Capturer le discours
Données, dictionnaires et automates

Mathilde Dargnat, UL & ATILF
Jacques Jayez, ENS de Lyon & LORIA
Paola Herreño Castañeda, UL & ATILF, ANR CODIM
Maeva Sillaire, UL & ATILF, ED SLTC

(merci aussi à Karolin Boczon, UL & IDMC, et à Yannick Parmentier, UL & LORIA)

JTTA, 22 mars 2024, ATILF, Nancy
« La linguistique de corpus à la croisée des questionnements théoriques, méthodologiques et empiriques »

(Mathilde)

Présentation

- **Enjeux du projet CODIM**
- **Préparation des données**
- **D'abord la description d'un marqueur isolé**
- **Domaines et fonctions des MD**
- **Perspectives**

Enjeux du projet CODIM (1)

- **CODIM** = Compositionality and Discourse Markers

- Étude des **combinaisons de marqueurs discursifs** (MD) :
  ah bon, mais enfin, donc du coup, mais quand même, merde alors, etc.

- Avant les combinaisons, les **marqueurs isolés**
  → une littérature très (trop) abondante, mais avec des constantes :

  – expressions invariables, plutôt courtes
  – souvent polyfonctionnelles (dans la langue et comme MD)
  – ne contribuent souvent pas au contenu vériconditionnel*
  – sens procédural plutôt que conceptuel
  – servent à :
    - organiser le discours (relations de cohérence) [DMC]
    - manifester les attitudes et émotions du locuteur |DMP]
    - gérer l'interaction [DMP]

(Mathilde)

Enjeux du projet CODIM (2)

(1) **parce que** [acte de langage → justification]

L1 – Qu'est-ce que vous
L2 – Oh
L1 – pensez en général des femmes qui travaillent ? **parce que** y en a de plus en plus qui travaillent [ESLO, lg. 361911]

(2) **tu parles** [marqueur épistémique d'évidence]

L1 – ben qu'est-ce que je veux dire ? ma copine elle lui avait fait croire que
L2 – (silence)
L1 – un de ses copains avait reçu une boule de bowling sur la tête
L2 – et puis qu'il était à l'hôpital
L1 – (silence)
L2 – une boule de bowling **tu parles** [ESLO, lg. 83 991]

(3) **hein** [demande de confirmation/appel à l'écoute]

L1 – soit ben là j'essaie de de ouais de rentrer un peu dans la communauté on va dire euh ben par l'intermédiaire d'associations euh notamment par euh ma fac
L2 – oui
L1 – et voilà donc là c'est c'est tout nouveau **hein** c'est depuis cette année que je me lance euh à aller vraiment dans des euh [TCOF, lg. 211]

Enjeux du projet CODIM (3)

Problèmes soulevés par l'analyse des combinaisons de MD

Récupérer les combinaisons pertinentes

Propriétés statistiques

Mesures d'association, résultats d'apprentissage automatique

cf. Brezina 2018 pour les mesures

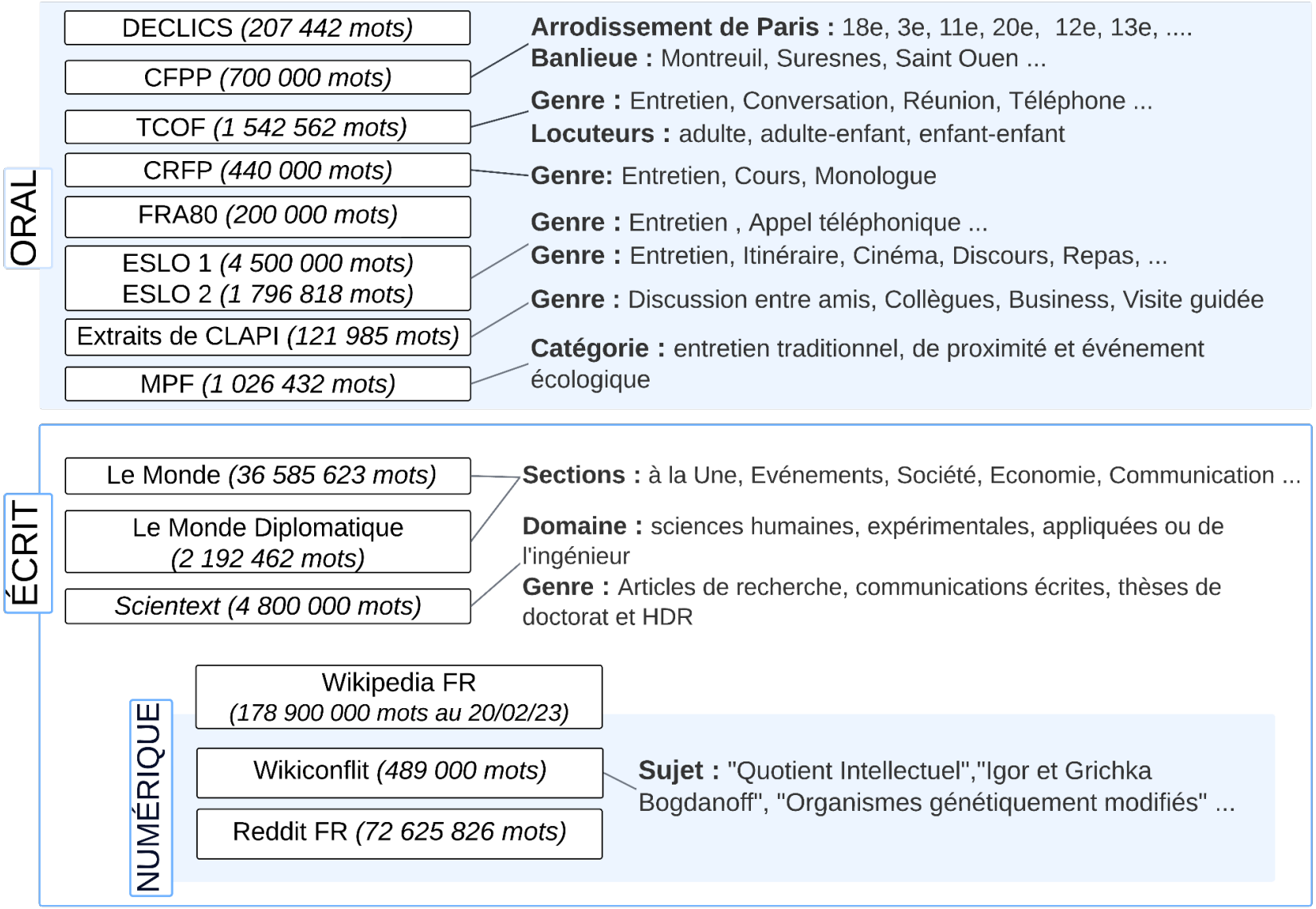
cf. Dargnat 2022, Dargnat & Jayez 2021 pour les résultats des 15 mesures pour *mais* en général et *mais enfin* spécifiquement

Expliciter les propriétés sémantico-pragmatiques des combinaisons

(Non-)compositionalité, interaction avec la prosodie, modélisation sémantique

(Maeva)

Préparation des données (1)



- **Sélection** et **récupération** des corpus
- **Uniformisation** des structures et transcriptions hétérogènes des corpus

≈ 300 000 000 mots

(Maeva)

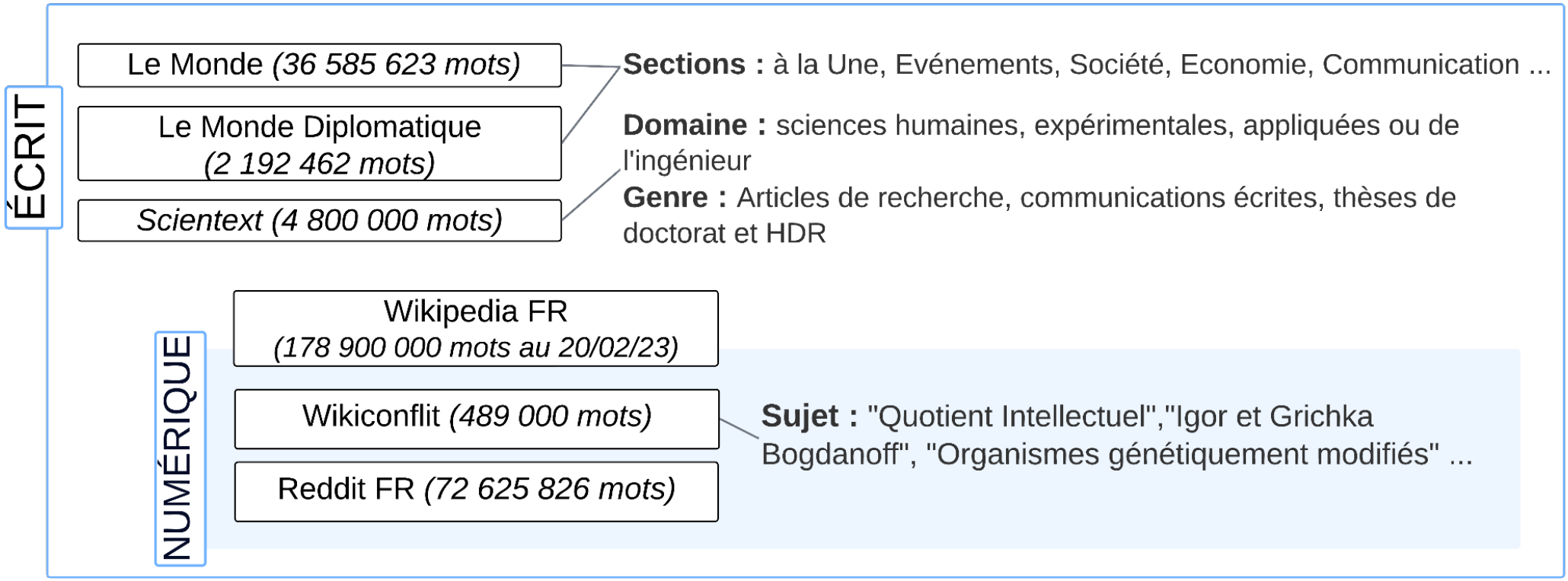
Préparation des données (2)

ORAL

DECLICS *(207 442 mots)*	**Arrodissement de Paris :** 18e, 3e, 11e, 20e, 12e, 13e, ... **Banlieue :** Montreuil, Suresnes, Saint Ouen ...
CFPP *(700 000 mots)*	**Genre :** Entretien, Conversation, Réunion, Téléphone ... **Locuteurs :** adulte, adulte-enfant, enfant-enfant
TCOF *(1 542 562 mots)*	**Genre:** Entretien, Cours, Monologue
CRFP *(440 000 mots)*	**Genre :** Entretien , Appel téléphonique ...
FRA80 *(200 000 mots)*	**Genre :** Entretien, Itinéraire, Cinéma, Discours, Repas, ...
ESLO 1 *(4 500 000 mots)* ESLO 2 *(1 796 818 mots)*	**Genre :** Discussion entre amis, Collègues, Business, Visite guidée
Extraits de CLAPI *(121 985 mots)*	**Catégorie :** entretien traditionnel, de proximité et événement écologique
MPF *(1 026 432 mots)*	

(Maeva)

Préparation des données (3)



(Maeva)

Préparation des données (4)

- **Produit final :** Corpus uniformisés - Forme tabulaire des données et métadonnées (fichier .csv et/ou base de données)

- **Objectifs :** Statistiques (R ou Python), meilleure transférabilité, longévité des données au sein de CODIM, bases de données

startTime	endTime	speaker_name	speaker_native	text	audio_filename	trs_filename	age_speaker	profession_speaker	...
0.000	31.210	Anita Musso	native	bon je reviens sur cette euh ce problème qui e...	Anita_MUSSO_F_46_11e.wav	Anita_MUSSO_F_46_11e-v2.trs	46.0	Auxiliaire de vie (accompagnement d'enfants a...	...
31.210	39.843	Sonia Branca-Rosoff (ENQ)	native	mais pour l'instant c'est une vraie question h...	Anita_MUSSO_F_46_11e.wav	Anita_MUSSO_F_46_11e-v2.trs	63.0	Professeur des Universités	...
39.843	49.601	Anita Musso	native	sur le coup je me dis je vais mettre cinq minu...	Anita_MUSSO_F_46_11e.wav	Anita_MUSSO_F_46_11e-v2.trs	46.0	Auxiliaire de vie (accompagnement d'enfants a...	...
49.601	51.075	Sonia Branca-Rosoff (ENQ)	native	et pourtant en	Anita_MUSSO_F_46_11e.wav	Anita_MUSSO_F_46_11e-v2.trs	63.0	Professeur des Universités	...

Préparation des données (5)

- Nettoyage pour l'application d'automates finis avec Unitex
  https://unitexgramlab.org

- Objectif : éliminer les occurrences non MD

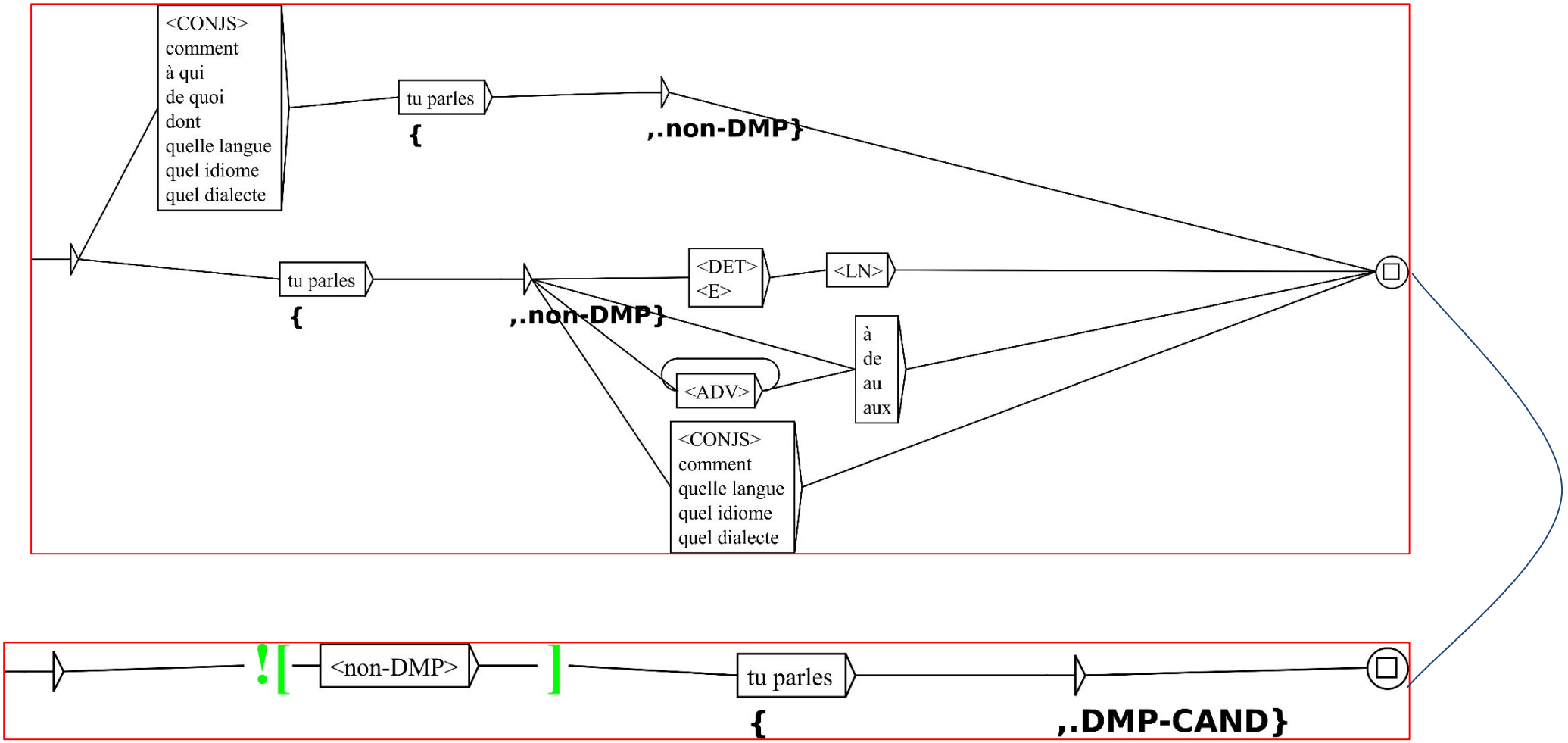
  ex. *bon* comme ADJ vs *bon* comme MD
  ex. *bien*, *dis*, *tu parles, tiens*, etc.

- Utilisation de ***cascades*** (= séquences d'application de graphes)

  Deux principes
  1. Élimination : trier d'abord les non MD
  2. Préemption : traiter les expressions englobantes d'abord
     ex. *bon courage/appétit* avant *bon*

Préparation des données (6)

Cascade pour le traitement de *tu parles*



Préparation des données (7)

État des données après nettoyage et avant exploration par les graphes, extrait de CFPP

CFPP_raw.snt (/mnt/12f5186a-36b4-4ab9-be9b-d0707b48077f/ANR_CODIM/Corpora)

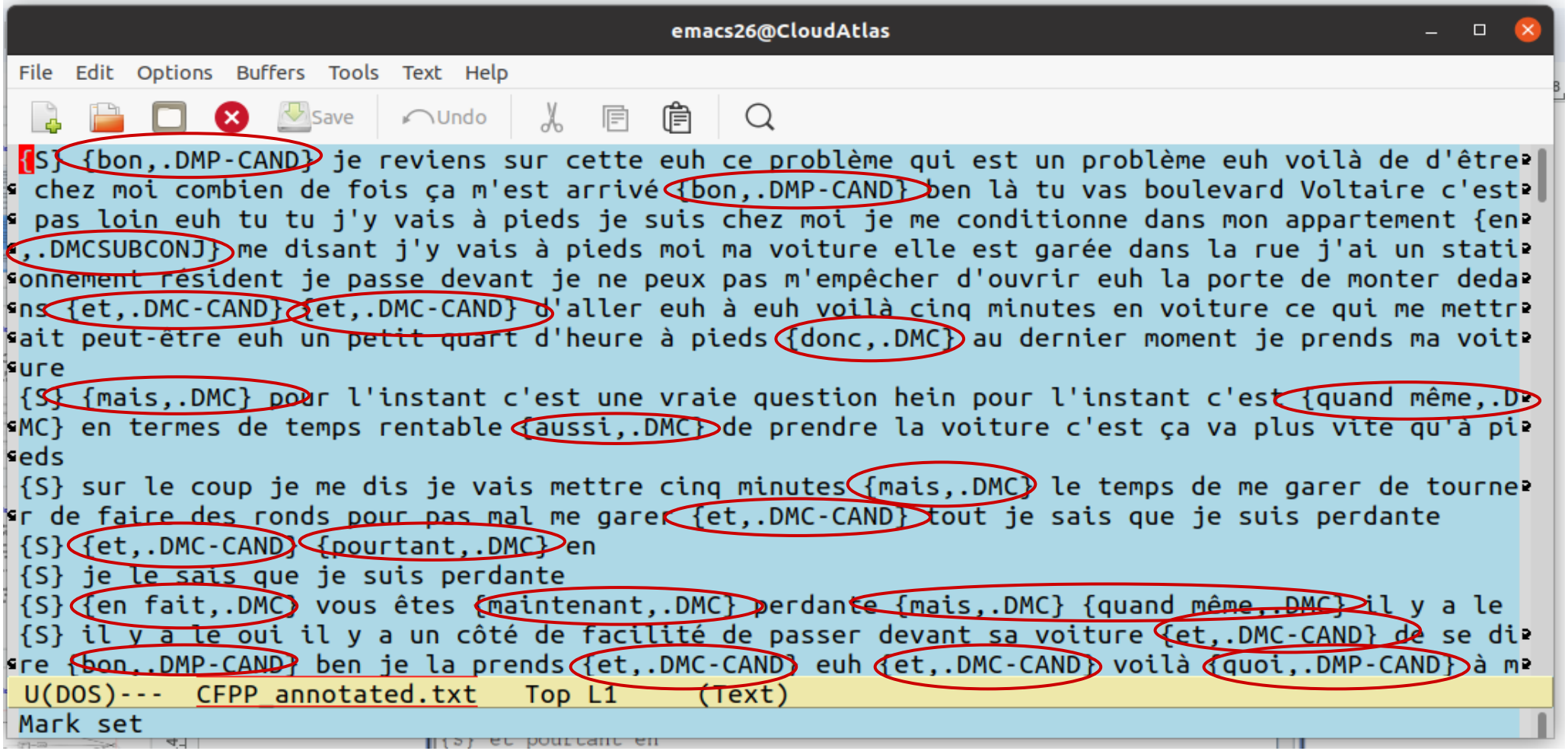
56098 sentence delimiters, 1296955 (15866 diff) tokens, 591963 (15832) simple forms, 181 (10) digits
578089 occurrences (19374 DLF entries) simple words, 3625 occurrences (3724 DLC entries) compound words, 10495 oc...

```
{S} bon je reviens sur cette euh ce problème qui est un problème
euh voilà de d'être chez moi combien de fois ça m'est arrivé bon
ben là tu vas boulevard Voltaire c'est pas loin euh tu tu j'y vais
à pieds je suis chez moi je me conditionne dans mon appartement en
me disant j'y vais à pieds moi ma voiture elle est garée dans la
rue j'ai un stationnement résident je passe devant je ne peux pas
m'empêcher d'ouvrir euh la porte de monter dedans et et d'aller euh
à euh voilà cinq minutes en voiture ce qui me mettrait peut-être
euh un petit quart d'heure à pieds donc au dernier moment je prends
ma voiture
{S} mais pour l'instant c'est une vraie question hein pour
l'instant c'est quand même en termes de temps rentable aussi de
prendre la voiture c'est ça va plus vite qu'à pieds
{S} sur le coup je me dis je vais mettre cinq minutes mais le temps
de me garer de tourner de faire des ronds pour pas mal me garer et
tout je sais que je suis perdante
{S} et pourtant en
{S} je le sais que je suis perdante
{S} en fait vous êtes maintenant perdante mais quand même il y a le
```

(Jacques)

Préparation des données (8)

Après exploration ≃ 100 graphes sur le corpus CFPP
(nbre total de graphes : 572, nbre total de MD : 798)



(Mathilde)

D'abord la description des MD isolés (1)

Objectifs

– Viser une description multidimensionnelle phonétique/syntaxe/sémantique/pragmatique
– Expliciter les contraintes d'interprétation d'un MD isolé
– Faciliter la comparaison entre MD et les contraintes de combinaison entre MD

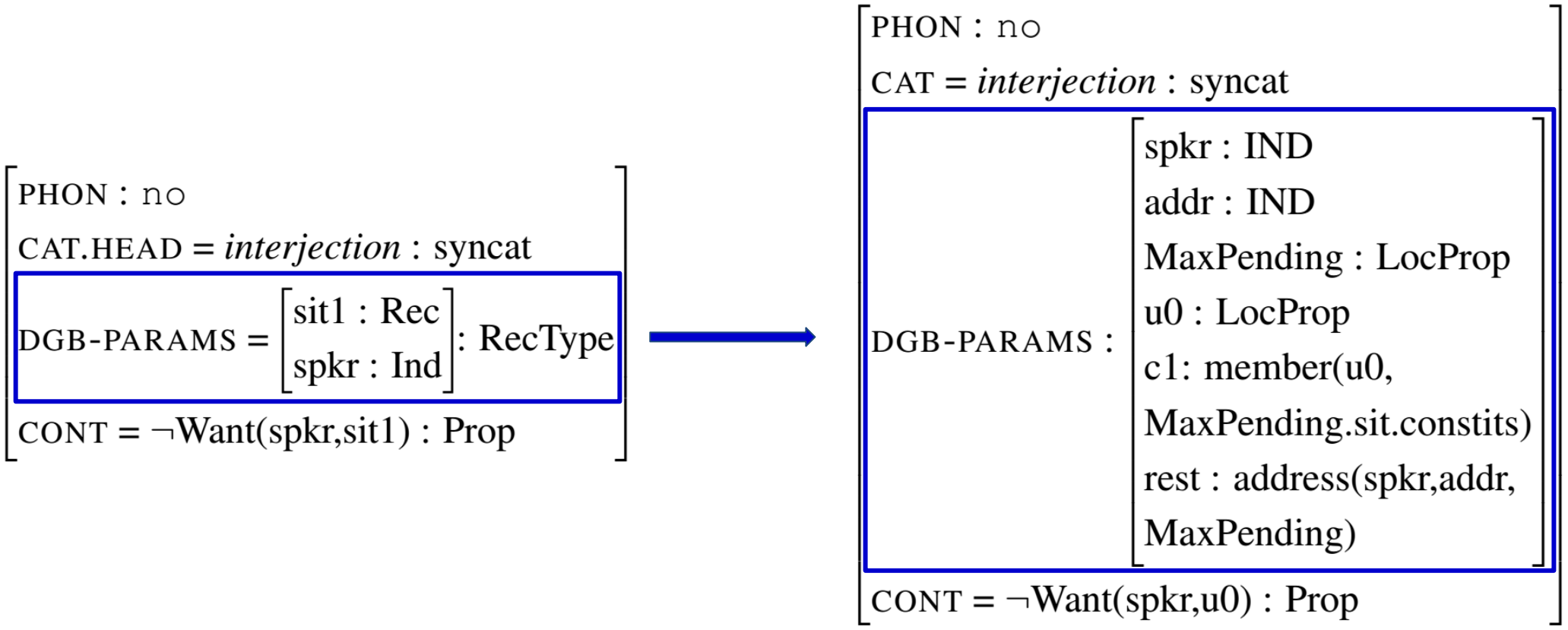
Cadre pour la description et la représentation des propriétés

– Structures de traits (AVM)
– Réflexion dans le cadre de la Type Theory with Records (TTR)
[cf. Ginzburg 2012, Cooper 2023]
– Dimensions :
- phonétique (PHON)
- syntaxique (CAT.HEAD)
- contenu propositionnel (CONT)
- contexte interactionnel (DGB-PARAMS / Dialogue Gameboard Parameters)

(Mathilde)

D'abord la description des MD isolés (2)

(4) *(L1 opens freezer to discover smashed beer bottle)*
L1 – **No !** *(I do not want this (the beer bottle smashing) to happen)*



Cf. Tian et al. 2015, Cooper & Ginzburg 2015, Ginzburg 2012.

(Mathilde)

D'abord la description des MD isolés (3)

(5) L1 – Show flights arriving in **uh** Boston [Shriberg 1994]

$$\begin{bmatrix} \text{PHON} : \text{uh} \\ \text{CAT} = \textit{interjection} : \text{syncat} \\ \text{DGB-PARAMS} : \begin{bmatrix} \text{spkr} : \text{IND} \\ \text{addr} : \text{IND} \\ \text{MaxPending} : \text{LocProp} \\ u0 : \text{LocProp} \\ \text{c1: member}(u0, \text{MaxPending.sit.constits}) \\ \text{rest} : \text{address(spkr,addr,MaxPending)} \end{bmatrix} \\ \text{CONT} = \left[\text{c1} : \text{FLDEdit(spkr,addr,MaxPending)}\right] : \text{Prop} \end{bmatrix}$$

Ginzburg et al. 2014 : 51 sqq.

(Paola)

Domaines et fonctions des MD (1) : Contexte

Objectifs

- Faire une généralisation permettant d'être appliquer à l'ensemble de MD
- Être capable de les intégrer dans les AVM

Dimensions des MD :

– Sémantique → Représentation du contenu propositionnel - état de choses
– Pragmatique → Expression d'attitudes et intentions

Niveaux du discours (Schiffrin, 1987) :

– Structure idéationnelle → Lien entre propositions
– Structure d'action → Lien entre actes du langage
– Structure d'échange → Tours de parole (prendre ou céder)
– État d'information → Organisation de connaissances
– Cadre de participation → Établir relations entre les interlocuteurs

(Paola)

Domaines et fonctions des MD (2)

Bühler (1934) → Composants d'une énonciation :

1. Représentation des idées
2. Expression de l'énonciateur
3. Appel à l'interlocuteur

Domaines → "Relations entre des unités discursives" (Redeker, 1990)

4. Idéationnelle → État de choses
5. Rhétorique → Motivation d'une énonciation en termes de croyances et intentions
6. Séquentielle → Structuration des segments

7. Interpersonnelle → Relation entre l'énonciateur et l'interlocuteur

(Paola)

Domaines et fonctions des MD (3)

Fonctions
– Signification procédurale
– Signification essentielle

Typologie de Dostie (2004):
– Connecteurs textuels (S1 + CT + S2)
– Marqueurs Discursifs
  Interprétation
  Réalisation d'un acte illocutoire
  Appel à l'écoute
  Écoute
  Balisage

Fonctions de Crible et Degand (2019):
Addition, Alternative, Cause, Concession, Condition, Consequence, Contrast, Hedging, Monitoring, Specification, Temporal, Agreeing, Disagreeing, Topic, Quoting

Merci de votre attention

Prochainement sur vos écrans :

METZ, 21-22 juin 2024 (site de Metz)
Colloque international **“Discourse Markers: Markers in discourse, Markers on Discourse”**
co-org. F. Berthe, M. Dargnat, A. Fetzer, I. Gaudy-Campbell

PARIS : 9-14 septembre 2024 (Sorbonne Université)
Workshop du GDR LLcD **“Coo-currences et marquage discursif”**
co-org. M. Dargnat et A. Tutin

NANCY : 13 septembre 2024 (ATILF)
JTTA **“Lexique et argumentation”**
co-org. M. Dargnat et H. Vinckel-Roisin