Building Lexicons of Discourse Connectives

Lucie Poláková

Institute of Formal and Applies Linguistics
Charles University
Prague, Czech Republic

Text Structure and Corpus Linguistics
2e Journée d’étude franco-tchèque
November 12, 2018
Outline

• Generally on lexicons and their purposes

• Overview and comparison of present-day electronic lexicons of discourse connectives in different languages

• CzeDLex: the Lexicon of Czech Discourse Connectives
Lexicons of Discourse Connectives?
Lexicons of Discourse Connectives?

Is that a new thing?
Lexicons of Discourse Connectives?

Is that a new thing?

Ball, 1993

Métrich et al., 1994

Buscha, 1989

Pasch et al., 2003
Lexicons of Discourse Connectives?

Is that a new thing?

- Ball, 1993
- Métrich et al., 1994
- Pasch et al., 2003
- Buscha, 1989
- DiMLex Stede/Umbach 1998
Lexicography of “function words”

• quite specific area of lexicography as a discipline
• requires different methodology than traditional dictionaries (functional definitions of the category of the lexicon, extent, nesting, the purpose?)
Lexicography of “function words”

• quite specific area of lexicography as a discipline
• requires different methodology than traditional dictionaries (functional definitions of the category of the lexicon, extent, nesting, the purpose?)

  *Don’t call before eight.*
  *He trembled before her.*
  *Before landing, fasten your seat belt.*
  *Think carefully before you choose.*
  *He will resign before accepting it.*
  *We have met before.*
Lexicography of “function words”

- quite specific area of lexicography as a discipline
- requires different methodology than traditional dictionaries (functional definitions of the category of the lexicon, extent, nesting, the purpose?)

  Don’t call before eight.

  He trembled before her.

  Before landing, fasten your seat belt.

  Think carefully before you choose.

  He will resign before accepting it.

  We have met before.

--> what is a discourse connective?
A Discourse Connective is...

Penn Discourse Treebank 2.0 Annotation Manual (Prasad et al. 2007):

“discourse-level predicates that take two abstract objects such as events, states, and propositions (Asher, 1993) as their arguments”

Handbuch der deutschen Konnektoren (Pasch et al. 2003):
1. X cannot be inflected.
2. X does not assign case features to its syntactic environment.
3. The meaning of X is a two-place relation.
4. The arguments of the relation (the meaning of X) are propositional structures.
5. The expressions of the arguments of the relation can be sentential structures.

Discourse markers/particles: one-place relation: yes, yeah, oh, well
A Discourse Connective is...

- Coordinating conjunctions (*and, but, or...*)
- Subordinating conjunctions (*because, although, if...*)
- Adverbs (*then, next, consequently, similarly...*)
- Particle-like expressions including rhematizers (*only, even, also...*)
- Numbers and letters as signals of list structures (*1, 2, a,b...*)
- Two interpuction signs: colon (:) and dash (—)

Lately: Danlos, et al. 2018:

**Primary vs. secondary** connectives: *therefore* x *that’s the reason why*

**Prepositions** (adpositions) as connectives:

*John left* *after taking a shower. Pour aller à l'école, je prends le tram.*
Electronic lexicons & their sources

- Machine-readability (and compatibility) is a HUGE advantage:
  - Linking of similar resources for different languages
  - Use in different NLP tasks (e.g. PDT Vallex, Czech SubLex)
From the Origins to a Connective Database

History and ways of development

• German DiMLex (Diskurs-Marker-Lexikon), 1998 – 2016
• French LexConn (Base lexicale des connecteurs discursifs du français), 2009 – 2016
• Spanish DPDE (Diccionario de partículas discursivas del español), 2003 – 2011
From the Origins to a Connective Database

History and ways of development

• German DiMLex (Diskurs-Marker-Lexikon), 1998 – 2016
• French LexConn (Base lexicale des connecteurs discursifs du français), 2009 – 2016
• Spanish DPDE (Diccionario de partículas discursivas del español), 2003 – 2011
From the Origins to a Connective Database

History and ways of development

**DiMLex 1998:** assembling pieces of information on German and English discourse markers from grammars (Quirk, 1972; Helbig/Buscha 1991), dictionaries (HdK in prep. at that time) and linguistic research literature

**DPDE:** oral: several oral corpora (CREA, CORPES, Val.Es.Co, ALCORE, COVJA...), written: any type of text, preferably journalistic

**LexConn:** manually constructed, the FRANTEXT corpus as empirical support

Recent projects mostly: annotated corpus, annotated for discourse relations
Frameworks of the Lexicons

• Morphosyntactic properties according to the specifics of individual languages (and their grammar tradition)
• Semantic description: frameworks of the underlying corpora
  ▪ SDRT (Segmented Discourse Representation Theory)
    -> LexConn
  ▪ PDTB (Penn Discourse Treebank)
    - -> common framework for lexicons in Connective-Lex
• ! different for spoken markers
<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Authors</th>
<th>Language</th>
<th>Primary aim</th>
<th>Items</th>
<th>Data format</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiMLex</td>
<td>Stede and Umbach 1998, Scheffler and Stede 2016</td>
<td>German</td>
<td>NLP: text generation and understanding</td>
<td>274</td>
<td>xml</td>
</tr>
<tr>
<td>LexConn</td>
<td>Roze et al. 2012</td>
<td>French</td>
<td>NLP: discourse parsing</td>
<td>328</td>
<td>xml</td>
</tr>
<tr>
<td>LDM - PT</td>
<td>Mendes, Lejeune 2016</td>
<td>Portuguese</td>
<td>NLP + lexicon linkage</td>
<td>228</td>
<td>excel -&gt; xml</td>
</tr>
<tr>
<td>DPDE</td>
<td>Briz et al., 2003</td>
<td>Spanish</td>
<td>theoretical research</td>
<td>229</td>
<td>plain text</td>
</tr>
<tr>
<td>PDTB-DiMLex</td>
<td>Scheffler 2017</td>
<td>English</td>
<td>NLP + lexicon linkage</td>
<td>142</td>
<td>xml</td>
</tr>
<tr>
<td>LICO</td>
<td>Feltracco et al. 2016</td>
<td>Italian</td>
<td>NLP + lexicon linkage</td>
<td>173</td>
<td>xml?</td>
</tr>
<tr>
<td>CzeDLex</td>
<td>Synková et al. 2017</td>
<td>Czech</td>
<td>NLP + lexicon linkage</td>
<td>140</td>
<td>xml (pml)</td>
</tr>
<tr>
<td>DisCoDict</td>
<td>Bourgonje et al. 2018</td>
<td>Dutch</td>
<td>NLP + lexicon linkage</td>
<td>142</td>
<td>xml</td>
</tr>
<tr>
<td>Arabic Discourse Cues</td>
<td>Benamara and Keskes 2018, Toulouse</td>
<td>Arabic</td>
<td>NLP + lexicon linkage</td>
<td>390</td>
<td>xml</td>
</tr>
<tr>
<td>Bangla</td>
<td>Das et al., 2018</td>
<td>Bangla</td>
<td>NLP + lexicon linkage</td>
<td>101</td>
<td>xml</td>
</tr>
</tbody>
</table>
Connective-Lex: A Multilingual Connective Database

- Potsdam: M. Stede, T. Scheffler, F. Dombek, D. Das 2017
- Web application
- 8 lexicons so far (German, French, Portuguese, English, Italian, Arabic, Dutch, Bangla)
- Semantic properties of connectives linked via PDTB 3.0 semantic taxonomy (so far only, not via lemmata)

http://connective-lex.info/
### Found 50 matching results.

| LexConrt: 9 | Eng-DiMLex: 23 | LICO: 2 | LDM-PT: 2 | DisCoDict: 1 | DiMLex: 3 | Arabic: 5 | DiMLex-Bangla: 5 |

#### autrement

**adv**

- EXPANSION:Disjunction, coord
  - Si le feu est rouge, tu t'arrêtes. (Autrement/Sinon), tu continues à rouler.
  - Si le feu est rouge, tu t'arrêtes. (Autrement/Sinon), tu auras une amende.

- EXPANSION:Exception, coord
  - J'ai quelques soucis avec mon boulot, en ce moment. (Autrement/Sinon/Moins) tout va bien.

- EXPANSION:Conjunction, coord

#### ou

**cco**

- EXPANSION:Disjunction, coord
  - Ou ça vient n'importe quel jour avec dix francs de fleurs et ça murmure des prières debout ou ça s'affole et ça parle à des pierres tombales.

#### ou bien

**cco**

- EXPANSION:Disjunction, coord
  - Ils nous attendaient là aussi. Ou bien, si ce n’était pas pour nous qu'ils étaient là, ils protégeraient quelque chose.
  - En réalité, la distance qu'il leur restait à couvrir était aussi peu sûre que la route choisie. Ou bien ils rencontraient le roi David avant d'attendre Samarconde. Ou

#### plutôt que de

**prep**

- EXPANSION:Disjunction, coord
  - Je me serais pendu plutôt que d’écrire un vaudeville.
  - Plutôt que de se rendre, ils disparaissent dans la nuit

#### pour le coup

**adv**
The way of CzeDLex

CzeDLex

- CzeDLex lexicon entries: based on manual discourse annotation in the Prague Discourse Treebank 2.0 (Penn Discourse Treebank 2.0 style)
- extraction of more than 900 connective types (incl. complex forms)
  - 30 DCs frequency > 100
  - Limited by register – written journalistic Czech
- lexicon structure largely inspired by the German DiMLex
- XML - encoded
- morphosyntactic and semantic properties of the connectives
- inclusion of secondary connectives
- inclusion of non-connective usages
CzeDLex online

- Version 0.5 released in 2017, online
- Ca. 20 most frequent connectives manually processed
  (= ca. 2/3 of all connectives in the corpus)
- linkage with Connective-Lex in progress
- new grant proposal for 2019 – 2021?
CzeDLex 0.5

**tedy** [so] (primary, single; count: 576)

variants: teda

**connective usages (59%; intra 17%)**

**reason-result** (proto [therefore], 60%; intra 17%; adverb)

[arg_semantics: reason-result:result; ordering: 2; integration: second]

complex_forms: a tedy / : tedy

examples:

- Nejsem fyzicky schopen pronést cokoli bez komentáře. Jak se tedy mohu stát politikem?
- [I am physically unable to state anything without a comment. How then can I become a politician?]
- Jsme tam vlastně už čtyři roky, vina tedy padá i na nás.
  [We have been there already for four years, so the guilt falls on us as well.]

**generalization** (tak [so; that way], 18%; intra 7%; adverb)

[arg_semantics: generalization:less specific; ordering: 2]

complex_forms: a tedy / - tedy

examples:

- Získovost u nás (a v celé Evropě) se pohybuje okolo dvou procent.
  [The thriftiness in us (and in the whole Europe) is around 2%]
- Je to tedy hluboko pod průměrnou úrokovou mírou všech našíh komerčních bank.
Selected References

- Danlos Laurence, Rysová Kateřina, Rysová Magdaléna, Stede Manfred: Primary and secondary discourse connectives: definitions and lexicons. In: Dialogue and Discourse, Vol. 9, No. 1, Copyright © Linguistic Society of America, ISSN 2152-9620, pp. 50-78, 2018
Thank you!
Merci!
Děkuji!

polakova@ufal.mff.cuni.cz

This work was supported by the Grant Agency of the Czech Republic (project GA17-06123S), by the LINDAT-Clarin project (LM2010013) and by the Ministry of Education, Youth and Sports of the Czech Republic (project COST-cz LD15052).
Treq – Translation Equivalent Database

• A Czech project by the Institute of Czech National Corpus, Charles University in Prague, Faculty of Arts
  
  http://treq.korpus.cz

• based on the large parallel corpus Intercorp
• -> not only translations of connectives!

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Proportion</th>
<th>French</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>75987</td>
<td>94.2</td>
<td>mais</td>
<td>but</td>
</tr>
<tr>
<td>2152</td>
<td>2.7</td>
<td>mais</td>
<td>and</td>
</tr>
<tr>
<td>415</td>
<td>0.5</td>
<td>mais</td>
<td>yet</td>
</tr>
<tr>
<td>260</td>
<td>0.3</td>
<td>mais</td>
<td>it</td>
</tr>
<tr>
<td>226</td>
<td>0.3</td>
<td>mais</td>
<td>well</td>
</tr>
<tr>
<td>204</td>
<td>0.3</td>
<td>mais</td>
<td>though</td>
</tr>
<tr>
<td>195</td>
<td>0.2</td>
<td>mais</td>
<td>although</td>
</tr>
<tr>
<td>156</td>
<td>0.2</td>
<td>mais</td>
<td>But</td>
</tr>
<tr>
<td>133</td>
<td>0.2</td>
<td>mais</td>
<td>only</td>
</tr>
</tbody>
</table>