

Annotation proposal for multilingual and multigenre spoken DRDs

Ludivine CRIBLE

TextLink Workshop on « Annotation of Discourse Relational Devices: Multilingual and multimodal challenges »

Madrid, 12-14 June 2017

Overview

1. Compiling multilingual, multigenre spoken corpora: the case of DisFrEn
2. Annotating DRDs in multilingual, multigenre spoken corpora
3. Functional taxonomy for spoken DRDs: a proposal and its revision
4. Results
5. Conclusion
6. Tool and technical format (tomorrow)

Compiling multilingual, multigenre spoken corpora

The case of DisFrEn

Corpus typology

○ Granger (2010)

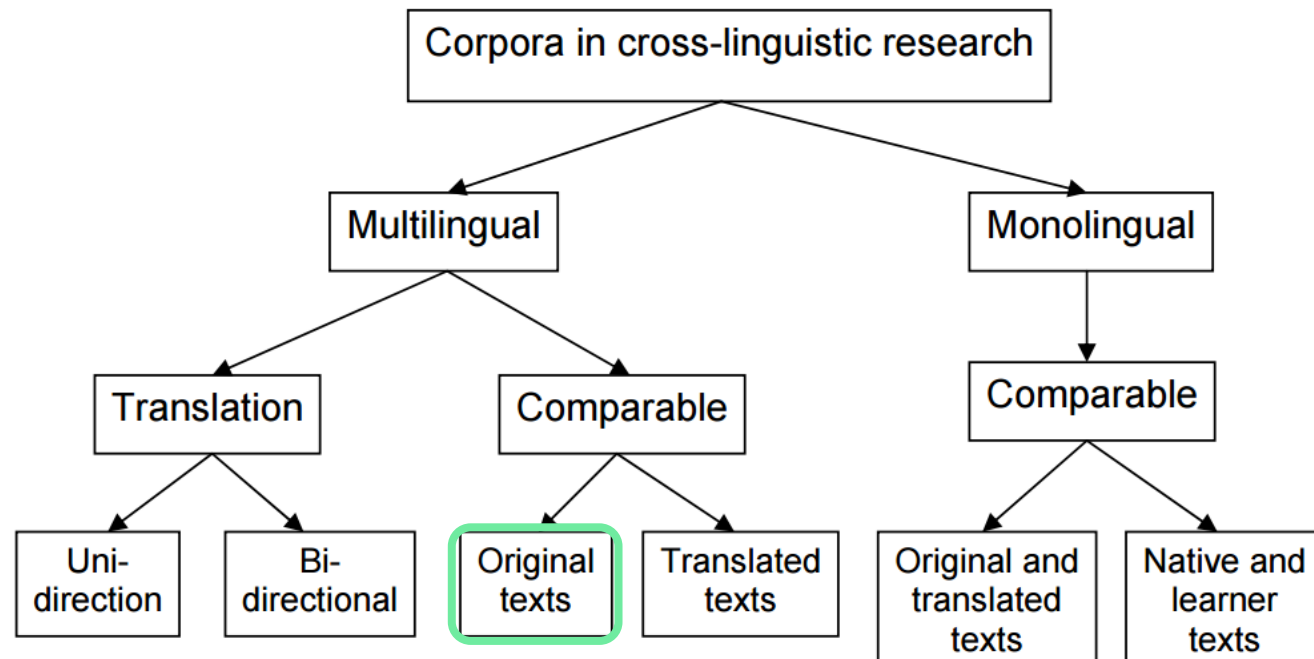


Figure 1: Corpora in cross-linguistic research

Dimensions of corpus comparability

Genre(s)

- same interaction types: conversation, interview...
- same parameters: dialogue vs. monologue, planned vs. unplanned...

Speakers

- sociolinguistic variables

Data quantity (words + minutes)

Date of the recordings

Audio availability and quality

Technical format: transcription conventions, segmentation, alignment...

...

Corpus recycling or new data collection?



time-saving

more control

- recording
- transcribing
- aligning

- copyrights
- date
- genres
- speakers...

Multilingual spoken corpora: examples

New corpora:

- EUROM: elicited (experimental) parallel corpus
- Euronews: TV news broadcasts in 10 languages
 - TED Talks?: scripted monologues + translated transcriptions (parallel)

+ comparable

Recycled corpora:

- C-Oral-Rom: spontaneous speech in Romance languages
- GECCo: written and spoken (academic, interviews) in English and German
- DisFrEn: comparable design of 8 genres in spoken English and French

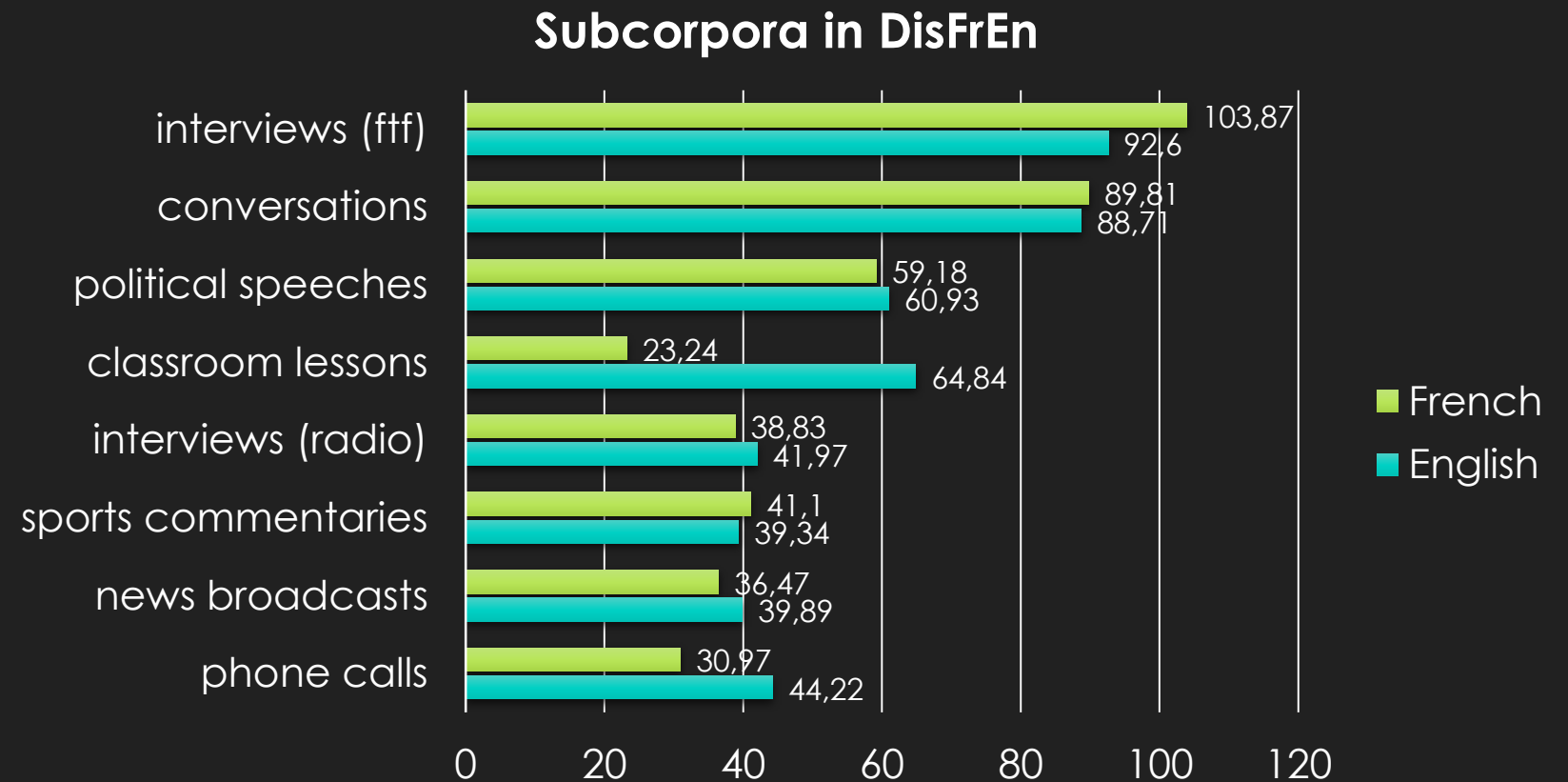
+ ecological

The DisFrEn corpus

- Recycled data
 - Bilingual English-French
 - Representative of many genres
 - Audio available
-
- dependent on available source corpora
 - English: ICE-GB
 - French: no reference corpus of spoken French!

Comparable design across 8 genres

- 15 hours
- 161.700 words
- 110 transcriptions
- Missing data



Annotating DRDs in multilingual, multigenre spoken corpora

Different challenges?

Annotation 101

1. Formulate research questions
2. Delimit the object of study
3. Design an annotation protocol
4. Cope with (un)expected problems
5. Cross fingers

Discourse relations vs. discourse markers

- Major divide in the field of discourse
- Roughly corresponds to written vs. spoken, but not necessarily
- Different challenges and benefits

Discourse relations

what counts as a relation
implicit and explicit
onomasiological
applied

mostly in written corpora

Discourse markers

what counts as a DM
relational and non-relational
semasiological
descriptive

mostly in spoken corpora

Marker-based approach

- Lower boundary: discourse vs. syntax (e.g. conjunctions)
- Higher boundary: discourse vs. other pragmatic categories (e.g. fillers)
- Disagreements on functional spectrum
- Two options: closed-list vs. bottom-up selection

From definition to identification

- Theoretical (abstract) definitions do not help with token identification
 - « sequentially dependent elements which bracket units of talk » (Schiffrin 1987: 31)
- Need for operational guidelines for DRD identification in authentic data
 - prescriptive criteria to be tested and revised
 - examples, tests, biases for borderline cases
 - every decision should be documented

Crible 2017a, Crible & Zufferey 2015

Tertium comparationis

- First step of any contrastive study

« make sure that we are comparing like with like »

(James 1980: 169)

- Common and neutral platform of comparison

- the case of bound morphemes in Turkish

- Requires rigorous definition of the object with formal and functional criteria

- Definition and functional spectrum are inter-related

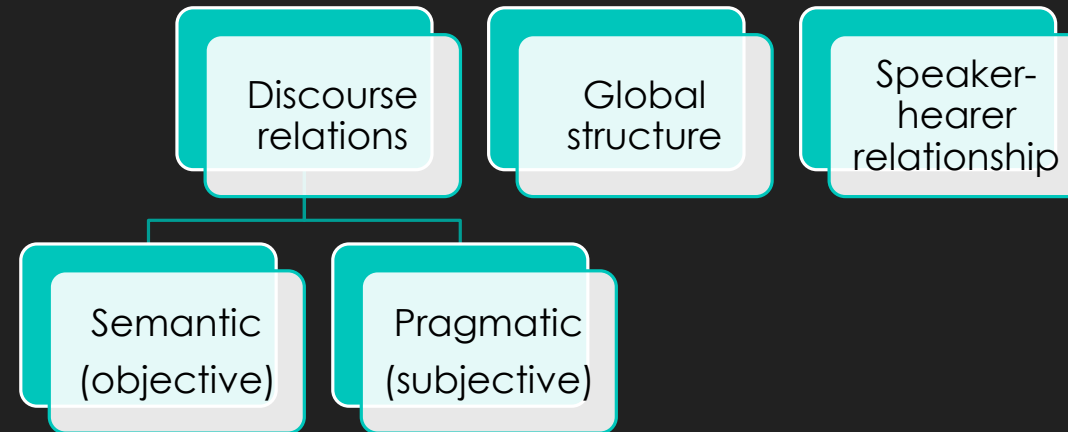
Multifunctionality of DRDs in speech

- the category covers items that perform many different functions
- a single member can perform different functions in different contexts
- a single member can perform different functions simultaneously in the same context

Cribble 2017

Tripartite functional models

- Halliday (1970): ideational, textual, interpersonal
- Redeker (1990): ideational, rhetorical, sequential
- Cuenca (2013): propositional, structural, modal
- ...



Discourse relations vs. topic relations

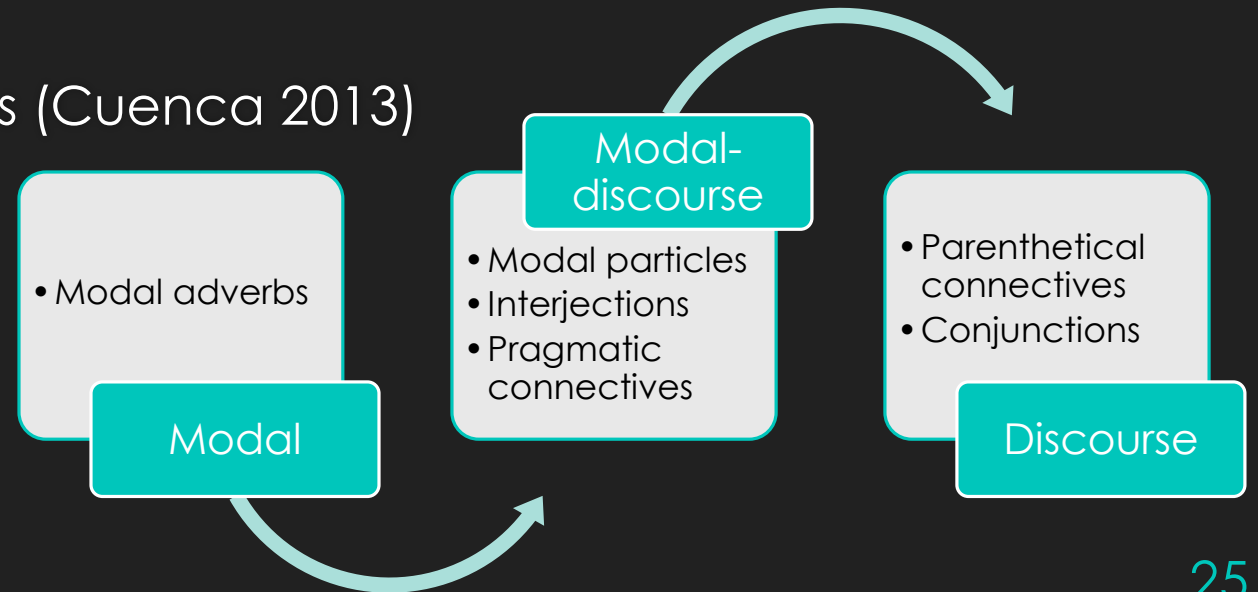
- No punctuation marks in speech
- No pre-defined structure in spontaneous speech: many changes of topic
- Different levels of coherence ? (Sanders et al. 1992: 2)
 - depends on your concept of what a discourse relation is
- Part and parcel of DRDs' functional spectrum
 - should be included in any marker-based approach

Intersubjectivity in dialogues

- Typically spoken « discourse markers » have intersubjective meanings
 - *you know, you see, look, listen, okay, right, and so on*
- Optional, procedural, fixed...
- Often combined with other functions (Cuenca 2013)

- Non-relational
- Different markers

➤ Borders of the category



Functions in speech and writing

- Common core of discourse relations, e.g. [cause] or [contrast]
 - 36,5% of all DMs in news vs. 20% in conversations in DisFrEn
- Structuring functions usually not included but exist in written-to-be-spoken
 - not necessarily the same structuring functions (global vs. local structure)
- Interpersonal functions are *a priori* absent from written texts and models
 - emerging interpersonal meanings of typical DRDs, e.g. French *mais* 'but'



Polysemy and underspecification

- Most DRDs are polysemous (e.g. *SO, BUT, ACTUALLY, WELL...*)
- Highly frequent multi-purpose DRDs: underspecified *AND*
elle avait un problème de boiler et j'ai été lui remettre un nouveau
'she had a boiler problem et 'and/so' I went to give her a new one'
- Different degree in planned vs. unplanned discourse?
 - higher type-token ratio of DMs in planned vs. unplanned speech in DisFrEn
 - *and* in DisFrEn vs. PDTB corpus: 11 vs. 4 function types; 60% [addition] vs. 91% [conjunction]
- Speaker-oriented strategy due to cognitive constraints of unplanned speech

Simultaneous multifunctionality

- Tendency of DRDs to perform more than one function (Petukhova & Bunt 2009)
- Exists in speech & writing
- Does not necessarily involves the same types of functions
 - writing: two discourse relations (e.g. conjunction-synchrony)
 - speech: discourse relation + structuring function (e.g. conclusion-topic resuming),

« The phenomenon of multifunctionality can be explained by considering participation in a dialogue as involving multiple activities at the same time, such as making progress in a given task or activity; monitoring attention and understanding; taking turns; managing time; and so on »

(Bunt 2012: 243)

Interim summary

- Multifunctionality in discourse exists in speech and writing but in different degrees
- More functions are needed to account for the full spectrum of DRDs
- Higher polysemy/underspecification of DRDs in speech
- Not only speech vs. writing but also (mostly?) planned vs. unplanned
- Definition and taxonomy depend on the research focus (relations vs. markers)

3 problems to expect

- Co-occurrence (boundary between co-occurring and complex DMs)
 - Underspecification in parallel vs. comparable corpora
 - Noise:
 - disfluencies
 - lack of context
 - missing Arg1 and/or Arg2...
- Topics for the round-table discussion!

Functional taxonomy for spoken DRDs

A proposal and its revision

Scope and goals

- Marker-based, bottom-up approach to relational & non-relational uses
 - large definition
 - large functional spectrum
- Reconcile models of discourse functions with discourse annotation schemes
 - generic level (domains): for quantitative analysis and summarization of data
 - specific level (functions): for descriptive accuracy

Main inspirations

- Generic level < Redeker 1990, González 2005
- Relational functions and format of the guidelines < PDTB 2.0 (Prasad et al. 2008)
- Other (speech-specific?) functions < González 2005, Cuenca 2013



Ideational

Rhetorical

Sequential

Interpersonal

Functions: ideational domain

Relations between real-world events

- cause – *because*
- consequence – *so, as a consequence*
- contrast – *but, by contrast, however*
- concession – *although, while*
- condition – *if*
- alternative – *or*
- temporal order – *when*
- exception – *unless*

Functions: rhetorical domain

Relations between epistemic and speech-act events; metadiscursive functions

- motivation – *because, since*
- conclusion – *so, as a conclusion, in sum*
- opposition – *but*
- relevance – *if*
- reformulation – *or, I mean, well*
- approximation – *sort of*
- comment – *actually*
- specification – *and, I mean, for example*
- emphasis – *actually, indeed*

Functions: sequential domain

Structuring of discourse segments

- turn-taking - *well*
- closing boundary – *so, FR voilà*
- topic-shifting – *so, now, by the way*
- topic-resuming - *anyway*
- quoting – *well, oh*
- enumerating – *and then, first*
- punctuating – *I mean, FR bon*
- addition - *and*

Functions: interpersonal domain

Interactive management of the speaker-hearer relationship

- monitoring – *you know, right*
- face-saving – *look, kind of*
- agreeing – *yeah, right*
- disagreeing – *well, FR ben*
- ellipsis – *and so on, etcetera*

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...



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...

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...

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...

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...

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...

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...

Reliability of the taxonomy

- Domains and functions are inter-dependent
- 30 labels, incl. very rare ones

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
 $\kappa = 0.406, 44.5\%$

intra-annotator agreement
 $\kappa = 0.74, 75.8\%$

- expertise
- training
- discussion

Less granularity, more reliability

- Revision of the taxonomy for **methodological** and **theoretical** reasons
 - fewer labels, more reliability
 - independence of the two annotation levels
- Assumption: any function can be expressed in any domain
 - extend the semantic-pragmatic divide to the other two domains
 - e.g. contrast can be ideational, rhetorical, sequential, interpersonal

(Crible & Degand in press)

Multifunctional *mais* 'but': cross-domain contrast

*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]

Independence of domains and functions

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

- Number and types of functions under testing
- Need to deal with domain-specific functions → sequential
 - Bunt (2011): « general-purpose » vs. « dimension-specific » functions

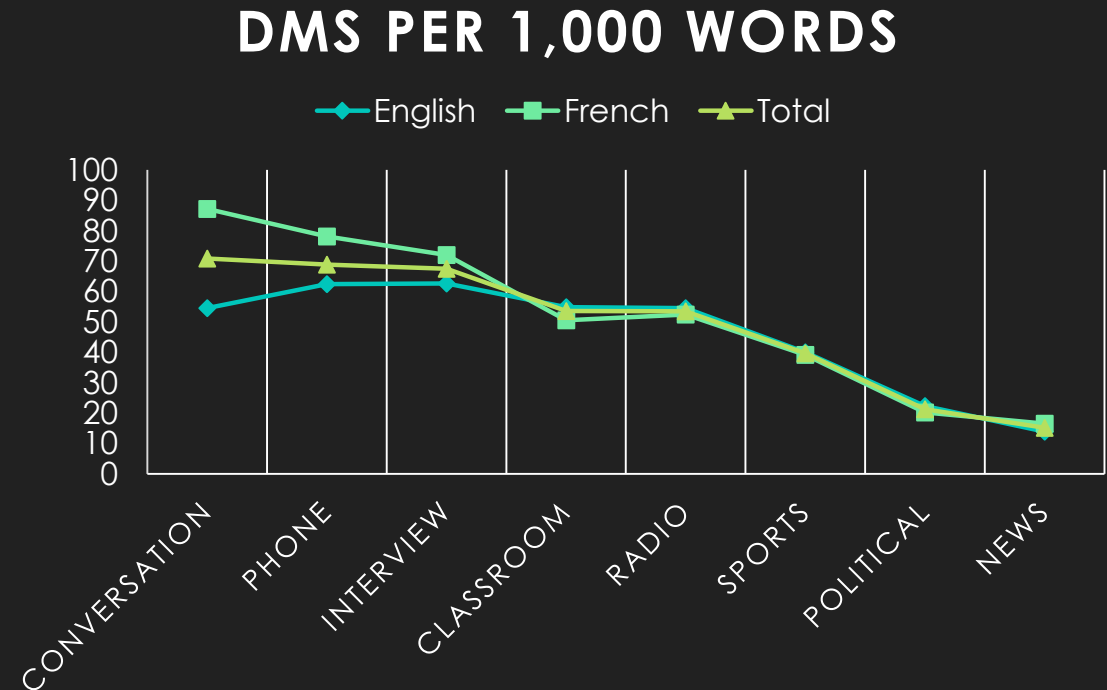
Results

Distribution across languages and genres

Distribution and variation of DMs

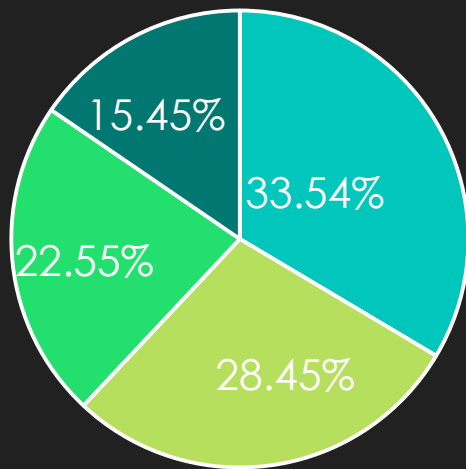
8,743 annotated DMs overall, 54 DMs ptw on average

- Most frequent DMs overall:
 - *and, well, but, so, you know*
 - *et, mais, donc, alors, hein*

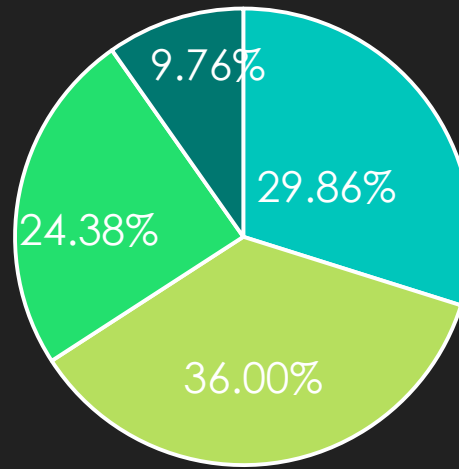


Domains from spontaneous to prepared

spontaneous

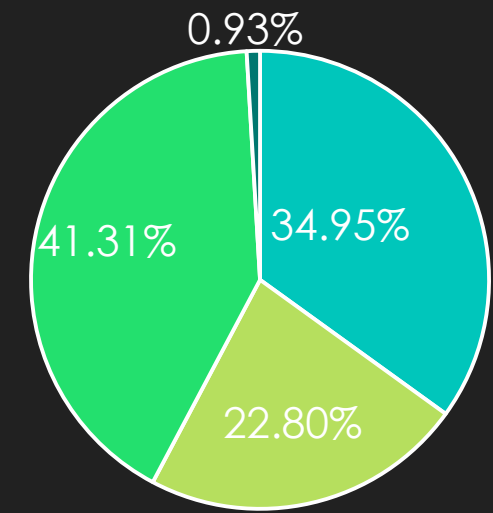


semi-prepared



Sequential
Rhetorical
Ideational
Interpersonal

prepared



5 most frequent functions by genre

	1 st	2 nd	3 rd	4 th	5 th
conversation	<i>addition</i>	<i>opening</i>	<i>opposition</i>	<i>monitoring</i>	<i>conclusion</i>
phone	<i>opening</i>	<i>addition</i>	<i>opposition</i>	<i>monitoring</i>	<i>conclusion</i>
interview	<i>specification</i>	<i>addition</i>	<i>consequence</i>	<i>conclusion</i>	<i>monitoring</i>
radio	<i>addition</i>	<i>specification</i>	<i>opposition</i>	<i>temporal</i>	<i>motivation</i>
classroom	<i>addition</i>	<i>conclusion</i>	<i>temporal</i>	<i>monitoring</i>	<i>opposition</i>
sports	<i>addition</i>	<i>consequence</i>	<i>temporal</i>	<i>concession</i>	<i>opposition</i>
political	<i>addition</i>	<i>temporal</i>	<i>concession</i>	<i>opposition</i>	<i>condition</i>
news	<i>addition</i>	<i>concession</i>	<i>temporal</i>	<i>opposition</i>	<i>topic-shift</i>

Four major form-function patterns

- coord. conj. in PRE-FIELD marking discourse structure (*and*)
 - no register restrictions
- subord. conj. in LEFT and RIGHT signaling discourse relations (*although*)
 - typical of formal registers
- adverbs in MEDIAL expressing meta-comments (*actually*)
 - no register restrictions
- interjections in INDEPENDENT serving interactional purposes (*okay*)
 - typical of dialogues

Conclusion

Summary and outlook

Challenges of multilingual spoken corpora

- Availability and technical treatment of source data
 - Definition of the *tertium comparationis* within lower and higher boundary
 - Formal and functional differences between planned and unplanned discourse
 - greater multifunctionality and ambiguity of DRDs in unplanned discourse
 - « Unclean » speech
- Every decision is research-specific and should be documented!

Outlook

- First spoken corpus fully annotated for DMs and their functions (in EN and FR)
- Taxonomy currently used in several languages, incl. the revised version
- Further data to explore: written, CMC, non-native
- Compatible with disfluency annotation and self-repair segmentation

Possible applications

- Language pathology, aging studies (Davis et al. 2013)
 - L1-L2 fluency, longitudinal studies (Denke 2009)
 - Processing studies with experimentation (Rohde & Horton 2014)
 - Transfer computational applications to the spoken mode
- Panel « Functions of pragmatic markers » at IPrA Belfast 2017!

Thank you for your attention!

Questions, feedback?

ludivine.crible@uclouvain.be