Methods and Tools for the analysis of discourse relational devices

Training School

COST Action IS1312

TextLink:

Structuring Discourse in Multilingual Europe

València (Spain), 18th-22nd January, 2016

Meeting Details

Host Institution: University of Valencia

Location: Valencia (Spain), Facultat de Filologia, Av. Blasco Ibáñez 32

TextLink Training School officer: Maria Josep Cuenca

Organizing committee coordinators: Maria Josep Cuenca & Salvador Pons

Aims

- Delivering intensive training on theoretical and empirical approaches to the description, identification and annotation of DRDs in multiple languages.
- Providing a theoretical and empirical introduction to discourse coherence with a focus on DRDs.
- Training in methods and tools for the identification and annotation of DRDs in authentic data, including statistical methods for data analysis.
- Improving trainer-trainee collaboration, fostering of ideas for future collaboration.

Structure

General sessions

Topic 1 (2 sessions): Coherence relations and DRD identification
Topic 2 (3 sessions): Cross-linguistic variation
Topic 3 (3 sessions): Corpus research, annotation theories and tools
Topic 4 (2 sessions): Dictionaries/Lexicons

Lab sessions

Hands-on sessions focusing on technical aspects and statistics. Specific topics: how to assemble a sound corpus, reliability, annotator agreement, experimental design, hypotheses, analysis of results, and so on.

Speed dating

Trainers and trainees talk for a short period of time about their current research or research interests. Meet up with other researchers in the same field, discuss crucial questions and methodology, share your insights and experiences. Trainees must be prepared to summarize their research in 5 minutes.

Schedule

Registration: Monday 18, 8:15-8:40 (4th floor)

Inaugural ceremony: Monday 18, 8:40-9:00 (Room 407, 4th floor)

	General session Room 406 9-10.30	General session Room 406 11-13.30	Lab session Room 406 15.00-18
Mon 18	Coherence relations and DRD identification: theory and analysis T. Sanders / W. Spooren	Annotation theories and tools A. Nedoluzhko	Lab: Corpus research: Methodology and statistics W. Spooren/ T. Sanders
Tues 19	Coherence relations and DRDs identification: converging evidence T. Sanders / W. Spooren	Corpus exploration of discourse relations in PDT 3.0 and PDTB J. Mírovský	Lab: Annotation tools A. Nedoluzhko
Wed 20	Cross-linguistic variation of DRDs J. Visconti	Corpus exploration of discourse relations in RST M. Iruskieta	*Research 'speed dating'* (Room: Espai cultural, 1st floor)**
Thur 21	Cross-linguistic variation: DRDs identification and annotation L. Degand/ S. Zufferey	Building discourse relational device lexicons L. Danlos	Lab: Working with cross-linguistic data S. Zufferey/L. Degand

Fri 22	Typology and DRDs V. Gast	Methodological issues on DRDs dictionary construction: The case of the DPDE S. Pons	Lab: Machine translation techniques to induce multilingual lexica of discourse markers D. Martin de Matos

**The *Research 'speed dating'* session will take place at Room: Espai cultural (1st floor).

Description

Coherence relations and DRD identification: theory and analysis

Trainers: Ted Sanders (Universiteit Utrecht) / Wilbert Spooren (Radboud Universiteit Nijmegen)

Language users communicate through discourse. The constituting property of discourse is that it shows coherence: people make a coherent mental representation of the information in the discourse. The discourse itself contains (more or less) overt signals that direct this interpretation process, among them discourse relational devices (DRDs) like connectives and cue phrases.
We focus on *coherence relations* that establish the relationship between discourse segments, such as *Cause-Consequence* and *Contrast*. These relations are conceptual and they can, but need not, be made explicit by DRDs (*because*, *so*, *however*, *although*) and lexical cue phrases (*For that reason*, *As a result*, *On the other hand*).
Some dominating accounts of coherence relations are introduced, compared and discussed. We will discuss their use in discourse annotation, as well as their theoretical backgrounds. Special attention will be paid to underlying dimensions on which various accounts converge. In the Lab session, we will work on concrete issues of annotation of corpus case.

Annotation theories and tools

Trainer: Anna Nedoluzhko (ÚFAL, Charles University Prague)

Discourse coherence is a complex natural language phenomenon which is achieved by different linguistic means (e.g., anaphoricity, information structure, discourse markers and connectives, rhetorical structure of text, etc.). Many approaches in computational linguistics are used to capture discourse relations and find practical applications.
This session focuses on discussing theories and approaches applied to the annotation of discourse phenomena in different languages, such as Rhetorical Structure Theory, Penn Discourse Treebank, Segmented Discourse Representation Theory and so on. Participants will have the opportunity to compare these approaches to see which phenomena are central to them and which ones are less prominent. We will also introduce the tools of discourse annotation and demonstrate esp. TrEd, PDTB and

MMAX2.

Corpus research: Methodology and statistics

Trainers: Wilbert Spooren (Radboud Universiteit Nijmegen) / Ted Sanders (Universiteit Utrecht)

This lab session focuses on corpus research at the discourse level: How can you assemble a corpus to do your investigations on, in a valid and reliable way? How can you search in existing corpora and what is best format to annotate and analyze cases? Participants discuss concrete cases taken from corpora of language use in various languages. We analyze coherence relations and DRDs. Annotation Tools – some of them automated – are introduced and put to use. The next methodological step is to determine whether several analysts agree: interrater reliability. Methods to compute this are introduced and participants will use them during the session. Further methodological implications are discussed. Finally, we discuss statistical methods (in SPSS or R) that help answer research questions, such as: do DRDs behave differently in one genre than in another?

Coherence relations and DRD identification: converging evidence

Trainers: Ted Sanders (Universiteit Utrecht) / Wilbert Spooren (Radboud Universiteit Nijmegen)

In a cognitive approach to coherence relations, it is important to account for the relationship between discourse as a linguistic object and the mental representation people have or make of it. Such an approach requires an interdisciplinary methodology of converging evidence. We discuss studies using both linguistic and psycholinguistic research methods and data, varying from text analysis to on-line discourse processing and language acquisition. Finally, we explore the relationship between discourse coherence and genre.

Corpus exploration of discourse relations in PDT 3.0 and PDTB

Trainer: Jiří Mírovský (Charles University Prague)

We will explore discourse relations in the Prague Dependency Treebank and in the Penn Discourse Treebank using the PML-Tree Query system, a general and powerful system for querying treebanks. We will learn basics of the query language on syntactic trees and use it later for searching for discourse relations. We will show that as a result of a query search, we can either get a sequence of individual occurrences of the query pattern in the data, or a summary of the occurrences in the whole data defined by a system of output filters.

Lab: Annotation tools

Trainer: Anna Nedoluzhko (ÚFAL, Charles University Prague)

Following the general session on “Annotation theories and tools”, trainees will be invited to use several annotation tools.

Cross-linguistic variation on DRD

Trainer: Jacqueline Visconti (University of Genoa)

Given the multifunctionality and context-boundness of DRDs, linguistic variation is a tricky question. The first part of this interactive lecture will focus on central theoretical issues in contrastive analysis, such as the notion of tertium comparationis, the balance between monolingual and comparative methods, the role of corpora in contrastive studies. In the second part, a selection of case-studies of contrastive investigations on DRDs in mostly European languages will be highlighted and discussed. methods will rely on multilingual corpora, such as Eurparl or VoxEurop, from which translation equivalents will be elicited as empirical data.

Corpus exploration of discourse relations in RST

Trainer: Mikel Iruskieta (University of the Basque Country)

In the RST framework, there are several discourse-annotated corpora available in individual languages, such as:
English <https://catalog.ldc.upenn.edu/LDC2002T07>,
Spanish <http://corpus.iingen.unam.mx/rst/>,
Brasilian Portuguese <http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm>, German <https://www.ling.uni-potsdam.de/acl-lab/Forsch/pcc/pcc.html>, and <http://ixa2.si.ehu.es/diskurtsoa/en/>, among others.
Some of them can be consulted and several tools have been developed for corpus exploration (exploration honekin zalantza, baina literaturan ikusi baduzu, aurrera!). There is also a small multilingual aligned RST corpus <http://ixa2.si.ehu.es/rst/>, which can be explored for getting information about different linguistic phenomena.
After the annotation process is over, evaluation is necessary to check reliability (precision and recall). In order to do so, a sound evaluation method and some search tools (which can be used in multilingual corpus) were developed: i) to study whether the annotators were consistent when looking for the relations or signals in a kwic style, ii) to check the aligned segments in different languages, iii) to check a kind of macrostructure of RS-tree looking for the RST relations that are linked to the most salient unit, and iv) to look for any information in the corpus based on part of speech.
In this session, I will present this method and the tools developed to consult the multilingual RST treebank we have developed in the Ixa research group at the University of the Basque Country.

Cross-linguistic variation: DRD identification and annotation

Trainers: Liesbeth Degand (Université catholique de Louvain) / Sandrine Zufferey (University of Fribourg)

While DRDs are found in (probably) all languages of the world, important variations exist in the number of DRDs languages display to express a given relation, even between typologically related languages. An overview will be given of the types of variation that exist along different types of dimensions: form vs. function, semasiology

vs. onomasiology, and how they can be empirically investigated on the basis of different types of data: parallel (translation, including fiction, non-fiction, subtitles, ...) or comparable (same text types in different languages), both in speech and writing.

Building discourse relational device lexicons

Trainer: Laurence Danlos (Université Paris Diderot)

Discourse relational devices (DRDs) are (simple or compounds) lexical items that express discourse relations between two discourse segments. For French we developed a lexicon of DRDs which records for each entry its syntactic category and its sense(s) (i.e. which discourse relation(s) it expresses) along with possible other information (e.g. constraint on its position) and examples. A first version, developed from linguistic knowledge, was revised after a discourse annotation experiment, and we will present the two methods. DRDs lexicons exist for other languages, German for example. We will make a comparison of the two resources with a reflection on the methods that was used to build them.

Lab: Working with cross-linguistic data

Trainers: Sandrine Zufferey (University of Fribourg) / L.iesbeth Degand (Université catholique de Louvain)

Following the general session on "Cross-linguistic variation: DRD identification and annotation", trainees will be invited to participate in a multilingual annotation experiment.

Typology and DRD

Trainer: Volker Gast (University of Jena)

The cross-linguistic, corpus-based study of discourse relational devices requires corpora annotated at various levels, minimally syntax and semantics. In this session I will demonstrate how we can create a cross-linguistic sample of sentences annotated at two levels by carrying out the following steps: (i) preparation of the data (e.g. syntactic parsing of a sample), (ii) manual corrections, (iii) enriching the sample with lexical-semantic information, and (iv)manually annotating it with sentence-semantic and pragmatic information.

Methodological issues on DRDs dictionary construction: The case of the DPDE

Trainer: Salvador Pons Bordería (Universitat de València)

Building a dictionary of DRDs seems a paradoxical activity, given the functional explanation assigned to them. Indeed, the lexicographical description of DRDs implies challenges both for the lexicographer and for the pragmatist; the former has to review received wisdom related to what counts as a definition, synonymy or even meaning. The latter has to transform an onomasiological approach, based on functions, into a semasiological approach, based on forms.
This session will reflect on the issues above on the experience from the *Diccionario de Partículas Discursivas del Español* (Briz, Pons and Portolés, online since 2003) (www.dpde.es).

Lab: Machine Translation Techniques to Induce Multilingual Lexica of Discourse Markers

Trainer: David Martin de Matos (University of Lisbon)

Discourse markers are universal linguistic events subject to language variation. This lab's work contemplates new methods and approaches for the description, classification, and annotation of discourse markers in the specific domain of the Europarl corpus. The study of discourse markers in the context of translation is crucial due to the idiomatic nature of these structures. Multilingual lexica together with the functional analysis of such structures are useful tools for the hard task of translating discourse markers into possible equivalents from one language to another. Using Daniel Marcu's validated discourse markers for English, extracted from the Brown Corpus, our purpose is to build multilingual lexica of discourse markers for other languages, based on machine translation techniques.

SOME TECHNICAL INFORMATION

Software

- SPSS. It is possible to download a Free 14-day trial version: https://www-01.ibm.com/marketing/iwm/iwmdocs/tnd/data/web/en_US/trialprograms/W110742E06714B29.html

- In case you do not have Linux installed on your laptop, it is possible to install a Virtual Linux OS (Ubuntu) Operating System inside Windows or Mac by using Oracle Virtualbox (https://www.virtualbox.org/). Once installed, additional software will be required:

- Giza++. http://www.statmt.org/moses/giza/GIZA++.html

- Moses decoder software. http://www.statmt.org/moses/

- Perl. https://www.perl.org/

Anna Nedoluzhko instructions

Install PDTB, RST-web and Brat according to the following instructions (for Linux):
1. Download the archived package from https://ufal.mff.cuni.cz/%7Enedoluzko/tools.tar.gz

2. In your computers, create TOOLS directory and unpack the content of the attached package to it. You will have three directories: PDTB, RST-web and brat.

3. PDTB
a) Make sure that Java is installed.
b) Run start.sh to make sure the tool works.

4. RST-web
a) Make sure Python 2.X is installed (preferably 2.6 or newer)
b) The Python package cherrypy must be installed if it isn't already (e.g. using pip install cherrypy from the command line)
c) Run start.sh to make sure the tool works.
d) Open rstWeb in your browser at: http://127.0.0.1:8080/ (I use Firefox)

5. Brat
a) Make sure Python 2.X is installed (preferably 2.6 or newer)
b) Run start.sh to make sure the tool works.
c) Open brat in your browser at: http://127.0.0.1:8001/ (I use Firefox)
d) To log in, use username: anot, password: anot
In case a red error message in browser arises, ignore it, it seems to have no effect.

However, if the tool still doesn't work, run ./install.sh -u in terminal. You will be asked to enter username and password. Use anot, anot, or any other but remember it.

Generally, for all tools, if any errors or problems arise, don't hesitate to describe them to me (nedoluzko@ufal.mff.cuni.cz), and we will try to solve them together

Classroom facilities

In the classroom, there will be laptops and desktop computers with the software specified above installed. Nevertheless, we encourage you to install the software previously on your own personal laptops.

Internet

Eduroam wireless service is available. All Desktop computers have also cable Internet access. If you do not have access to an institutional Eduroam account, we will provide you with an Eduroam username and password specifically created for the Training School.