"Towards a Model for Discourse Marker Annotation in spoken French: From potential to feature-based discourse markers"

Bolly, Catherine ; Crible, Ludivine ; Degand, Liesbeth ; Uygur, Deniz

ABSTRACT

Starting from the common observation that there is no recognized closed class of discourse markers (DMs) and that a number of linguistic markers may or may not count as DMs according to the definitions at stake (Schourup 1999: 228), we aim to present an empirical method for the identification and annotation of DMs in spontaneous spoken French (MDMA project). Central to our proposal is that DMs may be described as clusters of features that, in specific patterns of combination, allow distinguishing DM use from other uses. We proceeded in three steps: (i) using a very broad definition of DMs – i.e. items that "provide instructions to the hearer on how to integrate their host utterance into a developing mental model of the discourse in such a way as to make that utterance appear optimally coherent" (Hansen 2006: 25) – three analysts identified all potential DMs in an 800 words transcript; (ii) all types found were then extracted from a balanced 10,000 words corpus; and (iii) analyzed according to more than 10 features (including syntactic, semantic, collocational, and prosodic features). The hypothesis underlying our annotation experiment is that the analysis of the distributional constraints imposed on specific markers should uncover reliable features for the identification and categorization of DMs. The potential DMs extracted at the first step of analysis refer to those linguistic expressions that can, in one context or another, fulfill a DM function, i.e. be used at either of the following “levels” or “domains”: “the sequential structure of the dialogue,...
Towards a Model for Discourse Marker Annotation in spoken French:
From potential to feature-based discourse markers

Catherine Bolly
Ludivine Crible
Liesbeth Degand
Deniz Uygur-Distexhe

« Pragmatic markers, discourse markers and modal particles: What do we know and where do we go from here? » Oct. 16-17 2014, Como, Italy
Identifying discourse markers in spontaneous speech

• Common ground:
  • Lack of consensus
    « 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)
  • Lack of satisfying models (see Uygur-Distexhe 2012 and Crible 2014 for exceptions)
  • Proliferation of case studies

➢ Need for contrastive, categorical, onomasiological studies
➢ Towards a comprehensive model of annotation
  • From identification to description in context
  • Corpus-based selective criteria for inclusion/exclusion of DMs
DM definition

What are discourse markers?

“A pragmatic marker* is defined as a phonologically short item that is not syntactically connected to the rest of the clause (i.e., is parenthetical), and has little or no referential meaning but serves pragmatic or procedural purposes” (Brinton, 2008: 1)

DMs “provide instructions to the hearer on how to integrate their host utterance into a developing mental model of the discourse in such a way as to make that utterance appear optimally coherent” (Hansen 2006: 25)

- **Multifunctionality** within their “meaning potential” (Aijmer 2013) and in context
- **Metalinguistic** comments on the content/form of the linguistic utterance itself
- **Intersubjectivity** effects on interaction management
... et anorexique je ne parvenais pas à le retenir / j’ai / alors je pense à quelque ch/ je pensais à anus (rires) / comme c’est quand même le tube digestif hein qui est en bas (rires) et ça va depuis lors je n’oublie plus (rires) et encore l’autre jour aussi un mot / tiens je ne sais p/ tu vois / si / j’ai / j’oublie certains mots / ’fin / je retombe dessus après hein...

... and anorexic I couldn’t memorize it / I / so I think of someth / I thought of anus (laughing) / since it’s still the digestive tube right which is at the bottom (laughing) and it’s ok since then I don’t forget anymore (laughing) and again the other day too a word / see I don’t kn/ you see / yes / I / I forget certain words / well / I remember them afterwards right...

ID code: ageJM1; Pseudo: Jeanne; Age: 90; Source: Corpage 2012; Task 1 Line life; Time code: 1:13:10
MDMA Research group

- Check the relevance/reliability of identifying features of DMs
- Towards objective features to be automated (POS-tags, co-occurrence, position)

- Method overview:
  1. Intuitive selection of potential DMs
  2. Extraction of potential types in balanced corpus FR-BE
  3. Elaboration of the coding scheme from theory
  4. Annotation and revision of the scheme
  5. Statistical analysis of «predictive» parameters
3. Elaboration of the coding scheme from theory

- What to code and how?
- Selection of parameters to code from existing definitions
- Constant work-in-progress

(Schourup 1999, Schiffrin 1987, Brinton 2008...)

- Theory
- Selection
- Corpus annotation
4. Annotation and revision of the scheme

- Syntax
- Semantics
- Co-text
- Degree of certainty
4. Annotation and revision of the scheme

- **Syntax:**
  - **Category**
    « ben c’est-à-dire que bon comme bon quand je suis … » [« that is to say »] ➔ VP

- **Position in the utterance** (Lindström 2001)
  « si j’ai bien compris » [« correctly »] ➔ middle field

- **Position in turn**

- **Mobility**
  « il y avait d’autres problèmes non mais c’est vrai » [« no but »] ➔ no
4. Annotation and revision of the scheme

• Syntax
• Semantics:

  • Procedurality
  « non mais faut dire qu’en fait il a réfléchi » [« the thing is » VS. « we need to say that »] ➔ conceptual-procedural

  • Coded meaning
  « j’aimerais bien travailler dans ce genre de matériel » ➔ yes
4. Annotation and revision of the scheme

- Syntax
- Semantics
- Co-text:
  - Co-occurring DM
  - Pauses
4. Annotation and revision of the scheme

- Syntax
- Semantics
- Co-text
- Degree of certainty:

« hmm oui je pensais que on fait avec ça » ➔ 2, 3, 1, 1 ➔ 1,75 (mean)
4. Annotation and revision of the scheme

- **Syntax**:
  - Category: VP, clause, NP, adv …
  - Position in utterance: pre-field, initial, middle, end, post-field
  - Position in turn: initial, medial, final, autonomous
  - Mobility: yes / no
  - Optionality: yes / no

- **Semantics**:
  - Procedural meaning: procedural, conceptual, conc-proc
  - Prototypical meaning or not: yes / no

- **Co-text**:
  - Co-occurrence: no / left / right / both
  - Pause: no / left / right / both

- **Degree of certainty**: 1 – 3

Kappa-scores from 0.75 to 0.82
From theory to data and back again

- Corpus-based annotation scheme, moulded by the annotation experiment
- Scope over both DM tokens and their non-DM forms
- Qualitative alternative to kappa-scores

Statistical processing:
- Relative relevance of variables
- Clusters of co-occurring features
Overall distribution of some parameters

Both DMs & non-DMs

Degree of certainty (mean score) % of tokens
1 - 1,25 34 %
1,5 - 2,5 46,5 %
2,75 - 3 19,5 %
Ranking the variables

Conditional importance of variables

Synt.Pos
Sem.Coded.meaning.in.context
Synt.Categ
SyntMob
Turn
Pause

Procedurality removed because circular
Predict DM status on feature patterns

Pseudo-R² 0.78
Predict DM status on feature patterns

Pseudo-R² 0.78
Predict DM status on feature patterns

Pseudo-$R^2$ 0.78
Predict DM status on feature patterns

Pseudo-$R^2$ 0.78
Confirmed vs discarded DMs

Typical features of confirmed DMs:

- Conjunctions > particles > adverbs
- Pre-field
- Not coded meaning
- Procedural meaning

Typical features of discarded DMs:

- Adverbs > pronouns > VP
- Middle field, end field
- Coded meaning
- Conceptual-procedural

*et alors c'est très compliqué de pouvoir régler ces problèmes donc c'est même je dirais c'est presque plus du français*

*j'aimerais bien travailler dans ce genre de matériel*
High-profile clusters of features
Borderline clusters of features
Borderline cases

From 1.5 to 2.5 (included) degree of certainty

➢ Other PMs?
  • Modal adverbs: vraiment, encore, bien sûr
  • Response markers: c’est ça, c’est juste, oui, non, voilà
  • Hesitation particles: euh

➢ Typical features:
  • Autonomous position
  • Procedural meaning
Conclusion: a confirmatory method for operational annotation scheme

- Novelty of the approach to both DMs and their propositional (non-DMs) equivalents
- Provides corpus-based validation of the annotation model and its parameters
- Visualization of clusters of co-occurring features
DMs profiling

- Corpus-based profiles of typical features
  - Intuition-based potential DMs either confirmed or discarded
  - Borderline cases: different types of pragmatic markers

- Overall, syntactic position stands out as most predictive
  - Pre-field, i.e. not integrated: main difference with MPs
Perspectives

- Functional annotation (forthcoming)
  - same method but for confirmed and potential DMs only (>1,5)

- Test-bed for application of our corpus-based scheme to other data types:
  - Other corpora, genres, languages, modalities

- Semi-automatic annotation?
  - Surface features: more objective
  - Still difficult to automate
References


Hansen, M.-B. M. 2006. « A dynamic polysemy approach to the lexical semantics of discourse markers (with an exemplary analysis of French *toujours*) ». In K. Fischer (ed.), *Approaches to discourse particles*. Amsterdam : Elsevier. 21-41.


THANK YOU