# DISCOURSE RELATIONAL DEVICES IN TEXTLINK: FROM (CATEGORIAL) DESCRIPTION TO CORPUS ANNOTATION, AND BACK AGAIN

Liesbeth Degand

University of Louvain, Belgium

Institute for Language & Communication

liesbeth.degand@uclouvain.be

# Outline

- Background: the TextLink COST Action
- Theoretical and Methodological Challenges
- Research Questions
  - Spoken Discourse Segmentation
  - DM Annotation
    - Position
    - Function
- Conclusions

# TextLink: Structuring discourse in multilingual Europe

- COST Action
  - European framework that supports trans-national cooperation among researchers, engineers and scholars across Europe.
  - Half way (from April 2014 to April 2018)
- Aims
  - unify numerous but scattered linguistic resources on discourse structure by (1) identifying and creating a portal into such resources including annotation tools, search tools, and discourse-annotated corpora; (2) delineating the dimensions and properties of discourse annotation across corpora; (3) organising these properties into a sharable taxonomy; (4) encouraging the use of this taxonomy in subsequent discourse annotation and in cross-lingual search and studies of devices that relate and structure discourse; and (5) promoting use of the portal, its resources and sharable taxonomy.

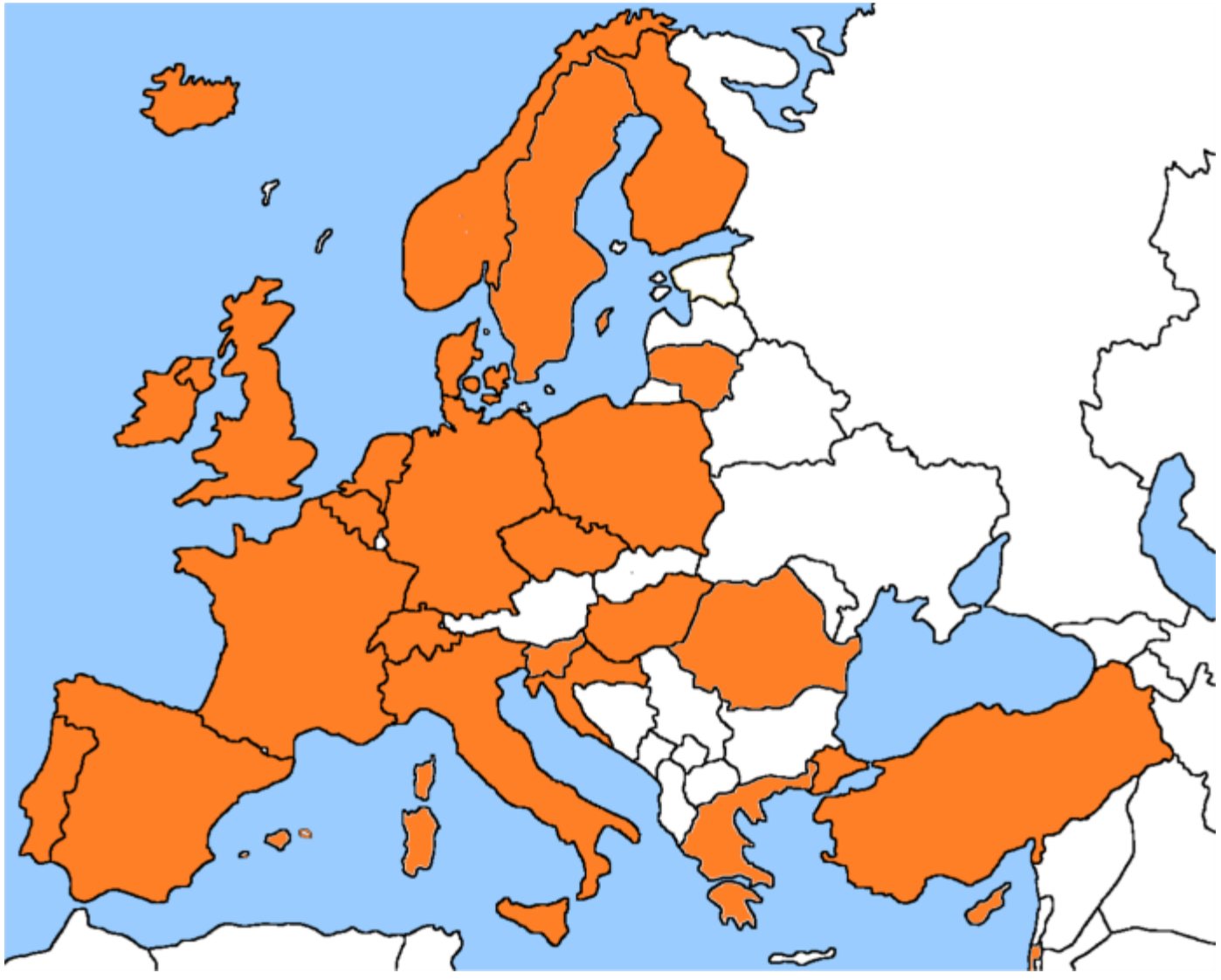
TextLink
Structuring Discourse in Multilingual Europe

# TextLink: Structuring discourse in multilingual Europe

Canada

Brazil

Argentina



# TextLink: Structuring discourse in multilingual Europe

- TextLink teams involved with over 20 languages
- Resources and/or linguistic analysis
  - Written
    - Czech, Dutch, English, Finnish, French, German, Greek, Italian, Lithuanian, Norwegian, Polish, (Brazilian) Portuguese, Romanian, Turkish,
  - Spoken
    - Catalan, English, French, German, Hebrew, Hungarian, Slovene, Spanish
  - Sign Language
    - French Belgian Sign Language, Catalan Sign Language

# TextLink: Sharing discourse annotated corpora and lexical resources

- Lexicons
  - French Lexicon of Discourse connectives (Roze et al. 2012)
  - German Discourse Marker Lexicon (Stede 2002)
  - GECCo List of conjunctive relations (English/German) (Lapshinova-Koltunski & Kunz 2014)
  - Inventory of Hebrew Discourse Markers (Maschler 2009)
  - Diccionario de partículas discursivas del español (Briz et al. 2003)
  - Portuguese Lexicon of Discourse Markers (under construction)
- Towards a portal of discourse annotated corpora
  - Imagine …

# TextLink Portal (to come!)

| | |
|---|---|
| Language | Czech, Dutch, English, Finnish, French, …. |
| Mode | Written, Spoken, Signed, …. |
| Text Type | Monologue, Dialogue, Newspaper, Political address, Spontaneous conversation …. …. |
| DRD | *alors, because, mais, want, pues, joten, ale* …. |
| Function | Ideational, Rhetorical, Sequential, Interactional,.. |
| Relation | Addition, Cause, Contrast, Topic Management, |

# TextLink Portal (to come!)

| | |
|---|---|
| Language | Czech, Dutch, English, Finnish, **French**, …. |
| Mode | Written, Spoken, Signed, …. |
| Text Type | Monologue, Dialogue, Newspaper, Political address, Spontaneous conversation …. …. |
| DRD | *alors, because, mais, want, pues, joten, ale* …. |
| Function | Ideational, Rhetorical, Sequential, Interactional,.. |
| Relation | Addition, Cause, Contrast, Topic Management, |

# TextLink Portal (to come!)

| | |
|---|---|
| Language | Czech, Dutch, English, Finnish, **French**, …. |
| Mode | Written, **Spoken**, Signed, …. |
| Text Type | Monologue, Dialogue, Newspaper, Political address, Spontaneous conversation …. …. |
| DRD | *alors, because, mais, want, pues, joten, ale* …. |
| Function | Ideational, Rhetorical, Sequential, Interactional,.. |
| Relation | Addition, Cause, Contrast, Topic Management, |

# TextLink Portal (to come!)

| | |
|---|---|
| Language | Czech, Dutch, English, Finnish, **French**, …. |
| Mode | Written, **Spoken**, Signed, …. |
| Text Type | **Monologue**, Dialogue, Newspaper, Political address, Spontaneous conversation …. …. |
| DRD | *alors, because, mais, want, pues, joten, ale* …. |
| Function | Ideational, Rhetorical, Sequential, Interactional,.. |
| Relation | Addition, Cause, Contrast, Topic Management, |

# TextLink Portal (to come!)

| | |
|---|---|
| Language | Czech, Dutch, English, Finnish, **French**, …. |
| Mode | Written, **Spoken**, Signed, …. |
| Text Type | **Monologue**, Dialogue, Newspaper, Political address, Spontaneous conversation …. …. |
| DRD | *alors, because, mais, want, pues, joten, ale* …. |
| Function | Ideational, Rhetorical, **Sequential**, Interactional,.. |
| Relation | Addition, Cause, Contrast, Topic, … |

# TextLink Portal (to come!)

[(nous en arrivons)SV <**ainsi**>md (à l'espace //C national et régional)SO]urv+ ///C
<**alors**>md //C <la //C euh représentation métrique>ag [(je vais passer)SV (assez rapidement)SRd (parce que c'est pas l'objet)SRd]urv <mais>md <pour clarifier les choses>ag ///S
<et>md //C <**en fait**>md //C [(vous corrigez)SV]urv //C [(vous les faites corriger)SV]urv ///C
<**mais**>md <la norme>ag //C [(qu')SO (est-ce)mi-I ///C (qu')SO-S (est-t-elle)SV //C (pour vous)SRd]urv ///T
<**et bien**>md //C [(il va //C introduire)SV ///C (dans sa musique)SRd ///C (le blues //C des des anciens esclaves de/ qu'il euh qui chantent //C dans les champs de coton)SO]urv ///C
<**de même**>md ///T [(il est rare)SV ///S (qu'un membre du corps scientifique ///S n'assure pas de charges d'enseignement)SO]urv ///T
<alors>md //C <**bon**>md <pour ce qui est //C de notre euh représentation de l'intonation>ag //C <en fait>md [(on reprend)SV (le terme de ///S profil mélodique)SO]urv //C <c'est-à-dire qu'>md [(on considère)SV (que tout

# TextLink Portal (to come!)

| | |
|---|---|
| Language | Czech, Dutch, English, Finnish, **French**, …. |
| Mode | Written, **Spoken**, Signed, …. |
| Text Type | Monologue, Dialogue, Newspaper, Political address, Spontaneous conversation …. …. |
| DRD | *alors, because, **mais**, want, pues, joten, ale* …. |
| Function | Ideational, Rhetorical, Sequential, Interactional,.. |
| Relation | Addition, Cause, Contrast, Topic, … |

# TextLink Portal (to come!)

<alors>md <cet auditeur //C vigilant>ag //C [(il va vous dire)SV]urv ///C <tiens>md euh [(encore jean d'ormesson)SN]ura //F <**mais**>md [(on entend)SV ///S (Jean d'Ormesson)SO //F (à chaque automne)SRd]urv ///T [(non non)SAdv]ura ///C <**mais**>md <la norme>ag //C [(qu')SO (est-ce)mi-I ///C (qu')SO-S (est-t-elle)SV //C (pour vous)SRd]urv ///T
[(c'est sûrement très grossier)SV]urv //C <d'ailleurs>md //T <**mais**>md [(vous savez)SV]urv //C [(quand on est écrivain //C et qu'on fait de la littérature)SRg //C (on est grossier)SV]urv //C
[(on est)SV (ce qu'on est)SO]urv //C <**mais**>md ///S [(c'est au moins //C essayer de se dire)SV //C (que ///S que on a une intention //C et que //S il y a //S une espèce //C de)SO //S (si on peut dire)insert //C (pensée //C derrière un livre)SO-S]urv+ ///T
<donc>md euh [(je je sais)SV //C (qu'il y a des prononciations réputées correctes //C d'autres incorrectes)SO]urv ///C euh<**mais**>md [(disons)SV]urv [(ça ne me tracasse pas)SV]urv //C
euh<ça>ag [(c'est //S c'est très remarqué)SV]urv ///C <**mais**>md ///S [(je me demande)SV //C (si //S on a conscience //C d'une norme ///C belge en

# TextLink Portal (to come!)

| | |
|---|---|
| Language | Czech, Dutch, **English**, Finnish, French, …. |
| Modality | Written, **Spoken**, Signed, …. |
| Text Type | **Monologue**, Dialogue, Newspaper, Political address, Spontaneous conversation …. …. |
| DRD | *alors, because, mais, want, pues, joten, ale* …. |
| Function | Ideational, Rhetorical, **Sequential**, Interactional,.. |
| Relation | Addition, Cause, Contrast, Topic, … |

# TextLink Portal (to come!)

the government comes under pressure to bring forward its plans for care in the community (0.360) **and** the impact on Israel of two hundred thousand soviet immigrants with a million more on their way (1.730)

I shall begin with a review of the economic situation and prospects (0.380) I shall **then** deal with monetary policy and public finances (0.220) **finally** I will present my tax proposals

England get an equaliser (0.220) England one soviet union one (0.380) well twenty seven goals Allan Smith has scored for arsenal this season or scored for arsenal during the season (0.647) but that in his sixth international is his first for England (0.380) **so** an equaliser equalising the own goal by mark wright

…

# TextLink Portal (to come!)

| | |
|---|---|
| Language | Czech, Dutch, **English**, Finnish, French, …. |
| Mode | Written, **Spoken**, Signed, …. |
| Text Type | Monologue, Dialogue, Newspaper, Political address, Spontaneous conversation …. …. |
| DRD | *alors, because, mais, want, pues, joten,* ***but*** . |
| Function | Ideational, Rhetorical, Sequential, Interactional,.. |
| Relation | Addition, Cause, Contrast, Topic, … |

# TextLink Portal (to come!)

that was good I wasn 't looking forward to doing it **but** I am now it 's going to be good (0.773)
you might not have got this (0.810) thing exactly right **but** you should have been able to do the rest
spk1: and you can tell the difference between those different accents
spk2: you can (0.090) Very much so (0.220) you know it 's uh (0.860) it 's it 's very interest- I don 't know how far out north Wales tends to now be thought of as part of (0.300) the Mersey (0.280) side uh (0.590) uh population because (0.400) people have gradually moved out and it 's become a computer belt (0.260) **but** taken the (0.170) accent with them (0.280) there 's less of a (0.020) a welsh accent (0.400) in there it 's more of the Liverpool ac (0.030) **but** I can certainly tell the difference between (0.420) somebody who 's truly Liverpudlian (0.330) and somebody who has a Cheshire accent
spk1: mm mm (0.290) Sherb
spk2: I know I 've been thinking of sherbet for a long time **but** <laughing/>
spk1: do you think…

# FROM CATEGORIAL DESCRIPTION TO DISCOURSE ANNOTATION

## Theoretical and Methodological challenges

# Challenges for categorial description

- It has become standard in any overview article or chapter on DMs to state that reaching agreement on what makes a DM is as good as impossible, be it alone on terminological matters (Degand, Cornillie, Pietrandrea, 2013: 5)
- “little consensus on whether they [DMs] are a syntactic or a pragmatic category, on which types of expressions the category includes, on the relationship of discourse markers to other posited categories such as connectives, interjections, modal particles, speaker-oriented sentence adverbials, and on the term “discourse marker” as opposed to alternatives such as “discourse connective” or “pragmatic marker” or “pragmatic particle” » (Lewis 2011, 419–20).

# Challenges for categorial description

- What is a Discourse Relational Device?
  - Linguistic expressions available in all the world's languages, including—but not restricted to—discourse markers and connectives, such as *because, but, however, I mean* or *well*, that help a speaker structure and organise their discourse
- Connective?
  - lexical items encoding a coherence relation between two abstract objects such as events, states or propositions (Asher, 1993, PDTB Research Group, 2007)
- Discourse Marker?
  - linguistic expression whose primary function lies at the discourse level, i.e. relating their host utterance to the discourse situation (…) contributing to the discourse organization (textual coherence), to the speaker/hearer interaction (interpersonal meanings), and/or to speaker attitudes (epistemic meaning)" (Degand 2014, 151)

# Challenges for categorial description

- What is a discourse (coherence) relation?
  - "A coherence relation is an aspect of the meaning of two or more discourse segments that cannot be described in terms of the meaning of the segments in isolation." The essential property of coherence relations is that they establish coherence in the cognitive representation people have or make of a discourse." (Sanders, Spooren & Noordman, 1992, 2)
  - "They are paratactic (coordinate) or hypotactic (subordinate) relations that hold across two or more text spans. When building a text or any instance of discourse (…) speakers choose among a set of alternatives that relate two portions of the text." (Taboada 2009, 125-126)

# A common ground

- The analysis of discourse markers is part of the more general analysis of discourse coherence—how speakers and hearers jointly integrate forms, meaning, and actions to make overall sense out of what is said. (Schiffrin, 1987:49)
- … the recognition of coherence relations by the hearer or reader enables them to assign coherence to a text. Discourse markers guide the text receiver in the recognition of those relations. (Taboada, 2006: 567-568)
- The problem of considering DMs to be the only type of signals is that DMs account for only a small fraction of relations present in a discourse, thereby leaving the majority of relations without DMs. This raises an obvious question: how are coherence relations signalled in the absence of DMs? (Das & Taboada, 2013)

# Criteria for the identification of DRDs

- Functional
  - Markers of an underlying discourse relation → any signal, including alternative lexicalizations (Das & Taboada, 2013; Prasad, Joshi, Webber 2010; Rysová, 2012; Taboada, 2009)
- Linguistic
  - Intra-sentential connectives form discourse segments that can be embedded under a matrix clause, Inter-sentential connectives form discourse segments that cannot be embedded under a matrix clause (Danlos et al. 2016)
- Feature-based (syntactic, semantic)
  - reliable, corpus-based selective criteria (semantic content, syntactic class, syntactic position), for the inclusion or exclusion of a particular item in the category dismissing “potential DMs” to retain only “confirmed DMS” (Bolly et al. 2015, in press, LPTS poster)

# Challenges for Discourse Annotation

- Onomasiological approach
  - Identification and Annotation of causal/contrastive/elaborative Discourse/Coherence relations
    - Explicit AND Implicit ?
  - Identification and Annotation of « all » DRDs
- Semasiological approach
  - Annotation of specific markers

# Challenges for Discourse Annotation

- Variety of taxonomies / Annotation schemes
  - Lexically grounded
    - Penn Discourse Treebank (Prasad et al. 2008, 2014)
  - Relation grounded
    - Rhetorical Structure Theory (Mann & Thompson 1988, Carlson et al. 2003)
    - Segmented Discourse Representation Theory (Asher & Lascarides 2003, Afantenos et al. 2012)
    - Cognitive approach to Coherence Relations (Sanders et al. 1992)
  - « Home-made » adaptions

Benamara & Taboada 2015

# Challenges for Discourse Annotation

- Different degrees of granularity
  - DRD-driven
  - Full discourse coverage
  - One-level vs. Multiple levels
  - Syntax, prosody, Morphology, …
- Cross-linguistic issues
  - Compatibility of annotation schemes (Benamara & Taboada 2015, Chiarcos 2014, Lapshinova et al. 2015, Sanders et al. LPTS2016)

# RESEARCH QUESTIONS

At the discourse level …

# Questions at the discourse level

- Discourse is a crucial notion for understanding human communication (Sanders & Spooren 2007)
- "Discourse is what makes us human, what allows us to communicate ideas, facts, and feelings across time and space." (Graesser, Millis, & Zwaan 1997: 164)
- Discourse phenomena to investigate:
  - The paradox of Discourse Markers
  - Spoken Discourse Segmentation

# The paradox of Discourse Markers

- Indices of fundamental cognitive processing during (spoken) language production, and genuinely constitutive of discourse/human communication.
- Sentence-level: optional, not a (morpho-)syntactic category
- Discourse-level: Communicatively obligatory (Diewald 2011) in ALL acts of human communication
  - written/spoken
  - formal/informal
  - CMC, even when space-constrained (texting, twitter, …)

# The paradox of Discourse Markers

→Human communication (in context; at the discourse level) is not possible without DMs.

→Study of DMs should learn us more about the underlying cognitive and functional principles of human communication.

# Spoken Discourse Segmentation

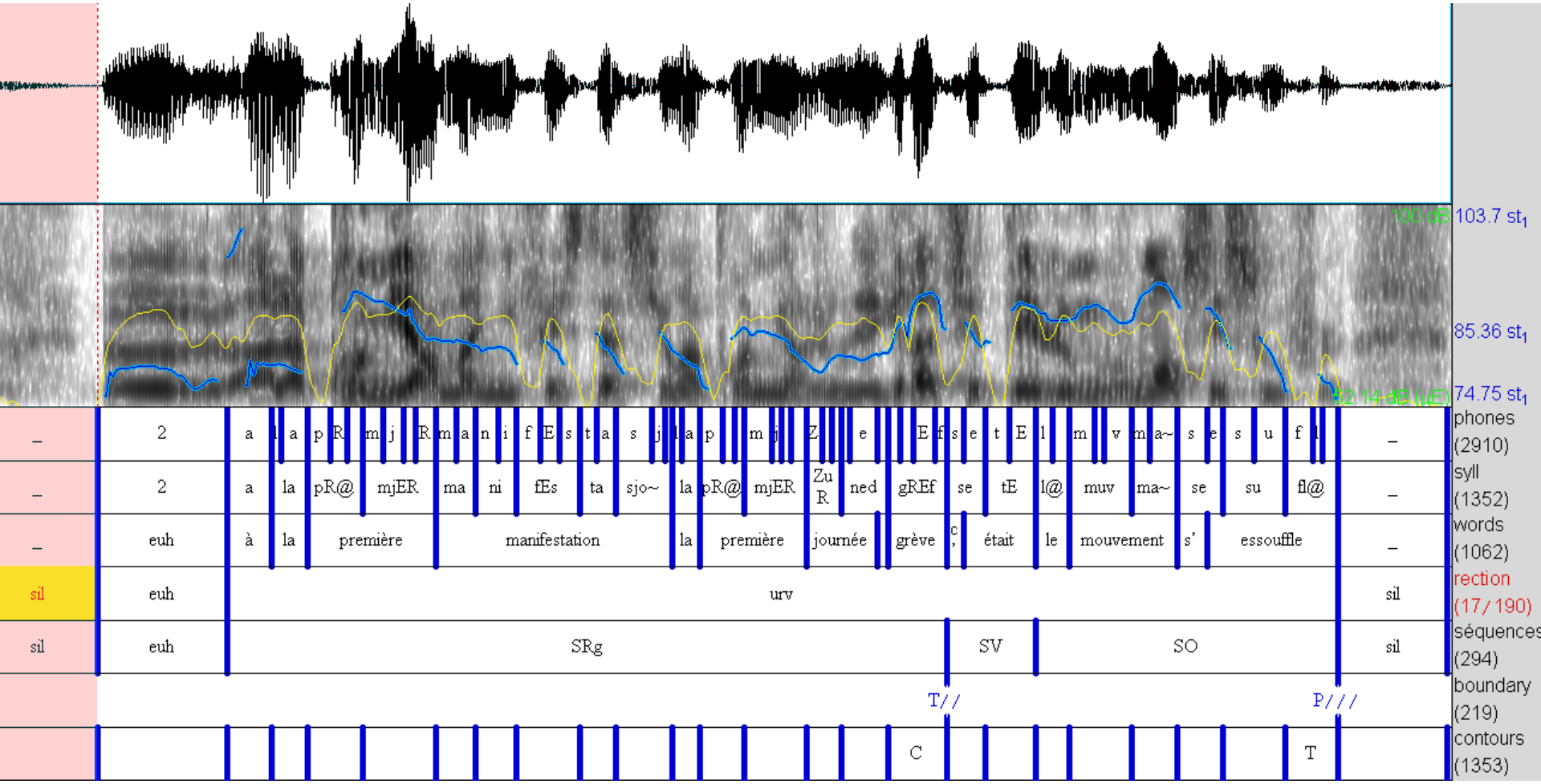
- "The stepwise structure of speech can give us many clues about cognitive processes. It suggests units for planning, production, perception and comprehension." (Holsánová 2008: 1)

→ Discourse segmentation is a crucial process in order to understand discourse production (planning) and discourse interpretation (coherent integration of spans within a structured, hierarchical schema).

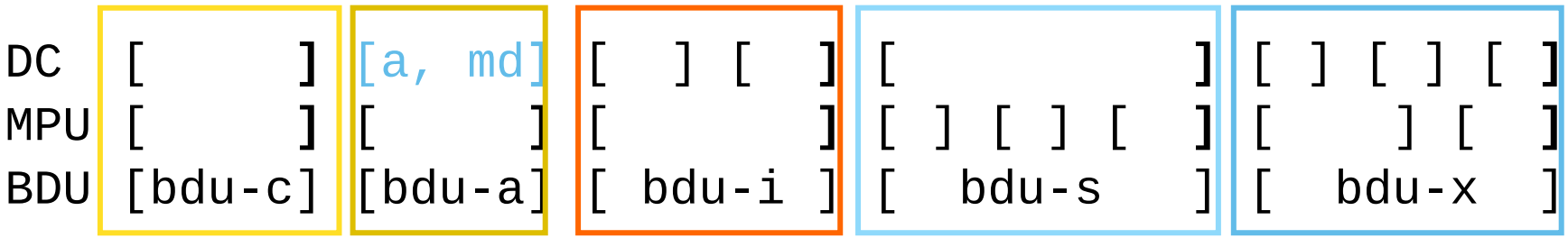
# Segmentation into Basic Discourse Units

- Mapping of two independently performed levels of analysis generates specific discourse organising level
  - **Syntactic analysis**: segmentation into dependency clauses (and functional sequences)
  - **Prosodic analysis**: segmentation into major prosodic units
- BDU **emerges** whenever a syntactic boudary coincides with a prosodic boundary. As long as the two levels do not coincide, the hearer awaits completion (Selting 2000)

# Syntactic and Prosodic Segmentation



# Types of BDU




(Degand & Simon 2009a)

- **Congruent BDU-c** : one dependency clause corresponds to one major prosodic unit
- **BDU-s grouped by syntax**: several major prosodic units correspond to one dependency clause
- **BDU-i grouped by intonation**: several dependency clauses correspond to one major prosodic unit
- **Adjunct BDU-a** (or regulatory): a non governed element is autonomized in a major prosodic unit
- **Mixed BDU-x** : sequence of mismatching syntactic and intonation units

# Types of BDUs: Examples

## Congruent BDU-c

- [(dans la majorité politique libanaise) (beaucoup de voix) (s'étaient élevées) (contre sa venue)]$^{urv}$ ///C

  in the lebanese political majority many voices raised against his coming ///

## Syntax-bound BDU-s

- [(pourquoi) (ce rôle majeur) (n'est-il pas dévolu) ///S (à la religion)]$^{urv+}$ ///C

  why isn't this major role devoted /// to religion

## Intonation-bound BDU-i

- euh [(on est devenu) (bien potes)]$^{urv}$ [(tout le monde) (se connaissait)]$^{urv}$ ///C

  uhm we've become good friends [urv] everybody knew one another [urv]

## Regulatory BDU-a

- <mais> <bon> ///T

  but well ///

# Functions of Basic Discourse Units

## Hypothesis

- BDUs are “the segments speakers and hearers rely on to construct and interpret the ongoing discourse, viz. segments on the basis of which inferential processes can take place.” (Degand & Simon 2009a: 82)

## Data-proof

- BDUs are discourse strategic units that vary with the communicative situation (Degand & Simon 2009ab, 2011; Simon & Degand 2011).
- Different genres call for different BDU distributions (1/3 congruent).

# Corpus LOCAS-F

- Louvain Corpus of Annotated Speech – French (collaboration with Anne Catherine Simon, Laurence Martin - Noalig Tanguy, Thomas Van Damme)
  - **12** « genres »
    - Communicative situations (degree of interactivity, degree of preparation, degree of broadcasting)
  - **42** audio samples (still growing)
    - 3 hours 11 minutes
    - ca. 15 min/genre
  - **62** speakers (**48** different speakers)
    - A mean of 4 speakers/genre (max. 11)
    - 12 speakers in more than one genre

Degand, Martin, Simon 2014

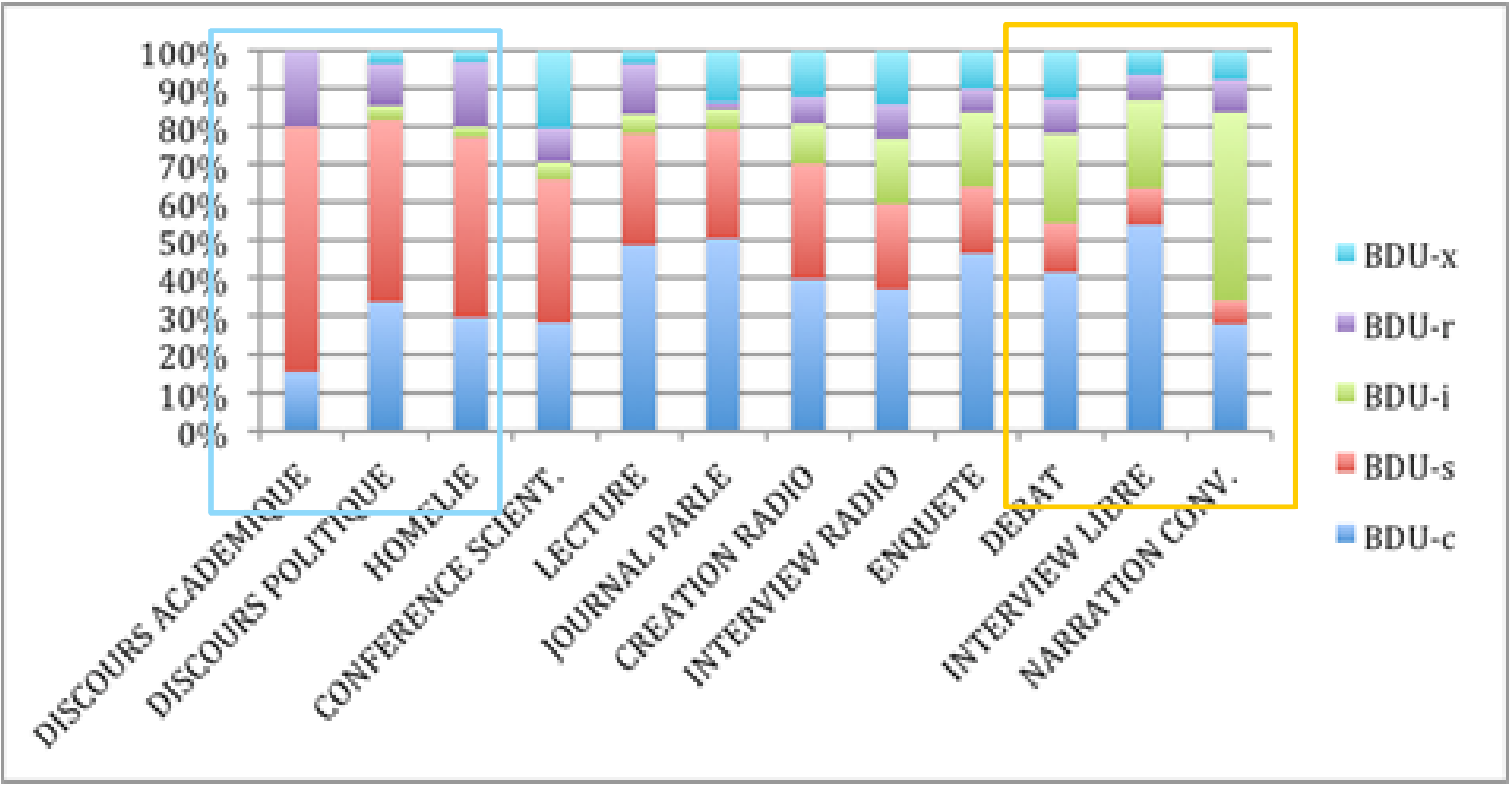
# LOCAS-F Situational features

| Genre | Duration | Interactivity | Preparation | Broadcast |
|---|---|---|---|---|
| Conference talk | 16:44 | 0 | 2 | 0 |
| Debate | 19:19 | 2 | 1 | 2 |
| Academic address | 15:18 | 0 | 2 | 1 |
| Political address | 20:26 | 0 | 2 | 2 |
| Homily | 13:23 | 0 | 2 | 1 |
| Survey | 16:04 | 1 | 1 | 0 |
| Free interview | 15:33 | 2 | 0 | 2 |
| Radio interview | 20:29 | 1 | 1 | 2 |
| Radio news | 14:45 | 0 | 2 | 2 |
| Reading | 15:18 | 0 | 2 | 0 |
| Free narration | 10:22 | 2 | 0 | 0 |
| Radio creation | 13:24 | 0 | 1 | 2 |
| Total | 3:11:05 | | | |

# LOCAS-F Situational features

| Genre | Duration | Interactivity | Preparation | Broadcast |
|---|---|---|---|---|
| Conference talk | 16:44 | 0 | 2 | 0 |
| Debate | 19:19 | 2 | 1 | 2 |
| Academic address | 15:18 | 0 | 2 | 1 |
| Political address | 20:26 | 0 | 2 | 2 |
| Homily | 13:23 | 0 | 2 | 1 |
| Survey | 16:04 | 1 | 1 | 0 |
| Free interview | 15:33 | 2 | 0 | 2 |
| Radio interview | 20:29 | 1 | 1 | 2 |
| Radio news | 14:45 | 0 | 2 | 2 |
| Reading | 15:18 | 0 | 2 | 0 |
| Free narration | 10:22 | 2 | 0 | 0 |
| Radio creation | 13:24 | 0 | 1 | 2 |
| Total | 3:11:05 | | | |

# LOCAS-F: BDU distribution




$\chi^2 = 602{,}84$ ; dl = 44 ; $p < 0{,}0001$

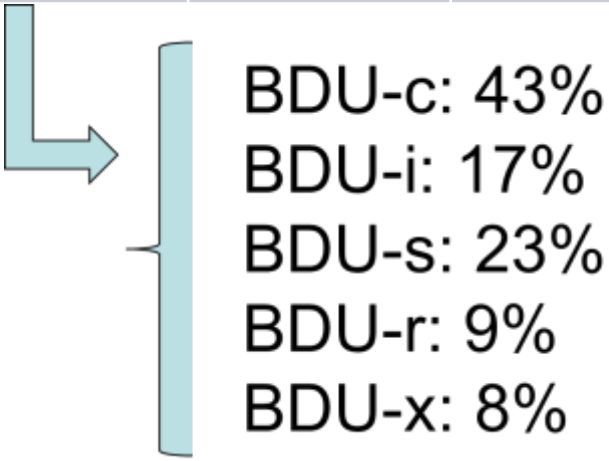
# LOCAS-F: BDU-distribution

- Neither the morpho-syntactic structure, nor the prosodic arrangement of spoken speech are sufficient conditions to define basic units in the discourse flow (cf. supra) →**comparison** of chi-square statistics
  - Distribution of BDU across genres
  - Distribution of (type of) syntactic clauses across genres
  - Distribution of intonation units (contours) across genres
- **Cramer's V** approx. 1,5 times higher for BDU distribution
- **BDU segmentation** is meaningful discourse segmentation

Martin, Degand, Simon 2014

# LOCAS-F in figures

| # | BDUs | Words | Length | Genres | Samples | DM |
|---|---|---|---|---|---|---|
| Total | 2875 | 41322 | 3:38 | 14 | 48 | 1780 |



# Discourse Markers in LOCAS-F

- **Claim**: the prosody-syntax interface gives rise to a distinctive discursive level of analysis contributing to the unfolding (linear) discourse.

→ Interaction between BDUs and other typical linguistic expressions working at the discourse level. A case in point are Discourse Markers

# Discourse Markers in LOCAS-F

- Linguistic expression whose primary function lies at the discourse level, i.e. relating their host utterance to the discourse situation.
- Can play a threefold role:
  - contributing to the discourse organization (textual coherence),
  - to the speaker/hearer interaction (interpersonal meanings),
  - and/or to speaker attitudes

# Discourse Markers in LOCAS-F

- Macro-syntactic constraints
  - It has to be syntactically detachable from a sentence (Schiffrin 1987)
  - They do not enter into the construction syntactically with other elements of the sentence (Sankoff et al. 2007)
  - [DMs are] “either outside the syntactic structure or loosely attached to it” (Brinton, 1996: 34).

→ **Weak clause association** (outside the dependency clause)

→ Identification of 1780 DM tokens (73 types)

# Variation in DM use

| | + interactive | - interactive |
|---|---|---|
| **- prepared** | 42 types<br>**649** tokens<br><br>*après, au fond, aussi, bien que, bon, en même temps, encore que, par contre, quoique, sinon* | 28 types<br>255 tokens<br><br>ensuite |
| **+ prepared** | 0 | 39 types<br>**314** tokens<br><br>*ainsi, au contraire, car, cependant, certes, de ce fait, de même, de plus, également, en effet, par ailleurs, par conséquent, pour autant, toutefois* |

# ANNOTATION OF DMS

Position

# DMs and (syntactic) position

- Initial position
  - It has to be commonly used in initial position of an utterance (Schiffrin 1987)
  - DMs “prototypically introduce the discourse segments they mark” (Hansen 1997)
  - “most items considered DMs are at least possible in initial position, and many occur there predominantly” (Schourup 1999)
  - Almost all DMs occur in initial position … , fewer occur in medial position and still fewer in final position (Fraser 1999)

# DM Position in LOCAS-F

- 1780 tokens (contiguous!)
- 73 types

| Position | Initial | Medial | Final | Isolated |
|---|---|---|---|---|
| BDU | 697 | 833 | 163 | 87 |
| Clause | 1345 | 88 | 179 | / |
| Intonation | 715 | 797 | 181 | 87 |

! Results of a script, partial manual correction

# DM types vs. Position in Clause

- Initial (left periphery)
  - 1345 tokens vs. 48 types
  - Top DMs: *et* (467), *mais* (278), *donc* (133), *alors* (79), *ben/bien* (77)
- Medial (parenthetical)
  - 88 tokens vs. 28 types
  - Top DMs: *quand même* (17), *donc* (14), *d'ailleurs* (8)
- Final (right periphery)
  - 179 tokens vs. 29 types
  - Top DMs: *quoi* (42), *hein* (38), *donc* (12), *voilà* (11)

# DM types vs. Position in Clause

- Exclusive uses (data + intuition)
  - LP: *ben, c'est-à-dire, car, c'est que, de même que, et, or, ou, puis*
  - RP: quoi
- Unexpected uses in LP
  - LP: *hein*

<les sentiments> [par moment ça me gonfle un peu] ///
<hein>// [autant de pages // pour savoir // si Titi va enculer Tata // comme dirait Céline]///

Those feelings, it sucks at times /// 'hein' (you know) // so many pages to know // whether X is going to f… Y // as Céline would say ///

# DM types vs. Position in Clause

- Unexpected uses in RP
  - RP: *mais*

L1:[je sais pas] // <moi>[je ne v/] [non] // [je di/ dis pas de chiffre] // <parce que j'en ai euh ///S une bonne dizaine en tête> // <mais>///C
L2: [ah] // <donc> <quand même> // [une bonne dizaine] ///

L1: I don't know I no I don don't say a figure because I have at least a dozen in my head <mais> ('but') ///

L2: ah so at least a dozen

[j'ai fait des vols] // <enfin> [j'ai volé dans les petit mag/ des petits bonbons dans les magasins] // [j'ai fait mes petits trucs] [c'est jamais des grands trucs euh de grand gangster] // <mais> ///T [j'ai fait mes petites conneries] // <quoi> ///T

I have stolen // well I have stolen in small sh/ sweets in shops // I have done my little things never big gangster things // <mais> I did my stupid things you see

# DM types between BDUs

- 87 tokens / 22 types
  - *alors, bon, et, mais, …*

<mais>$_{md}$ <avant>$_{ag}$ [il me semble qu'il serait bon ///C de réviser de rappeler les courants principaux du dix neuvième siècle]urv ///T <bon>$_{md}$ ///T [dans un premier temps ///C dans un premier temps ///S euh plantons je vais dire le décor ///C à la fois politique scientifique industriel ///C économique du dix neuvième siècle]urv+ ///T

but before this [it seems that it would be good to revise to recall the main currents of the nineteenth century] *bon* ('well; OK') [in the first place in the first place euh let's put I'd say the decor all in all political scientific industrial economical of the nineteenth century]

# A preliminary note on the prosody of DMs

- Fraser (1999: 933) DMs are “prosodically independent, being both accented and prosodically separated from their surrounding context by pauses, intonation breaks, or both”.
  - DMs do not systematically fulfill the assumption of prosodic independence. Rather prosodic independence results from syntactic and discursive constraints that need to be uncovered.
  - Correlate the prosodic characteristics of DMs with the other levels of annotation (BDU, syntactic clauses) and with the **function** these DMs fulfill in the discourse

# ANNOTATION OF DMS

## Functional domains

# Functional annotation of DMs

- What is more specific to DMs is their multifunctionality, which can be declined in three forms: (1) the category covers items that perform many different functions; (2) a single member can perform different functions in different contexts; and (3) a single member can perform different functions simultaneously in the same context, given the great polysemy of DMs. I have structured this multifunctionality into four functional “domains” (Sweetser 1990), inspired and revised mainly from Gonzalez (2005), Halliday and Hasan (1976), Redeker (1990) and Sweetser (1990). (Crible, in press).

# Functional annotation of DMs

- Ideational
  - linked to states of affairs in the world, semantic relations between real events. In other words, the relation between the two discourse objects exists independently in the real world.
- Rhetorical
  - linked to the speaker's metadiscursive work on the ongoing speech. This domain also includes pragmatic equivalents of certain ideational relations, when the relation is applied between two discursive events rather than world-events. These pragmatic relations apply to subjective claims, implicit assumptions or speech-acts. Unlike ideational relations, rhetorical functions cannot be reformulated without assigning mental states to one or both units

Crible, in press

# Functional annotation of DMs

- Sequential
  - linked to the structuring of discourse segments, both at macro- and micro-level. Sequential functions explicitly signal the progressing steps of speech and thought.
- Interpersonal
  - linked to the interactive management of the exchange, in other words to the speaker-hearer relationship. Interpersonal functions have a phatic function to call for attention or to manifest understanding.

Crible, in press

# Functional annotation of DMs

- Testing the annotation protocol in three rounds
- Double blind coding
- 3 x 100 random occurrences

| Round 1 | Round 2 | Round 3 |
|---|---|---|
| Kappa: .43<br>+ (abstract) coding scheme | Kappa: .59<br>+ bias<br>ideational > rhetorical > sequential > interpersonal | Kappa: .78<br>+ bias<br>+ DM cues/paraphr. |

Degand & Simon, 2015

# Functional annotation: Disagreements

| ACS Domains | LD ideational | rhetorical | sequential | interpersonal | Total |
|---|---|---|---|---|---|
| ideational | 41 | 11 | 12 | | 64 |
| rhetorical | 8 | 63 | 16 | 1 | 88 |
| sequential | 2 | 23 | 107 | 2 | 134 |
| interpersonal | | | | 12 | 12 |
| **Total** | **51** | **97** | **135** | **15** | **298** |

- Highest disagreement on rhetorical function (with ideational and sequential)
- Highest agreement on interpersonal function
- Bias as method (Spooren & Degand, 2010)

# Functional annotation of DMs: operationalisation

- **Ideational**: Objectifying domain
  - Relational: two explicit (adjacent) segments, order is (usually) relevant
  - Nature of the inference: Semantic underlying coherence relation (causal, temporal, conditional, …)
  - Nature of the DM: (adverbial) conjunction
  - Nature of the segments: states of affairs, no speaker involvement
  - Paraphrase: the situation in S1 (action, fact) [coherence relation] the situation in S2
- **Rhetorical**: Subjectifying domain, at the level of reasoning
  - Mainly relational
  - Inference: Pragmatic, speaker involvement, detour by speaker reasoning
  - DM: (adverbial) conjunction or adverb
  - Segments: At least one of the segments is opinion, point of view of speaker
  - Paraphrase: “according to me” (and I want you to believe this too)

# Functional annotation of DMs: operationalisation

- **Sequential**: At textual level
  - Relational or non-relational: Adjacency not required, often scope over larger span
  - Nature of the inference: Subjective, concerning the discourse structure itself
  - Nature of the DM: adverb (mainly); *bon, voilà, quoi, alors, …*
  - Nature of the segments: /
  - Paraphrase: explicitation of speech-act, “and now I add, say, reformulate, …”
- **Interpersonal**: At speaker/hearer level
  - Mainly non-relational
  - Inference: Subjective, speaker/hearer management
  - DM: phatic markers, parentheticals, *tu vois, hein*
  - Segments: /
  - Paraphrase: “I draw your attention to this aspect”, “I am listening”

# Functional annotation of DMs: operationalisation

| | ideational | rhetorical | sequential | interpersonal |
|---|---|---|---|---|
| **alors** | causal | JE en conclus que | et alors, de plus | |
| **c'est-à-dire** | | précision de la pensée, fonction de commentaire (Roulet) ou élaboration (PDTB/SDRT) | reformulation, efface une partie de la structure, la structure reformulée est celle qu'on retient | |
| **donc** | par conséquent | JE en conclus que<br>(ou précision de la pensée) | en d'autres mots, j'en reviens à (topic) | |
| **en fait** | (en réalité) | De mon point de vue, ce n'est pas ça (opposition)<br>OU précision de pensée | (reformulation complète), utilisé pour une prise de tour de parole (à propos) | |
| **et** | causal: et à cause de ça<br>temporel : et ensuite<br>conditionnel : si<br>contrastif : par contre<br>disjonctif : sinon, ou<br>additif : et aussi | sous-spécification d'une relation rhétorique : en fait, néanmoins, car | "et je vous annonce que", de plus, par ailleurs | |
| **mais** | par contre | néanmoins<br>« je ne sais pas mais je pense … » ; on s'oppose à un argument implicite, opposition à l'argumentation de l'autre<br>En début de TP : je tiens en compte ce que vous dites MAIS ne pensez-vous pas que… | Initiale du tp introduit | mais initial, discours indirect ou rapporte les pensées de l'autre |
| **puis** | et ensuite | | s'il y a des shifts de perspective (focalisation) | |

# Functional annotation: first results

| Domain | Frequency | Typical DM |
|---|---|---|
| Ideational | 82 (20%) | *et, puis* |
| Rhetorical | 116 (29%) | *mais, donc* |
| Sequential | 184 (46%) | *et, donc, alors, ben, quoi* |
| Interpersonal | 16 (4%) | *hein* |
| Total sample | 398 | |

# Functional annotation: first results

| | ideational | rhetorical | sequential | Inter-personal | Total |
|---|---|---|---|---|---|
| ***et*** | **45** | 5 | **69** | | 119 |
| ***mais*** | 7 | 41 | 19 | 1 | 68 |
| ***donc*** | 5 | 17 | 15 | | 37 |
| ***alors*** | 4 | 4 | 12 | 1 | 21 |
| ***ben*** | | | **15** | 2 | 17 |
| ***puis*** | **12** | | 3 | | 15 |
| ***en fait*** | | 10 | 5 | | 15 |
| ***enfin*** | | 4 | 10 | | 14 |
| ***quoi*** | | | **11** | | 11 |
| ***hein*** | | | | **11** | 11 |
| **Total** | 73 | 81 | 159 | 15 | 328/398 |

DM = 10 most frequent

# Functional annotation and position in BDU

| Domain | Position in BDU | | | | Total |
|---|---|---|---|---|---|
| | initial | medial | final | Isolated | |
| **ideational** | 35 | 45 | 1 | 1 | 82 |
| **rhetorical** | 30 | 70 | 7 | 9 | 116 |
| **sequential** | 80 | 77 | 14 | 13 | 184 |
| **interpersonal** | 2 | 9 | 4 | 1 | 16 |
| **Total** | 147 | 201 | 26 | 24 | 398 |

# Functional annotation and position in clause

| Domain | Position in clause | | | | Total |
|---|---|---|---|---|---|
| | LP | Medial | RP | / | |
| **ideational** | 76 | 2 | 1 | 3 | 82 |
| **rhetorical** | 80 | 22 | 2 | 12 | 116 |
| **sequential** | 141 | 3 | 20 | 20 | 184 |
| **interpersonal** | 5 | 1 | 8 | 2 | 16 |
| **Total** | 147 | 201 | 26 | 24 | 398 |

→ DMs play a relational and structuring role at the local level

# Functional annotation and situation

| | ideational | Inter-personal | rhetorical | sequential | Total |
|---|---|---|---|---|---|
| **interactive** | 27 | **9** | 38 | **81** | 155 |
| **non-prep** | 23 | 8 | 24 | 66 | 121 |
| **semi-prep** | 4 | 1 | 14 | 15 | 34 |
| **non-interactive** | **49** | 5 | **64** | **88** | 206 |
| **non-prep** | 17 | 2 | 6 | 26 | 51 |
| **prep** | 29 | 2 | 53 | 58 | 142 |
| **semi-prep** | 3 | 1 | 5 | 4 | 13 |
| **semi-interactive** | 6 | 2 | 14 | 15 | 37 |
| **semi-prep** | 6 | 2 | 14 | 15 | 37 |
| **Total général** | 82 | 16 | 116 | 184 | 398 |

# Functional annotation of DMs: first conclusions

- MACRO-functions are difficult to annotate BUT interesting → distribution across genres, registers, interaction with position
- Bias needed to disentangle: ideational/rhetorical, ideational/sequential, rhetorical/sequential
- ? Cross-linguistic validation of macro-functions: DMs as (linguistic) **cues** of domains increase agreement (cf. underspecified DMs), but cross-linguistic impact?
- Typical spoken domain: **sequential**

# Functional annotation of DMs: ongoing

- Functional and **relational** annotation (independently)
  - cause, consequence, contrast, temporal, reformulation, topic-shift, digression, confirmation request, …
- Co-occurrence of DMs
  - complex markers (*et puis, et alors, et voilà, ou bien, ou alors*) vs. compound (*mais bon*, voilà quoi, …) vs. sequences (*parce que bon alors …*) vs. repeated DMs (*et et, mais mais, …)*
    - Distinction between juxtaposition, addition, combination (Cuenca & Marin 2009)
- Compatibility with alternative annotation schemes

# Conclusions

- The TextLink COST Action offers a unique research environment for the investigation of discourse phenomena in multiple languages, with a focus on Discourse Relational Devices
- Insight into the discourse level gives us insight into human communication processes
  - Segmentation into BDUs gives us insight into the process of spoken language production
  - DMs are an inherent part of discourse production (and comprehension)
  - DMs structure language at a local level

# Conclusions

- DMs are multifunctional units operating at different functional levels/domains of the discourse: ideational, rhetorical, sequential, interpersonal
- Annotation of these domains is challenging, but doable!
  - Most frequent DMs are most multifunctional (underspecified DMs)
    - Complex annotation challenges (easy?) processing?
  - DMs as cues for functional domain (specific DMs)
  - Functional domains correlate with the discourse situation
- Awaiting annotation of discourse relations across domains
- Spoken language is different from written language, but similarities are more important!

# Acknowledgements


- Part of this research was carried out in collaboration with:
  - Anne Catherine Simon, Laurence J. Martin (University of Louvain)
- Financial support from grants:
  - ARC-project « A Multi-Modal Approach to Fluency and Disfluency Markers» granted by the Fédération Wallonie-Bruxelles (grant nr.12/17-044)
  - FRFC-project F 6/15 - MCF/OL - 20.472, Convention nr 2.4524.11 from the Belgian Science Foundation (FRS-FNRS).
- More information on the COST Action TextLink
  - http://textlink.ii.metu.edu.tr

References


- Afantenos, Stergos, Nicholas Asher, Farah Benamara, Myriam Bras, Cecile Fabre, Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Pery-Woodley, Laurent Prevot, Josette
- Rebeyrolles, Ludovic Tanguy, Marianne Vergez- Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: The ANNODIS corpus. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012).
- Asher, Nicholas and Alex Lascarides. 2003. Logics of Conversation. Cambridge University Press.
- Benamara, F. and M. Taboada (2015) Mapping different rhetorical relation annotations: A proposal. In Proceedings of the 4th Joint Conference on Lexical and Computational Semantics (*SEM), collocated with the Conference of the North American Association for Computational Linguistics. Denver. June 2015.
- Briz, Antonio, Pons Bordería, Salvador and Portolés, José (dirs.): Diccionario de partículas discursivas del español. www.dpde.es. Online since 2003.
- Bolly, Catherine T. , Ludivine Crible, Liesbeth Degand et Deniz Uygur-Distexhe (2015). « MDMA. Un modèle pour l'identification et l'annotation des marqueurs discursifs "potentiels" en contexte », *Discours* [En ligne], mis en ligne le 4 septembre 2015
- Bolly, Catherine, Crible, Ludivine, Degand, Liesbeth, Uygur-Distexhe, Deniz. (in press). Towards a Model for Discourse Marker Annotation in spoken French: From potential to feature-based discourse markers. In C. Fedriani & A. Sanso (eds), Discourse Markers, Pragmatic Markers and Modal Particles:
- New Perspectives, Amsterdam, John Benjamins.
- Brinton, Laurel J. 1996. *Pragmatic Markers in English: Grammaticalization and Discourse Functions*. Berlin: Mouton de Gruyter.
- Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Jan van Kuppevelt and Ronnie Smith, editors, Current Directions in Discourse and Dialogue, pages 85–112. Kluwer.
- Chiarcos, Christian. 2014. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).
- Crible, L. in press. Towards an operational category of discourse markers: A definition and its model. In C. Fedriani & A. Sanso (eds), *Discourse Markers, Pragmatic Markers and Modal Particles: New Perspectives*.
- Cuenca, M.J., Marín, M.J. (2009). Co-occurrence of discourse markers in Catalan and Spanish oral narrative. *Journal of Pragmatics* 41, pp. 899-914.

- Das, D. and M. Taboada (2013) Explicit and implicit coherence relations: A corpus study. In *Proceedings of the 2013 Annual Conference of the Canadian Linguistic Association.* Victoria, Canada.
- Degand, L. (2014). 'So very fast very fast then' Discourse markers at left and right periphery in spoken French. In Beeching, Kate and Ulrich Detges, eds. 2014. *Discourse Functions at the Left and Right Periphery: Crosslinguistic Investigations of Language Use and Language Change.* Brill: Leiden, 151-178.
- Degand, L., Cornillie, B., Pientrandrea, P. (Eds.) (2013). *Discourse Markers and Modal Particles: Categorization and description*. Amsterdam: John Benjamins.
- Degand, Liesbeth, Laurence J. Martin, and Anne-Catherine Simon. 2014. “Unités discursives de base et leur périphérie gauche dans LOCAS-F, un corpus oral multigenres annoté.” In *CMLF 2014 - 4ème Congrès Mondial de Linguistique Française 2014*, edited by EDP Sciences. Berlin, Allemagne.
- Degand, L. et Simon, A-C. (2009a). On identifying basic discourse units in speech: theoretical and empirical issues. In Discours 4 [En ligne]. URL http:// discours.revues.org/index.html.
- Degand, L. & Simon, A-C. (2009b). Mapping prosody and syntax as discourse strategies: How Basic Discourse Units vary across genres. In Wichmann, A., Barth- Weingarten, D. & Dehé, N. (eds) : Where prosody meets pragmatics : research at the interface, Studies in Pragma0tics. Bingley: Emerald, pages 79--105.
- Degand, L. & Simon, A-C. (2011). L’analyse en unités discursives de base : pourquoi et comment ?. In Langue française (2011), pages 45-59.
- Degand, L. & Simon, A.C. 2015. Variation of Discourse Markers across a multi-genre corpus of spoken French. Poster presentation at DiSpoL 2015. Identification and Annotation of Discourse Relations in Spoken Language, October 1-2 2015, Saarbrücken
- Diewald, Gabriele. 2011. « Pragmaticalization (defined) as grammaticalization of discourse functions ». *Linguistics* 49 (2): 365-90. doi:doi: 10.1515/LING.2011.011.
- Fraser, Bruce. 1999. « What are discourse markers? » *Journal of Pragmatics* 31 (7): 931-52. doi:10.1016/S0378-2166(98)00101-5.
- Graesser, A. C., K. K. Millis, et R. A. Zwaan. 1997. « Discourse Comprehension ». *Annual Review of Psychology* 48: 163-89. doi:10.1146/annurev.psych.48.1.163.
- Gonzalez, M. 2005. “Pragmatic markers and discourse coherence relations in English and
- Catalan oral narrative”. In: Discourse studies 7.1, 53–86.
- Halliday, M. A. K. and R. Hasan. 1976. Cohesion in English. London : Longman.
- Hansen, Maj-Britt Mosegaard. 1997. « Alors and donc in spoken French: A reanalysis ». *Journal of Pragmatics* 28 (2): 153-87. doi:10.1016/S0378-2166(96)00086-0.

- Holsánová, J. 2008. *Discourse, Vision, and Cognition*. Human Cognitive Processing 23. Amsterdam: John Benjamins Publishing.
- Lapshinova-Koltunski, E. and K. Kunz (2014). Conjunctions across Languages, Registers and Modes: semi-automatic extraction and annotation. In Diaz-Negrillo, A. and J. Diaz-Perez Francisco (eds). Specialisation and Variation in Language Corpora. Peter Lang.
- Lapshinova, E., Nedoluzhko, A., and Kunz, K. (2015). Cross languages and genres: Creating a universal annotation scheme for textual relations. In Rehbein, I. and Zinsmeister, H., editors, Proceedings of the Workshop on Linguistic Annotations, NAACL-2015, Denver, USA.
- Lewis, Diana M. (2011). A discourse-constructional approach to the emergence of discourse markers in English ». *Linguistics* 49 (2): 415-43.
- Mann, William C.  and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. Text, 8(3):243–281.
- Martin, Laurence J., Degand, Liesbeth & Simon, Anne Catherine. (2014). Forme et fonction de la périphérie gauche dans un corpus oral multigenres annoté. *Corpus* 13. 243-265. [available on-line : http://corpus.revues.org/2154]
- Maschler, Yael. (2009). Metalanguage in Interaction: Hebrew Discourse Markers. Amsterdam / Philadelphia: John Benjamins.
- Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In 6th International Conference on Language Resources and Evaluation (LREC).
- Prasad, Rashmi, Aravind Joshi and Bonnie Webber (2010). Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). Beijing, China, August 2010*
- Prasad, Rashmi, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. Computational Linguistics,
- 40(4):921–950.
- Redeker, G. 1990. “Ideational and pragmatic markers of discourse structure”. In: Journal
- of Pragmatics 14.3, 367–381.
- Roze, C., Danlos, L., and Muller, P. (2012). Lexconn: a French lexicon of discourse connectives. Revue Discours [Online] URL : http://discours.revues.org/8645.

- Rysová, Magdaléna. 2012. Alternative Lexicalizations of Discourse Connectives in Czech. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association, Istanbul, Turkey, pp. 2800–2807.
- Sanders, T.J.M., Spooren, W.P.M.S. & Noordman, L.G.M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15 (1), (pp. 1-35)
- Sankoff, Gillian, Pierrette Thibault, Naomi Nagy, Helene Blondeau, Marie-Odile Fonollosa, et Lucie Gagnon. 1997. « Variation in the Use of Discourse Markers in a Language Contact Situation. » *Language Variation and Change* 9 (2): 191-217.
- Schiffrin, Deborah. 1987. *Discourse markers*. Cambridge: Cambridge University Press.
- Schourup, Lawrence. 1999. « Discourse markers ». *Lingua* 107 (3-4): 227-65. doi:10.1016/S0024-3841(96)90026-1.
- Selting, M. (2000). The construction of units in conversational talk. Language in Society, 29, pages 477--517.
- Spooren, Wilbert & Degand, Liesbeth (2010). Coding coherence relations: reliability and validity. *Corpus Linguistics and Linguistic Theory* 6(2), 241–266
- Stede, M. (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, 2002.
- Sweetser, E. 1990. From etymology to pragmatics. Cambridge : CUP.
- Taboada, M. (2006) Discourse Markers as Signals (or Not) of Rhetorical Relations. *Journal of Pragmatics* 38(4): 567-592.
- Taboada, M. (2009) Implicit and Explicit Coherence Relations. In J. Renkema (ed.) *Discourse, of Course*. Amsterdam/Philadelphia: John Benjamins. 127-140.