*
pour (une/des...) raison(s) de	E	Causation, Evidence, FR_Reason (raison.n)	Explanation, Explanation_ep	
pour le coup	R			Result*
pour preuve	E	Evidence		Evidence
pour résumer	R			Result*
pour	R	Causation-X	Result	
pourquoi	E	**Causation, Evidence**		
premièrement	E			Explanation
preuve que	R	Evidence	Result_ep	Result*
puisque	E	Evidence, FR_Reason	Explanation, Explanation_ep	Explanation, Explanation*
résultat (adv)	R	Causation (résultat.n)	Result	Result
résultant de	E	**Causation** (résulter.v)		
sachant que	E			Explanation*
selon	E	**Evidence**		
si ... c'est que	E	cf. c'est que	Explanation	
si bien que	R	Causation	Result	Result
sinon	R	**Evidence-X**		
sitôt que	E			Explanation
sous l'effet de	E	**Causation, FR_Cont-Obj_inf**		
subséquemment	R	Causation-X		Result
suite à	E	Causation (suite.n)	Explanation	
surtout que	E			Explanation
tant et si bien que	R	Causation		Result
tant que	R		Result	
tel ... que	R		Result	
total (adv)	R			Result
tout d'abord	E			Explanation
vu	E	Evidence, **FR_Reason-X**	Explanation_ep	
vu que	E	Evidence, FR_Reason		Explanation, Explanation*

Table 1. Causal discourse markers (or clues) in Asfalda, Explicadis and LexConn, and their associated frames or DRs.

3 Causal frames in Asfalda and their interest for discourse annotation

Beyond Asfalda's decent size, the specific subset of 6 causal frames in Asfalda makes it an interesting DRD resource despite its annotation limitations.

The well-known content-level (or semantic or subjective) / epistemic-level distinction in the uses of causal discourse markers [9, 12, 7], is not present in Annodis nor in LexConn (except marginally with 6 lexical items associated with the Evidence relation), but has been introduced in Explicadis. It is also present in Berkeley's FrameNet through the distinction between the Causation and the Evidence frames, a distinction that has been much clarified in the Asfalda project as reflected in the annotation guide.[6] In addition, FrameNet distinguishes the frame Reason for triggers that are specific to content-level causation links in which the effect is an action or a mental attitude, a frame considered in Asfalda as semantically subsumed by Causation, and for this and other modifications morphed into FR-Reason. FR-Reason is closely related to the "volitional causal" relations of Degand and Pander Maat [7], so one step higher than Causation in their subjectivity scale, while this notion is completely absent from Explicadis and LexConn. Here is an excerpt of the frame FR-Reason in Asfalda, bold signalling the frame elements:

> **FR-Reason**
>
> *Definition*: A volitional **Agent** is responding to some situation **State_of_Affairs** by performing some **Action** (or holding some mental attitude). Alternatively, an **Actor**, a participant of some implicit **State_of_Affairs** stands in for the **State_of_Affairs**, in other words, an **Actor** volitionally or not pushes an **Agent** to perform some **Action** (or hold some mental attitude).
>
> *Distinctions with other frames*:
> $\neq$ Causation: In Causation the effect can be any sort of situation, not only actions and mental attitudes as in FR-Reason. Note though that FR-Reason is evoked only by those lexical units that have at least one subcategorization in which the **Agent** is subcategorized. Compare:
> *La crise de 1929 a amené la guerre* (*The crisis of 1929 brought the war*): Causation
> *La situation a amené le gouvernement à réagir* (*The situation has prompted the government to react*): FR-Reason
> $\neq$ Evidence: The main difference is that although volition or cognition is involved in the **Action**, FR-Reason is still a frame for factual objective causation, while Evidence is for epistemic causation or argumentation in which the "state-of-affairs" (cause) is presented as a support for a proposition (effect), which is a less established fact argued to be true. For a thorough examination of this distinction, which can be tricky, see the Evidence/Causation disambiguation guide.
>
> *Core Frame Elements*:
> **Action** The action that the **Agent** performs in response to a **State_of_Affairs**. This can also be a mental attitude held by the **Agent**.
> **Actor** An entity (not a situation, but not necessarily a sentient) which participates in an implicit **State_of_Affairs** (e.g., the **Actor**'s existence, presence, behaviour or action), perhaps volitionally and perhaps not.
> **Agent** (Semantic Type: Sentient) The person who responds to a **State_of_Affairs** by performing some **Action**.

[6] `http://asfalda.linguist.univ-paris-diderot.fr/documentation/asfalda_guide_desamb_Causation_Evidence.pdf`

State_of_Affairs The eventuality that motivates the **Agent**'s performing a particular **Action** in response to it.

Having introduced the new frame FR-Cause_Enunciation, Asfalda also includes a "pragmatic" or "meta-talk" causal frame in which the effect is a speech act, just as Explicadis does (and to some extent LexConn, although mixed up with epistemic causal relations). There are 19 occurrences of this frame, while only 3 Explanation_prag and Result_prag (unmarked) occurrences in Explicadis.

These 4 frames —Causation, Evidence, FR-Reason and FR-Cause_Enunciation— are the major ones able to encode causal DRs in Asfalda. The other 2, Explaining_the_facts and FR-Contingency-Objective_Influence, are only very marginally relevant to discourse; they contribute only with 2 lexical units and 3 occurrences.

The distribution of lexical units on this set of 6 frames in the corpus confirms earlier work on the famous French causal markers *parce que*, *car* and *puisque* (*because*, *since*). In particular, the distinction between Causation and FR-Reason allows to correctly account for the semantics of *puisque* which triggers only Evidence and FR-Reason in Asfalda and, crucially, not Causation. *Puisque* has been repeatedly shown not to be a simple content-level causal marker [9, 7, 14]; nevertheless, the lack of relation dedicated to volitional causation implied that examples (1) and (2) were considered occurrences of Explanation in LexConn and Explicadis respectively. (3) shows the single occurrence of *puisque* annotated with FR-Reason in Asfalda, as most are occurrences annotated with Evidence.

(1) **Puisqu**'il est mort je veux mourir (Since he is dead I want to die)

(2) Aujourdhui, les paléontologues donnent à Homo sapiens un âge d'environ 200 000 ans **puisque** les plus vieux ossements retrouvés sont deux crânes datés de -195 000 ans
(Today, paleontologists give Homo sapiens an age of about 200,000 years since the oldest bones found are two skulls dated to 195,000 years ago)

(3) Les syndicats s'y opposent [à la création d'un statut de cadre dirigeant] **puisqu**'ils prétendent représenter l'ensemble des employés face au patronat
(Trade unions oppose it [the creation of a senior management status] since they claim to represent all employees against employers)

Finally, the occurrences of the frame FR-Cause_Enunciation specific to pragmatic or speech-act level causal links show a phenomenon that, to the best of my knowledge, has not been described previously, except for a brief hypothesis in [3]. Only in few of these 19 occurrences is the effect a standard explicit speech act, e.g., an order, a recommandation or a rhetorical question. The majority are 13 cases of explanation of presupposition where the effect is an **implicit** speech act, the expression of that presupposition. Below are two examples.

(4) Cette mesure est justifiée par la fin de l'hyperinflation au Mexique. La hausse des prix de détail a **en effet** atteint 12 % seulement cette année, contre plus de 100 % par an à la fin des années 80.
(This measure is justified by the end of hyperinflation in Mexico. Indeed, the rise in retail prices reached only 12% this year, compared with over 100% a year in the late 1980s.)

(5) "Nous avions bon espoir d'obtenir d'elle un prêt-relais pour acheter les matières premières nécessaires au redémarrage de l'activité, **car** dans l'usine, les machines sont arrêtées depuis le 14 janvier dernier" explique le directeur d'EFI Michel Balandier.
("We were hopeful to get a bridge loan to buy the raw materials needed to restart the

production, because in the factory, the machines are stopped since last January 14th" explains EFI's CEO Michel Balandier.)

In (4), the second sentence including *en effet* (indeed) justifies the presupposition carried by the definite description *la fin de l'hyperinflation* (the hyperinflation ending): the inflation rate is indeed now considerably lower than it used to be. In (5), the proposition introduced by *car* (because) justifies the presupposition carried by the definite description *le redémarrage de l'activité* (production restarting): the production has indeed stopped.

Such examples show that sophisticated annotation tools for DRs should include the possibility to annotate spans that are not standard discourse units but any constituent that may carry a presupposition, as done in Asfalda for these cases. Moreover, one may wonder whether another DR dedicated to presupposition explanation could be necessary.

4 Conclusion

I believe Asfalda has a large potential to study discourse relational devices and their annotation. Beyond the few examples given here, the fact that Asfalda is originally not a DRD resource and includes nouns, verbs and adjectives in its lexicon makes it an excellent tool to study of the fuzzy boundary between causal discourse markers and the expression of a causal link within the propositional content of an elementary discourse unit.

I have here included in the sub-resource extracted from Asfalda prepositions and other constructs taking an event noun as complement, like *à cause de*, *en raison de*, *suite à*, *vu* as discourse markers, like done in Annodis and Explicadis. But this is still controversial and such lexical units are absent from LexConn. Further study of their behaviour in discourse is probably necessary to settle the issue. Because all sorts of parts of speech are annotated with the same set of frames, Asfalda provides an excellent starting point for this.

Acknowlegments

This work was funded by the French National Research Agency (ASFALDA project ANR-12-CORD-023). It owes much from colleagues involved in the Asfalda project; special thanks are due to those deeply involved in the development of the Causation domain together with me, Marie Candito and Philippe Muller. I gratefully acknowledge the work of the annotators, Marjorie Raufast and Anny Soubeille.

Thanks to Caroline Atallah and Myriam Bras for the many discussions on the analysis of causal DRs in general.

References

1. N. Asher and A. Lascarides. *Logics of Conversation.* Cambridge University Press, Cambridge, 2003.
2. C. Atallah. *Analyse de relations de discours causales en corpus: étude empirique et caractérisation théorique.* PhD thesis, Université de Toulouse, 2014.
3. C. Atallah. La ressource EXPLICADIS, un corpus annoté spécifiquement pour l'étude des relations de discours causales. In *Actes de TALN 2015*, 2015.
4. C. F. Baker, C. J. Fillmore, and J. B. Lowe. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August 1998. Association for Computational Linguistics.

5. M. Candito, P. Amsili, L. Barque, F. Benamara, G. De Chalendar, M. Djemaa, P. Haas, R. Huyghe, Y. Y. Mathieu, P. Muller, B. Sagot, and L. Vieu. Developing a French FrameNet: Methodology and first results. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Language Resources and Evaluation Conference (LREC)*, pages 1372–1379, Reykjavik, 2014. European Language Resources Association (ELRA).
6. L. Danlos, M. Colinet, and J. Steinlin. FDTB1, première étape du projet "French Discourse Treebank" : repérage des connecteurs de discours en corpus. *Discours*, 17, 2015.
7. L. Degand and H. Pander Maat. A contrastive study of dutch and french causal connectives on the speaker involvement scale. In A. Verhagen and J. van de Weijer, editors, *Usage-based Approaches to Dutch*, pages 175–199. LOT, Utrecht, 2003.
8. M. Djemaa, M. Candito, P. Muller, and L. Vieu. Corpus annotation within the French FrameNet: a domain-by-domain methodology. In N. Calzolari, K. Choukri, T. Declerck, Goggi, and Grobelnik, editors, *Language Resources and Evaluation Conference (LREC 2016)*, pages 3794–3801, Portoroz, Slovenia, 23-28 May 2016. European Language Resources Association (ELRA).
9. Groupe λ-l. Car, parce que, puisque. *Revue Romane*, 10:248–280, 1975.
10. M.-P. Péry-Woodley, N. Asher, P. Enjalbert, F. Benamara, M. Bras, C. Fabre, S. Ferrari, L.-M. Ho-Dac, A. Le Draoulec, Y. Mathet, P. Muller, L. Prévot, J. Rebeyrolle, L. Tanguy, M. Vergez-Couret, L. Vieu, and A. Widlocher. ANNODIS: une approche outillée de l'annotation de structures discursives (poster). In *Actes de TALN'09*, 2009.
11. C. Roze, L. Danlos, and P. Muller. LEXCONN: a French lexicon of discourse connectives. *Discours*, 10, 2012.
12. E. Sweetser. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge University Press, Cambridge, 1990.
13. L. Vieu, P. Muller, M. Candito, and M. Djemaa. A general framework for the annotation of causality based on FrameNet. In N. Calzolari, K. Choukri, T. Declerck, Goggi, and Grobelnik, editors, *Language Resources and Evaluation Conference (LREC 2016)*, pages 3807–3813, Portoroz, Slovenia, 23-28 May 2016. European Language Resources Association (ELRA).
14. S. Zufferey. 'car, parce que, puisque' revisited: three empirical studies on french connectives. *Journal of Pragmatics*, 44(2):138–153, 2012.

TED Multilingual Discourse Bank: A Parallel Resource Annotated in the PDTB Style

Deniz Zeyrek[1], Amália Mendes[2], and Murathan Kurfalı[1]

[1] Informatics Institute, Middle East Technical University, Ankara, Turkey
dezeyrek@metu.edu.tr, kurfali@metu.edu.tr
[2] Centro de Linguística da Universidade de Lisboa, University of Lisbon, Lisbon, Portugal
amaliamendes@letras.ulisboa.pt

Abstract. We describe TED Multilingual Discourse Bank, or TED-MDB, an effort of annotating TED talks transcripts of multiple languages in the PDTB style. The corpus involves transcripts of the selected talks in English, the pivot language, and the translations to five languages, i.e. European Portuguese, German, Turkish, Polish and Russian. We describe the steps in developing the corpus and how we adapt the PDTB guidelines according to the characteristics of the TED talks transcripts.

Keywords: multilingual corpus, TED talks, discourse, annotation

1 Introduction

A parallel corpus is a compilation of translated texts involving two or more languages. In recent years, there has been an upsurge of interest in parallel corpora for research purposes in linguistics, natural language processing and translation studies. Parallel corpora are useful for language teachers and researchers [7], particularly if they cover an array of genres or different languages. A well-known parallel corpus, Europarl [5] is rather limited in its coverage as it only includes European languages. Other parallel corpora, such as WIT3 [2], offers translated transcripts of TED talks in over 100 languages and can be used in comparative linguistics, translation studies or as an input to various language technology applications quite effectively.

TED Multilingual Discourse Bank (TED-MDB) is a recent effort that arose within the COST project Textlink[3]. It is a multilingual corpus of TED talks transcripts selected from the WIT3 website[4]. The corpus includes the pivot language, English and the translations of the transcripts to 5 languages (European Portuguese, German, Turkish, Polish and Russian) annotated at the discourse level in the PDTB style [8]. In the project, the linguistic characteristics of individual languages are kept in perspective so as to find common annotation solutions to discourse-level issues that arise. Given that TED talks are formal

[3] http://textlink.ii.metu.edu.tr/
[4] https://wit3.fbk.eu/

speeches delivered to a live audience, the transcripts involve aspects of spoken discourse. TED-MDB aims to annotate those aspects of spoken discourse to the extent they appear in the transcripts. In the rest of this work, we describe the stages in the development of the corpus, give an overview of its major annotation categories with relevant examples and describe its coverage.

2 TED Multilingual Discourse Bank

In selecting the TED talks transcripts for inclusion in the corpus, the content of the talks and the translation quality of the transcripts is considered. For example, the talks that rely on too many images and those which are not at the expected level of translation quality are avoided. The transcribed texts (both in English and the existing translations) are saved as text files for use in the annotation tool. All the annotations are created manually. In the rest of the paper, only English examples are provided from the corpus as representative of the annotations.

2.1 Principles and Major Annotation Categories

TED-MDB adopts PDTB's lexicalized approach to discourse connectives [11], i.e. it annotates discourse relations to the extent they are signalled by an anchor word or phrase. The anchor word or phrase is an overt explicit discourse connective (e.g. a coordinating conjunction, a subordinating conjunction, an adverb), or a potential discourse connective that can be inserted in an implicit relation. Discourse connectives are taken as discourse-level predicates with binary arguments, called Arg1 and Arg2. The arguments to a connective are determined on the basis of Asher's abstract object criterion [1]. Thus, TED-MDB annotates explicit discourse connectives, implicit discourse connectives and alternative lexicalizations (AltLex) [10] together with their binary arguments that have abstract object interpretations. A sense tag is assigned to these three discourse relation types from the PDTB 3.0 sense hierarchy [13]. Multiple senses can also be assigned to a single relation. In addition to these, entity relations (EntRel) and no relations (NoRel) are annotated (without assigning senses to them). The annotation tool is the PDTB annotation tool [6].

While presenting TED talks, the speakers often integrate (oral) rhetorical practices without losing sight of the formal aspects of their talk. This led us to extend the PDTB scheme and its principles in certain ways, particularly to capture the relations other than semantic (or informational) relations, which PDTB mainly aims to annotate. For example, we introduce a new category, namely, Hypophora to capture the relation of Question-Answer pairs which we come across in TED talks (example 1). In the examples throughout the paper, Arg1 is shown in italic fonts, Arg2 is rendered in bold fonts. The discourse connective or alternative lexicalization is underlined. The sense of the relation is provided in square brackets.

1. **Do companies that take sustainability into account really do well financially?** *The answer that may surprise you is yes.* [Hypophora]

Secondly, we use the NoRel tag for various purposes, e.g. to mark adjacent sentences that bear no semantic link at the local level (as in PDTB), to indicate topic shifts as in example (2), and to distinguish the discourse connective use of conjunctions such as (*but, so*) from their discourse marker use [12], as in example (3). In so doing, our aim is to record these tokens for analysis at a further stage. The NoRel tag, therefore, is a convenient label we chose to use to indicate relations other than semantic relations.

2. That's the equivalent of taking 21,000 cars off the road. *So awesome, right?* **Another example is Pentair.** [NoRel]
3. *Resist this if you can. Don't do this at home.* But **it makes me wonder if the investment rules of today are fit for purpose tomorrow.** [NoRel]

In the same spirit, if pragmatic markers, e.g. discourse particles used to fill pauses or to show attitudinal meanings [3,4] (e.g. *well*) appear in one of the arguments of the relation, the relation is annotated in the usual manner but the discourse particle itself is not annotated, being left for analysis later together with other aspects of spoken discourse integrated in the transcripts. Example (4) illustrates one of these tokens. Note that this example is an implicit relation, which can be made explicit with the connective *except*.

4. *..the odds that it's not completely wrong are better than the odds that our house will burn down or we'll get in a car accident.* Well, (Implicit=except) **maybe not if you live in Boston.** [Expansion:Exception:Arg2-as-exception]

2.2 The annotation procedure

As in all annotation projects, we start with a set of guidelines. In our case, the guidelines include a summary of PDTB guidelines, examples, and our project-internal principles prepared to familiarize the annotators with the task. The annotators are either experienced annotators or researchers in discourse. Hence, they function as the primary annotator; the annotations are checked by a secondary annotator or a researcher afterwards (this we refer to as the sanity check).

After the annotations have been created and a sanity check has been performed, they are compared with the annotations of other teams in regular meetings. Here, the aim is to control the annotations for correctness across languages (e.g. to make sure that no relation has been missed) and ensure their compatibility with the guidelines. Where needed, the annotation guidelines are updated and the cycle is repeated.

The annotation workflow involves searching and annotating explicit discourse connectives both at the inter-sentential position (example 5) and the intra-sentential position (example 6) (we consider explicit connectives 'easy' to find and annotate).

5. *..About 80 percent of global CEOs see sustainability as the root to growth in innovation and leading to competitive advantage in their industries. But 93 percent see ESG as the future, or as important to the future of their business.* So **the views of CEOs are clear.** [Contingency:Cause:Result+SpeechAct]
6. *Resist this*, if **you can.** [Contingency:Condition:Arg2-as-condition]

We also search and annotate implicit relations (example 7), EntRels (example 8) and NoRels (examples 2, 3 above) between two adjacent sentences delimited by a full stop, question mark or an exclamation mark, as in the PDTB [9]. Annotation of implicit relations that hold intra-sententially is futher work.

7. *The answer that may surprise you is yes.* (Implicit=In fact) **The data shows that stocks with better ESG performance perform just as well as others.** [Expansion:Level-of-detail:Arg2-as-detail]
8. *In blue, we see the MSCI world.* **It's an index of large companies from developed markets across the world.** [EntRel]

During the annotation procedure, we want the annotators to pay attention to the incremental flow of discourse just as in real life. Hence, we ask the annotators to go over the whole text sentence by sentence spotting and annotating the discourse relations as they appear in the text. In order to avoid bias from the pivot language, we chose to annotate the talks (6 in total) without annotation projection.[5] This slows down the process but we believe pace can be compromised for annotation quality.

The current coverage of TED-MDB is visualized in Table 1, showing the distribution of major discourse relation types and Hypophora across the transcripts considered.

Table 1. Absolute frequencies of top level senses and Hypophora across the transcripts in six languages [15]

Language	Comparison	Contingency	Expansion	Hypophora	Temporal
English	71	132	281	11	46
Russian	56	114	270	12	30
Polish	82	108	183	8	44
Portugese	71	143	288	14	54
German	56	120	259	9	31
Turkish	74	146	307	14	41

3 Conclusion

We described TED-MDB, a corpus of TED talks transcripts in English and their translations to multiple languages. We explained the steps in the development of

[5] The term annotation projection refers to projecting linguistic analysis from one language to another via word-aligned parallel bilingual corpora [14].

this resource. Our aim is to capture the features of spoken discourse integrated in the transcripts. We described how we extended the PDTB scheme to fulfill this aim. In the future, our goal is to analyze and annotate more transcripts to reach a more complete understanding of the discourse of TED talks through their transcripts.

4 Acknowledgment

Thanks to the members of the TED-MDB team (Maciej Ogrodniczuk, Yulia Grishina, Sam Gibbon) and the annotators (Nuno Martins, Celina Heliasz, Joanna Bilińska, Daniel Ziembicki). We also thank Bonnie Webber and two anonymous reviewers for constructive suggestions, although any errors are our own.

References

1. Asher, N.: Reference to abstract objects in discourse, vol. 50. Springer Science & Business Media (2012)
2. Cettolo, M., Girardi, C., Federico, M.: Wit3: Web inventory of transcribed and translated talks. In: Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT). vol. 261, p. 268 (2012)
3. Cuenca, M.J., Marín, M.J.: Co-occurrence of discourse markers in catalan and spanish oral narrative. Journal of Pragmatics 41, 899–914 (2009)
4. Fischer, K.: Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume. Approaches to discourse particles pp. 1–20 (2006)
5. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit. vol. 5, pp. 79–86 (2005)
6. Lee, A., Prasad, R., Webber, B.L., Joshi, A.K.: Annotating discourse relations with the pdtb annotator. In: COLING (Demos). pp. 121–125 (2016)
7. McEnery, T., Xiao, R.: Parallel and comparable corpora: What is happening. Incorporating Corpora. The Linguist and the Translator pp. 18–31 (2007)
8. PDTB Group: The Penn Discourse Treebank 2.0 annotation manual. Tech. rep., Institute for Research in Cognitive Science, University of Philadelphia (2008)
9. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A.K., Webber, B.L.: The penn discourse treebank 2.0. In: LREC (2008)
10. Prasad, R., Joshi, A., Webber, B.: Realization of discourse relations by other means: Alternative lexicalizations. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 1023–1031. Association for Computational Linguistics (2010)
11. Prasad, R., Webber, B., Joshi, A.: Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. Computational Linguistics (2014)
12. Schiffrin, D.: Discourse markers. No. 5, Cambridge University Press (1988)
13. Webber, B., Prasad, R., Lee, A., Joshi, A.: A discourse-annotated corpus of conjoined VPs. In: Proc. of the 10th Linguistics Annotation Workshop. pp. 22–31 (2016)

14. Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: Proceedings of the first international conference on Human language technology research. pp. 1–8. Association for Computational Linguistics (2001)
15. Zeyrek, D., Mendes, A., Kurfalı, M.: Multilingual extension of pdtb-style annotation: The case of ted multilingual discourse bank. In: LREC (2018)

Adding Senses and New Discourse Relations to Turkish Discourse Bank: Recent Updates

Deniz Zeyrek[1], Nihan Soycan,[1], Arzu Burcu Güven,[1], and Murathan Kurfalı[1]

Informatics Institute, Middle East Technical University, Ankara, Turkey[1]
dezeyrek@metu.edu.tr, kurfali@metu.edu.tr, nihansoycan@gmail.com,
arzuburcuguven@gmail.com

Abstract. We describe the recent enhancements on Turkish Discourse Bank, namely, the updates we implemented following the revised PDTB 3.0 sense hiearchy and the addition of converbs (suffixal connectives) along with their senses. We explain the automatic revision phase and the manual annotation phase that took place and provide examples. We conclude with an evaluation of the converbs' senses and corpus statistics describing the coverage of the corpus. The enrichments are aimed to contribute to the development of a new version of corpus.

Keywords: Turkish, discourse, annotation

1 Introduction

In the last decade, the release of discourse-annotated corpora, such as RST Discourse Treebank [3], Discourse Graph Bank [10] and Penn Discourse Tree Bank (PDTB) [8] have been highly useful in understanding the phenomena surrounding discourse. Among these, the PDTB annotation scheme has led to reliable results in annotation projects in a number of languages (e.g. Arabic [2], Hindi [6], Chinese [15],Turkish [12]) and has revealed interesting discourse-level phenomena specific to those languages.

Turkish Discourse Bank (TDB) version 1.0 is a 400.000-word multi-genre corpus of written Turkish following the rules and principles of PDTB. It annotates 8483 relations for major discourse connective types and their binary arguments [4].[1] It postpones the annotation of other relation types and their senses to a later stage. Recently, a 40.000-word-sub-corpus of TDB has been enriched with PDTB 2.0 sense hierarchy. This version is referred to as TDB 1.1 [12]. On the other hand, the PDTB team has introduced a new, revised scheme for senses, called PDTB 3.0. The revised PDTB scheme preserves the earlier top-level senses (TEMPORAL, COMPARISON, CONTINGENCY, EXPANSION), updates the name of some earlier sense tags, eliminates certain subsenses and adds new subsenses that are missing in the previous version. These revisions result in a richer sense scheme with a flatter hierarchy [9]. We believed the revised

[1] In TDB 1.0, approximately 5% of the annotated relations consist of phrasal expressions, such as *bu nedenle* 'for this reason'.

PDTB hierarchy would be useful to capture the senses in other languages and hence, we decided to enhance TDB's coverage with the revised PDTB sense hierarchy. In addition to the updates implemented following PDTB 3.0, we added new intra-sentential explicit discourse connectives to the TDB corpus. These are called converbs (suffixal connectives) which are missing in the earlier version. We describe these enrichments with examples and provide relevant corpus statistics.

2 Enrichment of TDB 1.1 with PDTB 3.0 sense hierarchy

TDB 1.1 is a corpus of 20 text files, each with approximately 2.000 words with the genre distribution as shown in Table 1 below.

Table 1. The genre distribution in TDB 1.1

Genre	Number of files	%
Fiction	7	35%
News	6	30%
Research monograph	2	10%
Magazine article	2	10%
Memoir	2	10%
Interview	1	5%
Total	20	100%

Following the PDTB, TDB 1.1 annotates relations made salient by an overt connective ('explicit connectives'), relations not marked by a connective ('implicit connectives'), alternative lexicalizations [7] and entity relations.[2] The enrichments on TDB took place in two steps: an automatic update phase on the name of the sense tags, and a manual update phase involving the newly introduced senses. Table 2 provides a list of the senses affected by the revisions. 28.68% of the updates are implemented manually and the remaining 71.32% are performed automatically through a simple script.

In what follows, we describe these phases with examples. In the examples throughout the paper, the discourse connective is underlined, Arg2 (the text span that hosts the connective) is rendered in bold, Arg1 (the argument that is semantically linked to Arg2) is in italics. As in PDTB, we insert an overt connective in the sentence to make an implicit relation explicit and show the implicit discourse connectives in parentheses.

[2] Entity relations are not assigned sense. They are annotated in TDB 1.1 but are out of the scope of the current work.

Table 2. List of the senses updated automatically or manually

Old Sense	Updated Sense	Method
Exp.Alternative.Chosen alternative	Exp.Substitution.Arg2-as-subst	Auto
Exp.Alternative.Conjunctive	Exp.Conjunction	Auto
Exp.Alternative.Disjunctive	Exp.Disjunction	Auto
Exp.Restatement	Exp.Level-of-detail	Auto
Exp.Restatement.Equivalence	Exp.Equivalence	Auto
Exp.Restatement.Generalization	Exp.Level-of-detail.Arg1-as-detail	Auto
Exp.Restatement.Specification	Exp.Level-of-detail.Arg2-as-detail	Auto
Comp.Concession.Expectation	Comp.Concession.Arg1-as-denier	Auto
Comp.Contrast.Juxtaposition	Comp.Contrast	Auto
Comp.Contrast.Opposition	Comp.Contrast	Auto
Comp.Pragmatic concession	Comp.Concession+SpeechAct	Auto
Cont.Cause.Reason	Cont.Cause.Reason(Arg1-as-result)	Auto
Cont.Cause.Result	Cont.Cause.Result(Arg2-as-result)	Auto
Cont.Pragmatic cause.Justification	Cont.Cause+belief.Reason	Auto
Cont.Pragmatic condition.Implicit assertion	Cont.Condition+SpeechAct	Auto
Cont.Pragmatic condition.Relevance	Cont.Condition+SpeechAct	Auto
Cont.Cause.Reason	Cont.Cause+Belief.Reason	Manual
Cont.Cause.Reason	Cont.Cause+SpeechAct.Reason	Manual
Cont.Cause.Result	Cont.Cause+Belief.Result	Manual
Cont.Cause.Result	Cont.Cause+SpeechAct.Result	Manual
Cont.Condition	Condition.Arg1-as-cond	Manual
Cont.Condition	Condition.Arg2-as-cond	Manual
Cont.Condition	Condition+SpeechAct	Manual
Comp.Concession.Contra-expect	Comp.Conc+SpeechAct.Arg2-as-den	Manual

2.1 Automatic updates

The changes that involve a simple revision of the sense label are automatically implemented. For example, in (1) and (2), the older sense tags are easily replaced by the new labels.

1. *Öğütme taşları çok büyük,* ama **bir kısmı hem öğütme işleminde, hem de belki taş işçiliğinde kullanılıyor.**
   The ground stones are huge but **some of them were used in the grinding process and perhaps in stonemasonry too.** [Genre: Interview]
   COMPARISON: Concession: Arg2-as-denier
   (was COMPARISON: Concession: Contra-Expectation)

2. *Bu evler kerpiçten yapılıyor, içten ve dıştan sıvalı.* (IMP=Ayrıca) **İki gözlü, üç gözlü, hatta beş gözlü olanları var.**
   These houses are made of adobe, they are plastered inside and outside. (IMP=In fact) **Some of them have two, three and even five rooms.** [Genre: Interview]
   EXPANSION: Level of Detail: Arg2-as-detail
   (was EXPANSION: Restatement: Specification)

Example (3) below involves the elimination of the lowest sense level, which is also updated automatically.

3. *Birçok Iraklı yaşamını, devletin her ay karneyle yaptığı gıda yardımı sayesinde sürdürebiliyor.* Oysa **Bağdat'ın açık pazarlarında, parası olan herkes, aradığı her şeyi bulabiliyor.**
   Many Iraqis maintain their lives thanks to the monthly food aid rations the government is providing. On the other hand, **anyone who has money can find anything they want in the open bazaars of Baghdat.** [Genre: News]
   (COMPARISON: Contrast)
   (was COMPARISON: Contrast: Opposition)

2.2 Manual updates

In the second phase of the revisions, the corpus was updated by adding the speech-act or belief features on some of the existing senses. To this end, two native speaker annotators (graduate students in Cognitive Science program at Middle East Technical University) were recruited.[3] They were novice annotators who had good knowledge of language. They took a graduate-level course on discourse, which familiarized them with discourse mechanisms. Before the annotations started, they were trained in the annotation guidelines, the annotation tool (Discourse Annotation Tool for Turkish, [1]) and were familiarized with the PDTB 3.0 scheme to reassign the updates on the already existing sense labels. They also checked the relations in terms of correctness to ensure that the previous argument span selection is aligned with the TDB annotation guidelines and is free of human errors.[4]

The annotators worked individually (approximately 3 hours per week) and had regular adjudication meetings including the project manager, where a unanimously agreed version is created for each relation token under examination. Sense revisions were completed in 2 months. Examples (4) and (5) illustrate some of the changes implemented in this phase.

4. *...tırmanmaya başlandı mı bitirilmeli!* Çünkü **her seferinde acımasız bir geriye dönüş vardı.**
   ...one must finish when he starts climbing! Because **each time there was a relentless comeback.** [Genre: Novel]
   CONTINGENCY: Cause + Belief: Reason
   (was CONTINGENCY: Pragmatic Cause: Justification)

5. *Sana kahve yapacağım.* Ama **çok içmedim.**
   I will make you some coffee. But **I haven't drunk much.** [Genre: Novel]
   COMPARISON: Concession + Speech Act: Arg2-as-denier
   (was COMPARISON: Pragmatic Contrast)

[3] The annotators are also contributing to the current work as the second and third coauthors.
[4] See [11] for the major principles that guide the TDB annotation manual.

In the manual annotation phase, the relations missing in the PDTB 2.0 sense hiearchy are also searched and added to the corpus. An example is provided in example (6).

6. *Analizler aynı sonuçları verirse, bakırın oradan alındığını öğreneceğiz.* Değilse, **başka yerlerde arayacağız.**
   If the analyses give the same results, we will conclude that the use of copper was learnt there. If not, **we will examine other places.** [Genre: Interview]
   CONTINGENCY:Negative Condition (New Sense)

3 Addition of a new explicit intra-sentential connective type: Converbs

In Turkish, discourse relations can be signaled both lexically and morphologically [14]. In TDB 1.0, only lexically signaled discourse connectives are annotated, such as conjunctions, adverbs and phrasal expressions (see footnote 1). The clearest case of marking a relation morphologically is converbs, e.g. -(I)ncA 'when', -(y)ken 'while', -Ip 'and (then)' etc., which are a typical aspect of Turkic languages corresponding to English subordinating or coordinating conjunctions (see examples 7 and 8). [5]

7. **.. bir bakıma kendine de gönderme yap**[arak], *yazılış mantığını sorgular.*
   .. he questions the wording (by) **referring to himself in a way**-[Conv] [Genre: Research]
   EXPANSION: Manner: Arg1-as-Manner

8. **Raif Bey bu kez masama gel**[ip] *önüme mor bir iki buçukluk atarak...*
   This time, Mr. Raif came by my desk-[Conv] (and) *throwing a purple two-and-a-half lira banknote...* [Genre: Novel]
   EXPANSION: Conjunction

We added 104 converb tokens to TDB 1.1 together with their arguments and PDTB 3.0 sense labels. Using the exact match criterion [5], we calculated inter-annotator agreement on the sense labels and obtained $\geqslant 0.8$ on each level of the PDTB 3.0 sense hierarchy (see Table 3) [13]. Table 4 provides the distribution of the senses annotated for the converbs.

The addition of the converbs yielded a total of 867 explicit relations with 671 (77.4%) intra-S DC tokens and 196 inter-S DC tokens (22.6%).

4 Summary and conclusion

We described the recent enhancements in TDB 1.1, which primarily focuses on the updates implemented according to the revised PDTB 3.0 sense hierarchy. It

[5] The capital letters indicate that the vowels are rendered differently in each word depending on vowel harmony rules.

Table 3. IAA results of converb senses in TDB

Sense	IAA
Level-1	89.5%
Level-2	81.9%
Level-3	80.0%

Table 4. Distribution of senses in converbs

Sense	Frequency
Comparison: Concession : Arg1-as-denier	3
Contingency: Cause: Reason(Arg1-as-result)	9
Contingency: Cause: Result (Arg2-as-result)	6
Contingency: Condition: Arg2 as condition	3
Contingency: Negative-condition: Arg2-as-negcond	2
Expansion: Conjunction	13
Expansion: Manner: Arg2-as-manner	12
Expansion: Substitution: Arg1-as-subst	2
Temporal: Asynchronous: Precedence	1
Temporal: Asynchronous: Succession	13
Temporal: Synchronous	48

also involves the addition of converbs together with their binary arguments and senses. We described the annotation cycle of the manual revision phase, which can be summarized as follows:

- Familiarization with TDB guidelines.
- Familiarization with PDTB 3.0 senses.
- Individual annotation of predetermined relations.
- Adjudication meetings to give a final unanimous decision for each relation token under consideration.
- Quality check of the revised annotations.
- Creation of an agreed version for each token to be added to the corpus.

The new version of TDB thus being developed is aimed to be a modest but a relatively more complete version of local discourse structure and semantics involving the major relation types along with their binary arguments and senses.

5 Acknowledgment

This research has been partially supported by METU Project Funds no. BAP-07-04-2017-001. Thanks to two anonymous reviewers for useful comments.

References

1. Aktaş, B., Bozsahin, C., Zeyrek, D.: Discourse relation configurations in Turkish and an annotation environment. In: Proc. of the 4th Linguistic Annotation Workshop. pp. 202–206. ACL (2010)

2. Al-Saif, A., Markert, K.: The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic. In: LREC (2010)
3. Carlson, L., Okurowski, M.E., Marcu, D.: RST Discourse Treebank. Linguistic Data Consortium, University of Pennsylvania (2002)
4. Demirşahin, I., Zeyrek, D.: Pair annotation as a novel annotation procedure: The case of Turkish Discourse Bank. In: Handbook of Linguistic Annotation, pp. 1219–1240. Springer (2017)
5. Miltsakaki, E., Prasad, R., Joshi, A.K., Webber, B.L.: The Penn Discourse Treebank. In: LREC (2004)
6. Oza, U., Prasad, R., Kolachina, S., Sharma, D.M., Joshi, A.: The Hindi Discourse Relation Bank. In: Proc. of the 3rd Linguistic Annotation Workshop. pp. 158–161. Association for Computational Linguistics (2009)
7. Prasad, R., Joshi, A., Webber, B.: Realization of discourse relations by other means: Alternative lexicalizations. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 1023–1031. Association for Computational Linguistics (2010)
8. Prasad, R., Webber, B., Joshi, A.: Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. Computational linguistics (2014)
9. Webber, B., Prasad, R., Lee, A., Joshi, A.: A discourse-annotated corpus of conjoined VPs. In: Proc. of the 10th Linguistics Annotation Workshop. pp. 22–31 (2016)
10. Wolf, F., Gibson, E., Fisher, A., Knight, M.: The Discourse Graphbank: A database of texts annotated with coherence relations. Linguistic Data Consortium (2005)
11. Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., Çakıcı, R.: Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. Dialogue and Discourse 4(2), 174–184 (2013)
12. Zeyrek, D., Kurfalı, M.: TDB 1.1: Extensions on Turkish Discourse Bank. LAW XI 2017 p. 76 (2017)
13. Zeyrek, D., Kurfalı, M.: An assessment of explicit inter- and intra-sentential discourse connectives in Turkish Discourse Bank. In: LREC (2018)
14. Zeyrek, D., Webber, B.L.: A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. In: IJCNLP. pp. 65–72 (2008)
15. Zhou, Y., Xue, N.: The Chinese Discourse Treebank: a Chinese corpus annotated with discourse relations. Language Resources and Evaluation 49(2), 397–431 (2015)