Building Lexicons of Discourse Connectives

Lucie Poláková

Intitute 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

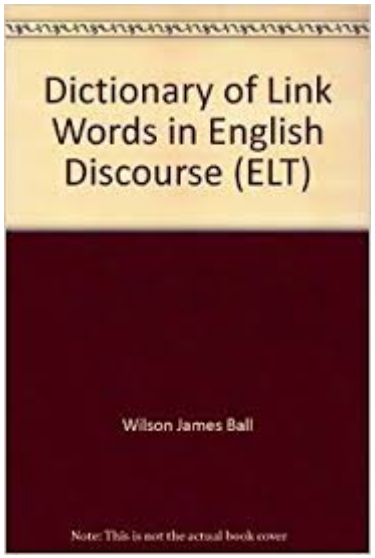
- Generally on lexicons and their purposes
- Overview and comparison of present-day electronic lexicons of discourse connectives in different languages
- CzeDLex: the Lexicon of Czech Discourse Connectives

Lexicons of Discourse Connectives?

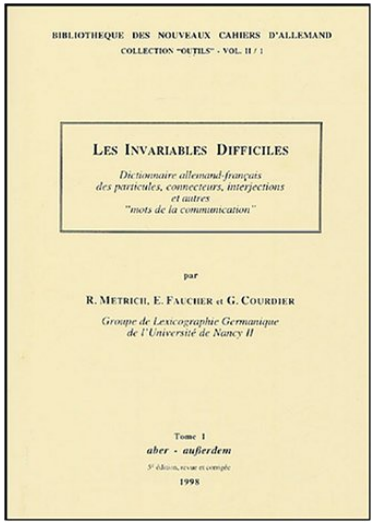
Lexicons of Discourse Connectives?

Lexicons of Discourse Connectives?

Is that a new thing?



Ball, 1993



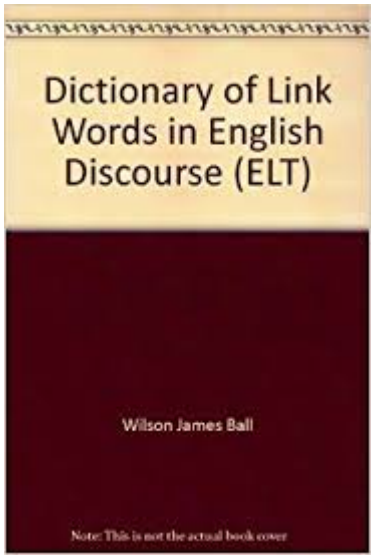
Métrich et al., 1994

Pasch et al., 2003

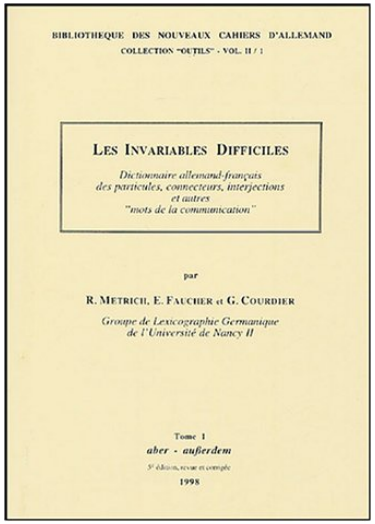
Buscha, 1989

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

Current Lexicons

Lexicon	Authors	Language	Primary aim	Items	Data format
DiMLex	Stede and Umbach **1998**, Scheffler and Stede **2016**	**German**	NLP: text generation and understanding	274	xml
LexConn	Roze et al. **2012**	**French**	NLP: discourse parsing	328	xml
LDM - PT	Mendes, Lejeune **2016**	**Portuguese**	NLP + lexicon linkage	228	excel -> xml
DPDE	Briz et al., **2003**	**Spanish**	theoretical research	229	plain text
PDTB-DiMLex	Scheffler **2017**	**English**	NLP + lexicon linkage	142	xml
LICO	Feltracco et al. **2016**	**Italian**	NLP + lexicon linkage	173	xml?
CzeDLex	Synková et al. **2017**	**Czech**	NLP + lexicon linkage	140	xml (pml)
DisCoDict	Bourgonje et al. **2018**	**Dutch**	NLP + lexicon linkage	142	xml
Arabic Discourse Cues	Benamara and Keskes **2018**, Toulouse	**Arabic**	NLP + lexicon linkage	390	xml
Bangla	Das et al., **2018**	**Bangla**	NLP + lexicon linkage	101	xml

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/

Lexicon Selection

- [x] Arabic Arabic about
- [x] DiMLex-Bangla Bangla about
- [x] DiMLex German about
- [x] DisCoDict Dutch about
- [x] Eng-DiMLex English about
- [x] LDM-PT Portuguese about
- [x] LexConn French about
- [x] LICO Italian about

Search Options

Filter | Word

Syntactic Category

- [ ] cco
- [ ] csu
- [ ] adv
- [ ] prep
- [ ] other

Discourse Relation (PDTB3)

- [ ] COMPARISON
  - [ ] Contrast
  - [ ] Similarity
  - [ ] Concession ([ ] Arg1-/ [ ] Arg2-as-denier)
- [ ] CONTINGENCY
  - [ ] Cause ([ ] Reason/ [ ] Result)
  - [ ] Condition ([ ] Arg1-/ [ ] Arg2-as-cond)
  - [ ] Negative condition ([ ] Arg1-/ [ ] Arg2-as-negcond)
  - [ ] Purpose ([ ] Arg1-/ [ ] Arg2-as-goal)
- [ ] EXPANSION
  - [ ] Conjunction
  - [x] Disjunction
  - [ ] Equivalence
  - [ ] Instantiation
  - [ ] Level-of-detail ([ ] Arg1-/ [ ] Arg2-as-detail)
  - [ ] Substitution ([ ] Arg1-/ [ ] Arg2-as-subst)
  - [ ] Exception ([ ] Arg1-/ [ ] Arg2-as-excpt)
  - [ ] Manner ([ ] Arg1-/ [ ] Arg2-as-manner)
- [ ] TEMPORAL
  - [ ] Synchronous
  - [ ] Asynchronous ([ ] Precedence/ [ ] Succession)

[ ] +Belief [ ] +SpeechAct [ ] Other

"We will EITHER go to Wacken OR decorate our balcony."

Found **50** matching results. LexConn: **9** Eng-DiMLex: **23** LICO: **2** LDM-PT: **2** DisCoDict: **1** DiMLex: **3** Arabic: **5** DiMLex-Bangla: **5**

autrement

Variants

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/Mais) tout va bien.

EXPANSION:Conjunction coord

ou

Variants

cco

EXPANSION:Disjunction coord
Je brûle d'être à la hauteur. Mais je bégaie. Je bafouille. Je balbutie quelque évidence. Ou je lâche une incohérence que je regrette aussitôt.
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

Variants

cco

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

plutôt que de

Variants

prep

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

pour le coup

Variants

adv

The way of CzeDLex

- Mírovský Jiří, Synková Pavlína, Rysová Magdaléna, Poláková Lucie: *CzeDLex – A Lexicon of Czech Discourse Connectives.* In: The Prague Bulletin of Mathematical Linguistics, No. 109, Czech Republic, pp. 61-91, Oct 2017

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

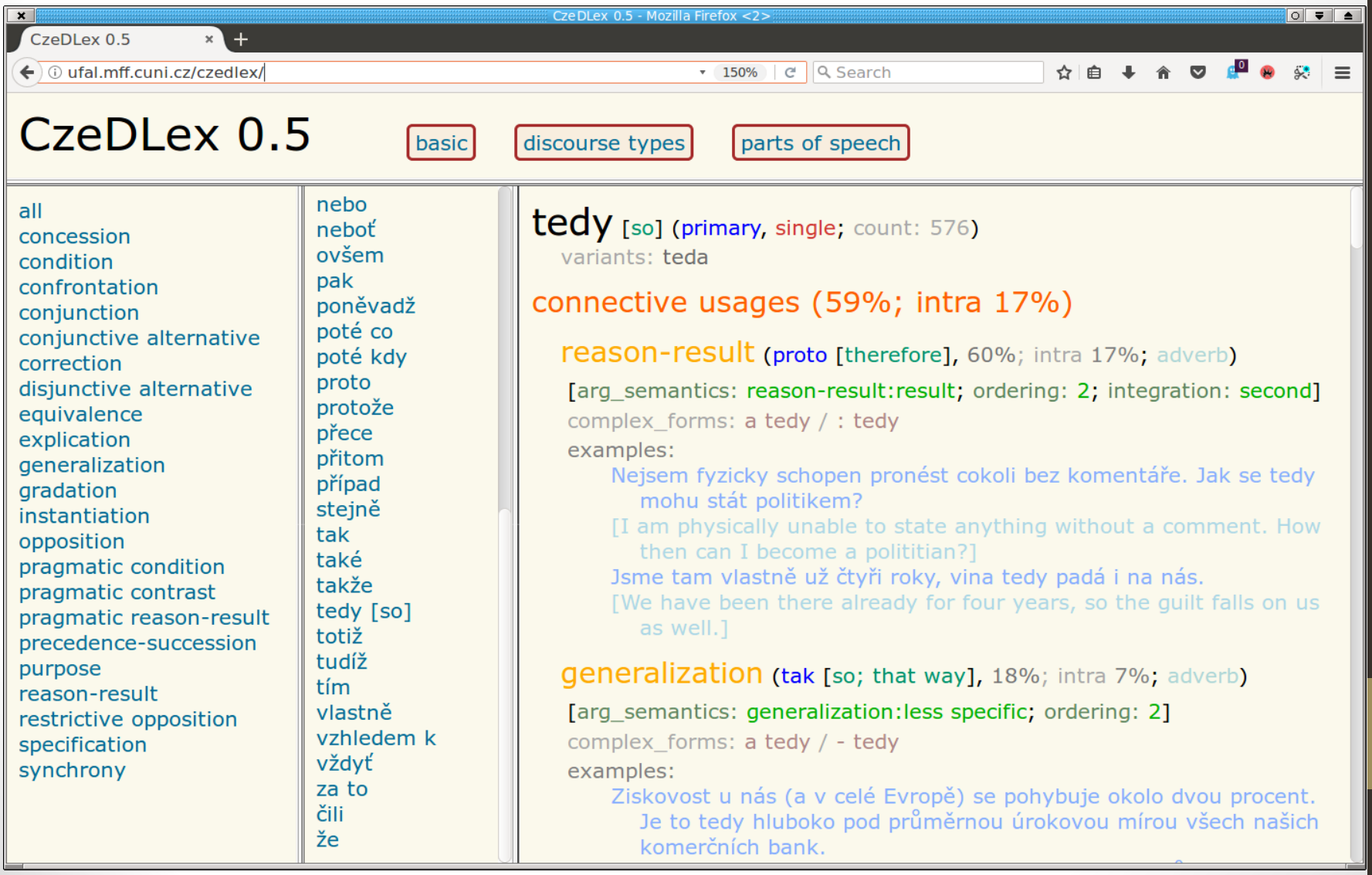
CzeDLex online

- Version 0.5 released in 2017, online

  http://ufal.mff.cuni.cz/czedlex

- 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 online



Selected References

- Antonio Briz, Salvador Pons Bordería, and José Portolés. 2003. Diccionario de partículas discursivas del español. Data/Software, www.dpde.es. Online since 2003.
- 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
- Feltracco Anna; Jezek Elisabetta; Magnini Bernardo; Stede Manfred. "LICO: A Lexicon of Italian Connectives" In: Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016), Napoli, December 5-7, 2016.
- Mendes, Amália and Pierre Lejeune (2016) "LDM-PT. A Portuguese Lexicon of Discourse Markers." In: Degand, Liesbeth, Csilla Dér, Péter Furkó, Bonnie Webber (eds.) Conference Handbook of TextLink – Structuring Discourse in Multilingual Europe Second Action Conference, Budapest, 11-14 April 2016, 89-92.
- Mírovský Jiří, Synková Pavlína, Rysová Magdaléna, Poláková Lucie: *CzeDLex – A Lexicon of Czech Discourse Connectives.* In: The Prague Bulletin of Mathematical Linguistics, No. 109, Czech Republic, pp. 61-91, Oct 2017
- Rashmi Prasad and Nikhil Dinesh and Alan Lee and Eleni Miltsakaki and Livio Robaldo and Aravind Joshi and Bonnie Webber. The Penn Discourse Treebank 2.0. Proceedings of LREC 2008, pp.2961–2968, Marrakech, Morocco.
- Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LEXCONN: A French Lexicon of Discourse Connectives. Discours [En ligne], 10|2012, http://discours.revues.org/8645.
- Tatjana Scheffler and 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 2016)*. European Language Resources Association (ELRA), Paris, France.
- Manfred Stede and Carla Umbach. 1998. DiMLex: A Lexicon of Discourse Markers for Text Generation and Understanding In: *Proceedings of the 17th International Conference on Computational linguistics.* pp. 151–184. Association for Computational Linguistics.
- Manfred Stede. 2002. DiMLex: A Lexical Approach to Discourse Markers In: *A. Lenci, V. Di Tomaso (eds.): Exploring the Lexicon - Theory and Computation.* Alessandria (Italy): Edizioni dell'Orso.
- Michal Škrabal, Martin Vavřín: The Translation Equivalents Database (Treq) as a Lexicographer's Aid. Electronic lexicography in the 21st century. *Proceedings of eLex 2017 conference*, Leiden, 2017

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!

Michal Škrabal, Martin Vavřín: The Translation Equivalents Database (Treq) as a Lexicographer's Aid. Electronic lexicography in the 21st century. *Proceedings of eLex 2017 conference*, Leiden, 2017

Treq

TRANSLATION
EQUIVALENTS
DATABASE

Ver. 2.0

Source language: French ⇄ Target language: English

Restrict to ?: Collection(s): 6

mais | Search

☐ Lemma ? ☐ Multiword ? ☐ RegEx ? ☐ A = a ?

▲ Frequency ▼	▲ Proportion ▼	▲ French ▼	▲ English ▼
75987	94.2	mais	but
2152	2.7	mais	and
415	0.5	mais	yet
260	0.3	mais	it
226	0.3	mais	well
204	0.3	mais	though
195	0.2	mais	although
156	0.2	mais	But
133	0.2	mais	only