Designing a corpus-based lexicon for spoken DRDs

SEMANTIC CONSIDERATIONS

Ludivine CRIBLE (Université catholique de Louvain)
Amalia MENDES (Universidade de Lisboa)

TextLink Final Action Conference, 19-21 March 2018, Toulouse

Overview of the talk

Introduction

- Variation of DRDs
- Existing DRD lexicons

Spoken corpus and functional classification

- Data
- Taxonomy in domains and functions

Proposal for the semantics of spoken DRDs

- Different types of polyfunctionality
- Next steps for the lexicon

Introduction

SYNTACTIC AND SEMANTIC VARIATION IN EXISTING LEXICONS

Variable terminology

- Written and spoken language is built upon relations of coherence
- These relations are often signalled by Discourse-Relational Devices (DRDs)
  - « connectives » : relation between two (or more) units, two-position markers (e.g. *because*)
  - « discourse markers » : not necessarily relational, one-position markers (e.g. *I mean*)

➢ « DRDs » as the umbrella term to cover formal and functional variability ?

Variable form

- Typically short and fixed expressions with a (primarily) procedural meaning
- Common core : conjunctions (*and*, *but*, *although*) and adverbials (*so*, *however*, *in fact*)
- Other categories : verb phrases (*I mean*), interjections (*oh*)
- Other devices : alternative lexicalizations (*It results that*), syntactic forms (gerund)

Variable function

- DRDs are highly polyfunctional as a category : cause, contrast, specification, topic…
- Individual DRDs can be quite polyfunctional/ambiguous too : e.g. *actually*, *so*, *and*
  - depends on degree of granularity in semantic distinctions

- Translation equivalents are not necessarily used in the same way across languages
- Challenging to teach, to acquire and to translate

➢ Need for DRD lexicons to be consulted or applied automatically

Building lexicons

- Automatically extract information from annotated discourse banks
  - the case of the English section of Connective-Lex (PDTB)
- Manually inspecting texts and grammars
  - the case of LEXCONN (French) and DIMLex (German)
- Automatically extract information + manual verification and additions
  - the case of the *Diccionario de partículas discursivas del español – DPDE* (Spanish)*, the* LDM-PT (Portuguese) and the CzeDLex (Czech)
- Most lexicons focus on written data : Czech, French, German, Italian, Portuguese
- Exceptions : the *DPDE and the Maschler Inventory of Hebrew Discourse Markers*

Encoding the polyfunctionality of DRDs in lexicons

- Different typologies to label the semantic relations expressed by DRDs :
  - LEXCONN → SDRT
  - DIMLex, LDM-PT → PDTB 3.0
  - DPDE → lexicographic definition

- Different solutions to encode polyfunctionality :
  - DIMLex → list of senses in a POS entry
  - LEXCONN, LDM-PT → individual entries of form-meaning pairs
  - DPDE : distinguishes between distinct uses (homonyms) and « other uses » (contextual senses)

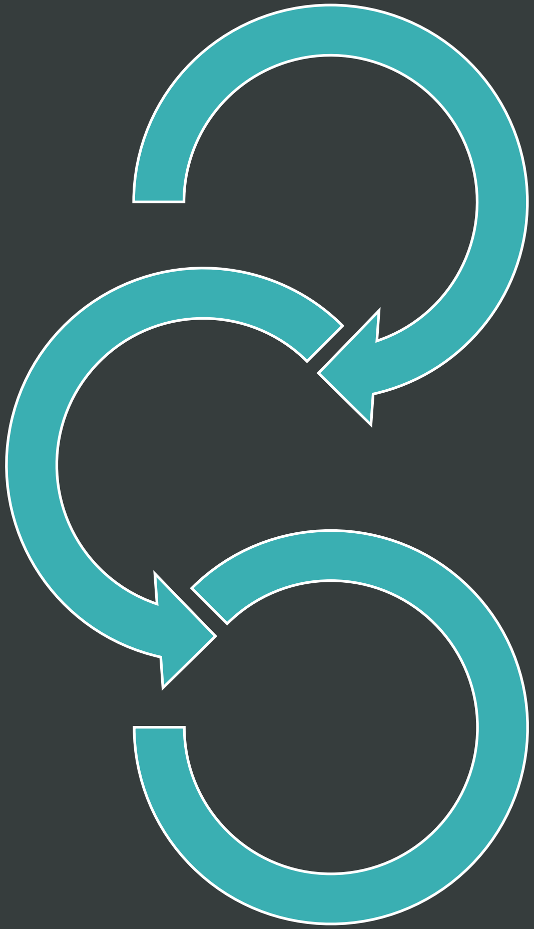
Our proposal

- First steps for the semantic structure of a lexicon of spoken DRDs
- Based on annotations in the DisFrEn dataset (English-French)

- Which semantic labels to use
- How to account for ambiguity and polysemy

Why turn to speech?

- Several DRDs and relations are shared across speech and writing
  - *and, so, but, because, actually, in fact, for example…*
- Some shared DRDs perform additional functions
  - *so* → exemplification, topic-resuming
- Some shared relations are expressed by additional markers
  - restatement → *well, I mean, you know*



The DisFrEn dataset

CORPUS DATA AND FUNCTIONAL TAXONOMY

English-French comparable dataset

- 80,000 words (abt 7.5 hours) in each language
- 8 spoken genres, such as conversation, interview, classroom lesson, news broadcast…
- Sampled from existing corpora, mainly *International Corpus of English* and VALIBEL

- Text-to-sound aligned, audio available during annotation
- Manually annotated under EXMARaLDA

(Crible 2017)

Identification of DRDs (discourse markers)

- Bottom-up (no closed list) and manual identification
- Three main criteria :
  - syntactic optionality
  - formal fixedness
  - procedural meaning
- 100+ DRD types in each language

actually; after; after all; albeit; alright; although; and; and so on; and still; and that kind of stuff; and then; and things; anyway; as; as it were;as long as; as soon as; because; before; but; but then; by the way; considering; either; etcetera; even if; even though; finally; first; first of all; for; for example; for instance; having said that; however; I don't know; I mean; I suppose; if; if you like; in addition; in fact; in other words; indeed; insofar ; as; instead; kind of; like; listen; look; meanwhile; nevertheless; no; now; oh; ok; okay; on the other hand; once; only; or; or something; otherwise; plus; provided; right; say; second; secondly; see; since; so; so that; sort of; then; therefore; though; till; unless; until; well; when; whenever; where; whereas; while; whilst; yeah; yes; yet; you know; you see

(Crible 2017)

Sense disambiguation (1)

- New taxonomy designed to reconcile models of discourse functions (speech) with discourse annotation schemes (writing)
- Two inter-dependent semantic-pragmatic layers :
  - domains (generic level) : for quantitative analysis and summarization of data
  - functions (specific level) : for descriptive accuracy
- Generic level mainly inspired by Redeker (1990), González (2005)
- Relational functions and guidelines inspired by the PDTB 2.0
- Additional functions inspired by González (2005), Cuenca (2013)

Sense disambiguation (2)

- 4 domains, 30 functions

Ideational	Rhetorical	Sequential	Interpersonal
cause	motivation	punctuation	monitoring
consequence	conclusion	opening boundary	face-saving
concession	opposition	closing boundary	disagreeing
contrast	specification	topic-resuming	agreeing
alternative	reformulation	topic-shifting	elliptical
condition	relevance	quoting	
temporal	emphasis	addition	
exception	comment	enumeration	
	approximation		

inter-annotator agreement κ = 0.406, 44.5%

intra-annotator agreement κ = 0.74, 75.8%

An example

BB1: could you talk a little bit about the Wirral accent I I know that um (0.200) there’s obviously quite a um range of accents in that part of the country

BB4: yeah (0.520) uh well I (0.290) consider myself to have a Cheshire accent because when I was born (0.300) and I lived in (0.110) on the Wirral (0.287) uh (0.333) i- (0.460) it was a Cheshire accent which is (0.440) the accent I have now though (0.270) there are overtones of (0.230) the Liverpudlian accent (0.290) however over the years certainly it has changed (0.270) and now it’s very much (0.110) a Liverpool accent (0.340) and uh you know which (0.430) I’m not (0.300) I’m not saying I disapprove of it but I think it’s a lazy speech and you need to (0.440) actually um (0.530) think about what you’re saying I know my nephew sometimes’ll to speak to me in the Liverpool accent (0.350) and I’ll say please speak to me in English <laughing/> (0.160) but it’s things like “yeah” and “you what” and (0.230) whereas you know mine is “yes” “pardon” or whatever <noise/> I’m a bit old-fashioned in that way so I do find the accent (0.440) is a bit harsh and it’s interesting that actually that accent is spread out into the (0.270) uh (0.390) the parts of north Wales that are very near to the Wirral…

(EN-intf-03)

An example

BB1: could you talk a little bit about the Wirral accent I I know that um (0.200) there’s obviously quite a um range of accents in that part of the country

BB4: yeah (0.520) uh well I (0.290) consider myself to have a Cheshire accent because when I was born (0.300) and I lived in (0.110) on the Wirral (0.287) uh (0.333) i- (0.460) it was a Cheshire accent which is (0.440) the accent I have now though (0.270) there are overtones of (0.230) the Liverpudlian accent (0.290) however over the years certainly it has changed (0.270) and now it’s very much (0.110) a Liverpool accent (0.340) and uh you know which (0.430) I’m not (0.300) I’m not saying I disapprove of it but I think it’s a lazy speech and you need to (0.440) actually um (0.530) think about what you’re saying I know my nephew sometimes’ll to speak to me in the Liverpool accent (0.350) and I’ll say please speak to me in English <laughing/> (0.160) but it’s things like “yeah” and “you what” and (0.230) whereas you know mine is “yes” “pardon” or whatever <noise/> I’m a bit old-fashioned in that way so I do find the accent (0.440) is a bit harsh and it’s interesting that actually that accent is spread out into the (0.270) uh (0.390) the parts of north Wales that are very near to the Wirral…

(EN-intf-03)

An example

BB1: could you talk a little bit about the Wirral accent I I know that um (0.200) there's obviously quite a um range of accents in that part of the country

BB4: yeah (0.520) uh well I (0.290) consider myself to have a Cheshire accent because when I was born (0.300) and I lived in (0.110) on the Wirral (0.287) uh (0.333) i- (0.460) it was a Cheshire accent which is (0.440) the accent I have now though (0.270) there are overtones of (0.230) the Liverpudlian accent (0.290) however over the years certainly it has changed (0.270) and now it's very much (0.110) a Liverpool accent (0.340) and uh you know which (0.430) I'm not (0.300) I'm not saying I disapprove of it but I think it's a lazy speech and you need to (0.440) actually um (0.530) think about what you're saying I know my nephew sometimes'll to speak to me in the Liverpool accent (0.350) and I'll say please speak to me in English <laughing/> (0.160) but it's things like "yeah" and "you what" and (0.230) whereas you know mine is "yes" "pardon" or whatever <noise/> I'm a bit old-fashioned in that way so I do find the accent (0.440) is a bit harsh and it's interesting that actually that accent is spread out into the (0.270) uh (0.390) the parts of north Wales that are very near to the Wirral…

(EN-intf-03)

An example

BB1: could you talk a little bit about the Wirral accent I I know that um (0.200) there’s obviously quite a um range of accents in that part of the country

BB4: yeah (0.520) uh well I (0.290) consider myself to have a Cheshire accent because when I was born (0.300) and I lived in (0.110) on the Wirral (0.287) uh (0.333) i- (0.460) it was a Cheshire accent which is (0.440) the accent I have now though (0.270) there are overtones of (0.230) the Liverpudlian accent (0.290) however over the years certainly it has changed (0.270) and now it’s very much (0.110) a Liverpool accent (0.340) and uh you know which (0.430) I’m not (0.300) I’m not saying I disapprove of it but I think it’s a lazy speech and you need to (0.440) actually um (0.530) think about what you’re saying I know my nephew sometimes’ll to speak to me in the Liverpool accent (0.350) and I’ll say please speak to me in English <laughing/> (0.160) but it’s things like “yeah” and “you what” and (0.230) whereas you know mine is “yes” “pardon” or whatever <noise/> I’m a bit old-fashioned in that way so I do find the accent (0.440) is a bit harsh and it’s interesting that actually that accent is spread out into the (0.270) uh (0.390) the parts of north Wales that are very near to the Wirral…

(EN-intf-03)

An example

BB1: could you talk a little bit about the Wirral accent I I know that um (0.200) there's obviously quite a um range of accents in that part of the country

BB4: yeah (0.520) uh well I (0.290) consider myself to have a Cheshire accent because when I was born (0.300) and I lived in (0.110) on the Wirral (0.287) uh (0.333) i- (0.460) it was a Cheshire accent which is (0.440) the accent I have now though (0.270) there are overtones of (0.230) the Liverpudlian accent (0.290) however over the years certainly it has changed (0.270) and now it's very much (0.110) a Liverpool accent (0.340) and uh you know which (0.430) I'm not (0.300) I'm not saying I disapprove of it but I think it's a lazy speech and you need to (0.440) actually um (0.530) think about what you're saying I know my nephew sometimes'll to speak to me in the Liverpool accent (0.350) and I'll say please speak to me in English <laughing/> (0.160) but it's things like "yeah" and "you what" and (0.230) whereas you know mine is "yes" "pardon" or whatever <noise/> I'm a bit old-fashioned in that way so I do find the accent (0.440) is a bit harsh and it's interesting that actually that accent is spread out into the (0.270) uh (0.390) the parts of north Wales that are very near to the Wirral…

(EN-intf-03)

An example

BB1: could you talk a little bit about the Wirral accent I I know that um (0.200) there’s obviously quite a um range of accents in that part of the country

BB4: yeah (0.520) uh well I (0.290) consider myself to have a Cheshire accent because when I was born (0.300) and I lived in (0.110) on the Wirral (0.287) uh (0.333) i- (0.460) it was a Cheshire accent which is (0.440) the accent I have now though (0.270) there are overtones of (0.230) the Liverpudlian accent (0.290) however over the years certainly it has changed (0.270) and now it’s very much (0.110) a Liverpool accent (0.340) and uh you know which (0.430) I’m not (0.300) I’m not saying I disapprove of it but I think it’s a lazy speech and you need to (0.440) actually um (0.530) think about what you’re saying I know my nephew sometimes’ll to speak to me in the Liverpool accent (0.350) and I’ll say please speak to me in English <laughing/> (0.160) but it’s things like “yeah” and “you what” and (0.230) whereas you know mine is “yes” “pardon” or whatever <noise/> I’m a bit old-fashioned in that way so I do find the accent (0.440) is a bit harsh and it’s interesting that actually that accent is spread out into the (0.270) uh (0.390) the parts of north Wales that are very near to the Wirral…

(EN-intf-03)

An example

BB1: could you talk a little bit about the Wirral accent I I know that um (0.200) there's obviously quite a um range of accents in that part of the country

BB4: yeah (0.520) uh well I (0.290) consider myself to have a Cheshire accent because when I was born (0.300) and I lived in (0.110) on the Wirral (0.287) uh (0.333) i- (0.460) it was a Cheshire accent which is (0.440) the accent I have now though (0.270) there are overtones of (0.230) the Liverpudlian accent (0.290) however over the years certainly it has changed (0.270) and now it's very much (0.110) a Liverpool accent (0.340) and uh you know which (0.430) I'm not (0.300) I'm not saying I disapprove of it but I think it's a lazy speech and you need to (0.440) actually um (0.530) think about what you're saying I know my nephew sometimes'll to speak to me in the Liverpool accent (0.350) and I'll say please speak to me in English <laughing/> (0.160) but it's things like "yeah" and "you what" and (0.230) whereas you know mine is "yes" "pardon" or whatever <noise/> I'm a bit old-fashioned in that way so I do find the accent (0.440) is a bit harsh and it's interesting that actually that accent is spread out into the (0.270) uh (0.390) the parts of north Wales that are very near to the Wirral…

(EN-intf-03)

From meaning to function

- Not only the encoded semantics but also contextually enriched interpretations
- Includes functions related to the management of speech turns, topics or relationships
- Allows double labels for simultaneous functions

➢ Not only what they mean but also what they do

From corpus to lexicon

- Number and types of semantic labels for each DRD can be very high
  - 19 different (combination of) labels for 429 occurrences of *so* : CONS, CCL, SPE, TS, REFOR…
  - only two labels (result, reason) in the PDTB 2.0

➢ Not only due to data type but also a difference in the coverage of the taxonomy

- Such rich information cannot be directly implemented in the lexicon
  - needs to be filtered or structured
  - needs to be reduced so that it can be useful for the lexicon user

➢ Distinguish between different types of polyfunctionality

Our semantic proposal

POLYFUNCTIONALITY AND FUTURE STEPS

Ambiguity is ambiguous

Different types of polyfunctionality concern different DRDs or the same DRD :

1. Polysemy
2. Multifunctionality
3. Underspecification
4. Multidimensionality

- Ambiguity corresponds to homonymy (Tuggy 1993, e.g. *banks*) : not relevant for DMs

(Crible forthc.)

Polysemy

- Single lexeme with clearly distinct yet related meanings (Lyons 1977)
- A DRD encodes more than one meaning
  - *but* = contrast, concession
  - *so* = consequence, specification

➢ The lexicon should reflect all of these meanings

Multifunctionality

- Two or more simultaneous functions in a given context
    - e.g. temporal + consequence
- Annotation instructions often limit to one or two labels
- Multifunctionality can be easily extracted from *DisFrEn* (double labels)

➢ Double labels are not practical for lexicons

   ➢ choose the more prominent sense, if any

Underspecification

- The DRD expresses a meaning that is richer, more specific than its basic meaning
- Mostly applies to *and* (most frequent DRD in written and spoken English)
  - only encodes addition, not polysemous
  - can be used in contexts with enriched interpretations of consequence, concession, contrast…
  - only 57% of all *and* tokens express addition in *DisFrEn* (1140 total)
  - 91% of all *and* tokens express addition in the PDTB (3000 total) + list (7%), result (1%), juxtap. (0.4%)
- Underspecified labels of *and* can be easily extracted from *DisFrEn*

➢ Either do not include in the lexicon (semantic spectrum only, not pragmatic functions)

➢ Or do not lose the information but distinguish underspecified uses from core meaning

Multidimensionality

- Applies to types, not tokens in context
- Some senses of the DRD belong to different domains (or dimensions)
- In *DisFrEn*, some labels have equivalents in other domains
  - contrast – opposition (ideational – rhetorical)
  - cause – motivation (ideational – rhetorical)
  - condition – relevance (ideational – rhetorical)
  - alternative – reformulation (ideational – rhetorical)
  - temporal – enumeration (ideational – sequential)
- These pairs are not formally identified in the corpus, simply listed as different labels

Multidimensionality : independent layers

- Inspired by Crible & Degand’s (in press) revision of the taxonomy
- From 30 to 11 functions :

<table>
<tr><th>Ideational</th><th>Rhetorical</th><th>Sequential</th><th>Interpersonal</th></tr>
<tr><td colspan="4">[addition] [alternative] [cause] [condition] [consequence] [contrast] [opening] [punctuation] [specification] [temporal] [topic]</td></tr>
</table>

- Assumption : any function in any domain
- One core meaning (or more if polysemous) expressed in several domains

➢ The lexicon only includes the core meaning(s) and specifies possible domains

Multidimensional contrast with *mais 'but'*

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*

[ideational contrast]

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 le cours*

[rhetorical contrast]

L2 *euh j'aime les néologismes j'aime les les régionalismes mais euh je mets le point d'exclamation dessus euh pour dire euh attention*

L1 ***mais*** *la norme qu'est-ce qu'est-elle pour vous*

[sequential contrast]

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

[interpersonal contrast]

From annotations to lexicon entries

- Current annotations in *DisFrEn* do not allow to distinguish between polysemy, underspecification and multidimensionality
- We need to decide
  - whether other DRDs besides *and* can be underspecified (*actually ? I mean ?*)
  - whether we want to include the enriched interpretations of underspecified DRDs
  - ✓ which labels are multidimensional equivalents
  - whether everything else is polysemous

➢ Reduce the polyfunctionality of DRDs in the lexicon

➢ Maintain a large coverage of their functional spectrum in speech (and writing)

Possible semantic structure

Entry	Core meaning	Domains of use	Underspecified uses
AND	addition	ideational, rhetorical, sequential	consequence, contrast, specification, topic…
SO	consequence	ideational, rhetorical, sequential	NA
	specification	ideational, rhetorical	NA
BUT	contrast	ideational, rhetorical, sequential, interpersonal	NA
WHEREAS	contrast	ideational	NA

- Corpus annotations from *DisFrEn* not directly applicable
- Requires some top-down semantic decisions

Conclusions

- Semantic framework necessary to structure DRDs polyfunctionality
  - in particular, to formalize the entries in the lexicon
  - in general, to revisit classifications and semantic-pragmatic descriptions
- Building a corpus-based lexicon is complex
  - *DisFrEn* was not specifically designed for lexicographic applications
  - However it offers a broader and more flexible view of the functional spectrum of DRDs
  - ➢ Importance of the purpose and research question behind any annotation endeavor

- Work in progress!

Thank you for your attention

ludivine.crible@uclouvain.be
amaliamendes@letras.ulisboa.pt