
# 24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)

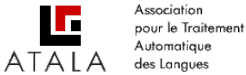
**Orléans, France – 26-30 juin 2017**

https://taln2017.cnrs.fr

# Actes de l'atelier « ACor4French – Les corpus annotés du français » (ACor4French 2017)

**Laurence Danlos, Karën Fort, Bruno Guillaume, Sylvain Kahane (Eds.)**

**Sponsors :**


ATALA
Association
pour le Traitement
Automatique
des Langues


Ortolang


cnrs
dépasser les frontières


Liberté · Égalité · Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE
DE LA CULTURE


acatus
informatique


aktan


proxem


inbenta


Orléans
Mairie


ORLÉANS
MÉTROPOLE


LOIRE&ORLEANS
ECO


FRENCH TECH
LOIRE
VALLEY


UNIVERSITE D'ORLEANS


Laboratoire
Ligérien de
Linguistique


LABORATOIRE
D'INFORMATIQUE
LIFO
FONDAMENTALE
D'ORLEANS


LI
Laboratoire d'Informatique
EA 2101

# Préface

Dans de nombreuses tâches du TAL, les corpus annotés (semi-)manuellement sont utilisés comme données d'apprentissage et/ou comme données de référence pour l'évaluation des outils. Dans les deux cas, le fait de disposer de corpus annotés de qualité est un enjeu essentiel.

Pour la langue française, les corpus annotés ont fait l'objet de nombreux projets pendant ces dix dernières années (FTB, PFC, Valibel, Sequoia, FDTB, Rhapsodie, Annodis, Orféo, ...) que ce soit pour l'analyse syntaxique (en constituants ou en dépendances), pour l'analyse du discours, pour les anaphores pronominales, pour la prosodie, etc. Ces projets ont fait des choix linguistiques souvent indépendants les uns des autres et les données ne sont pas toujours facilement convertibles d'une ressource à l'autre. De plus, ces ressources ne sontpas forcément libres ou n'ont pas toutes des licences compatibles qui permettent de redistribuer de nouvelles versions corrigées.

Par ailleurs, un nouveau projet international de création de corpus annotés en syntaxe de dépendances pour différentes langues a démarré (Universal Dependencies) qui pousse à réfléchir et à repositionner les données spécifiques au français. De plus, la création de ressources de ce type par myriadisation par le jeu (jeux ayant un but) pose de nouvelles questions, par exemple quant à l'expertise des annotateurs.

L'objectif de cet atelier est que les chercheurs impliqués dans ces différents développements pour la langue française se rencontrent pour faire un état des lieux des corpus disponibles, des besoins futurs et des nouvelles initiatives qui pourraient se mettre en place pour coordonner les prochains projets de développement de corpus afin qu'ils s'enrichissent mutuellement.

# Comités

## Comité d'organisation

Laurence Danlos (LLF, Paris 7)
Karën Fort (STIH, Paris-Sorbonne)
Bruno Guillaume (Loria, Nancy)
Sylvain Kahane (Modyco, Nanterre)

## Comité scientifique

Christophe Benzitoun (ATILF, Nancy)
Philippe Blache (LPL, Aix-Marseille)
Marie Candito (LLF, Paris 7)
Mathieu Constant (ATILF, Nancy)
Laurence Danlos (LLF, Paris 7)
Marie-Catherine de Marneffe (OSU, Ohio)
Iris Eshkol-Taravella (LLL, Orléans)
Carole Etienne (ICAR, Lyon)
Cécile Fabre (ERESS, Toulouse)
Karën Fort (STIH, Paris 4)
Kim Gerdes (LPP, Paris 3)
Bruno Guillaume (Loria, Nancy)
Sylvain Kahane (Modyco, Nanterre)
Anne Lacheret (Modyco, Nanterre)
Frédéric Landragin (Lattice, Paris)
Marie-Claude Lhomme (OLST, Montréal)
Yann Mathet (GREYC, Caen)
Philippe Muller (IRIT, Toulouse)
Alexis Nasr (LIF, Aix-Marseille)
Guy Perrier (Loria, Nancy)
Sophie Rosset (Limsi, Paris)

Djamé Seddah (Alpage, Paris 4)
Anne-Catherine Simon (UCL, Louvain)

# Table des matières

## Session « oral »

# Annotations de particules de discours en français sur une large variété de corpus

Katarina Bartkova[1], Mathilde Dargnat[1], Denis Jouvet[2], Lou Lee[1]
(1) CNRS, ATILF, UMR 7118, Nancy, F-54063, France
(2) Université de Lorraine, CNRS, Inria, LORIA, UMR 7503, Nancy, F-54000, France
{katarina.bartkova;mathilde.dargnat}@univ-lorraine.fr,
lou.lee4@etu.univ-lorraine.fr, denis.jouvet@loria.fr

## RESUME

En français, certains mots et expressions sont fréquemment utilisés en tant que particules de discours dans le langage parlé, en particulier en parole spontanée. Comme la signification sémantique de tels mots varie selon qu'ils sont utilisés en tant que particule de discours ou non, l'identification correcte de leur fonction discursive est importante. Pour étudier les particules de discours, ainsi que leurs corrélats prosodiques, une large variété de corpus de parole correspondant à différents degrés de spontanéité sont considérés. Cela va de la parole préparée (e.g., contes et journaux d'information radiodiffusés) à la parole spontanée (e.g., interviews et interactions entre personnes). Ce papier présente les corpus considérés, la sélection d'occurrences des mots, l'annotation de leur fonction discursive, les paramètres prosodiques calculés, ainsi que la fréquence d'utilisation de quelques mots en tant que particule de discours sur les divers corpus.

## ABSTRACT

**Annotation of discourse particles In French over a large variety of speech corpora.**
In French, some words and expressions are frequently used as discourse particles in spoken language, especially in spontaneous speech. As the semantic meaning of such words differ whether they are used as discourse particles or not, the correct identification of their discourse function is of great importance. To study such discourse particles, as well as their prosodic correlates, a large variety of speech corpora exhibiting various degrees of spontaneity are considered. They range from prepared speech (e.g., storytelling and broadcast news) to spontaneous speech (e.g., interviews and interactions between people). This paper presents the speech corpora considered, the selection of word occurrences, the annotation of their discourse particle function, the computed prosodic features, as well as the frequency of usage of a few selected words as discourse particle on the various corpora.

MOTS-CLES : Particules de discours, français, paramètres prosodiques, annotation
KEYWORDS: Discourse particle, French language, prosodic parameters, annotation

# 1 Introduction

Ce que nous ciblons par *particules discursives* (DPs, pour « *Discourse Particles* ») correspond à un ensemble de mots ou de syntagmes fonctionnant au niveau du discours, entendu comme acte de communication verbale. Bien qu'on puisse globalement inscrire les DPs dans la classe des adverbes,

leur catégorisation pose problème. D'une part, car certains items peuvent avoir d'autres usages dans la langue (pronom, adjectif, nom, etc.). D'autre part, car leur signification est souvent plus complexe que celle des représentants traditionnels de la classe des adverbes, à savoir les adverbes de manière ou d'énonciation. Pour une discussion sur la catégorisation, se reporter entre autres à (Bouchard, 2002 ; Degand & Fagard, 2011 ; Dostie, 2004 ; Fernandez-Vest, 1994 ; Fischer, 2006 ; Hansen, 1998 ; Métrich et al., 2002 ; Paillard, 1998).

Les DPs sont massivement représentées dans les échanges oraux spontanés, d'où leur grande variabilité, à tous les niveaux (Gadet, 2003 ; Koch & Osterreicher, 2001). On peut néanmoins citer quelques exemples prototypiques pour le français : *bon, alors, ben, ah, voilà, quoi*, *tu sais, tu vois/vous voyez*, etc. Les DPs peuvent aussi se combiner de manière plus ou moins compositionnelle (*ah bon, bon ben alors, mais enfin*, *voilà quoi*, etc.). Nous ne nous intéressons ici qu'aux formes simples les plus représentées dans les corpus à notre disposition : *alors, bon, donc* et *quoi*.

Les DPs, seules ou dans une classe plus générale appelée *marqueurs discursifs*[i], ont fait l'objet d'un assez grand nombre d'analyses ces vingt dernières années, notamment à la suite du développement des études sur le discours et la langue parlée reposant sur des données authentiques. Ces analyses ont majoritairement porté sur des aspects sémantiques (y compris en diachronie[i]), pragmatiques et parfois syntaxiques, mais plus rarement sur les propriétés prosodiques. Bien que différant sur certains points, elles permettent de dresser un portrait général. Du point de vue sémantico-pragmatique, les DPs ne font pas partie du contenu propositionnel de l'énoncé-hôte, elles ont pour fonctions principales la gestion de l'interaction, la manifestation des attitudes mentales du locuteur – épistémiques ou affectives – et la structuration de l'énoncé en cours. Du point de vue syntaxique, les DPs peuvent occuper plusieurs positions dans l'énoncé-hôte, souvent à l'initiale ou en finale ; elles ne sont pas complément et rarement ajout[ii] d'un élément particulier de l'énoncé-hôte, pour des synthèses, cf. (Aijmer 2013 ; Dargnat, à par. ; Denturck, 2008).

Il existe un certain nombre de corpus de français oral spontané, plus ou moins accessibles et dans des formats divers. Les études portant sur les particules de discours ont généralement été faites sur des données authentiques orales, mais celles-ci sont circonscrites et le plus souvent personnelles et non diffusées ou non exploitables dans leur totalité, notamment du point de vue prosodique, e.g., (Brémond, 2004 ; Bastien, 2015; Bruxelles & Traverso, 2006 ; Castadot, 2014 ; Chanet, 2001 ; Col et al., 2015 ; Denturck, 2008 ; Hansen, 1998 ; Noda, 2011 ; Teston, 2006).

Notre but est d'étudier les relations potentielles entre l'interprétation des DPs et leurs propriétés prosodiques dans des corpus oraux variés. Cette problématique se retrouve dans de nombreux travaux, mais elle n'a pas fait l'objet d'une investigation systématique et quantifiée. Ceci implique un travail de sélection, d'homogénéisation et d'enrichissement de données existantes, de manière automatique et semi-automatique, en particulier l'alignement parole-transcription, l'annotation des valeurs sémantico-pramatiques et des propriétés prosodiques. Une première étude de corrélats prosodiques de quelques mots fréquemment utilisés comme particules de discours, ainsi qu'une première détection automatique de la fonction discursive, ou non, de ces mots à partir des corrélats prosodiques a été présentée dans (Dargnat et al., 2015). Toutefois cette étude a été menée sur le corpus ESTER2 (Galliano et al., 2009) qui correspond essentiellement à de la parole préparée. Entre temps le projet ORFEO a rassemblé une large variété de corpus de parole en français allant de la parole très préparées (contes) à de la parole spontanée (interviews et dialogues), voire très spontanée (interactions entre personnes). Ces corpus ont été alignés au niveau des mots dans le cadre du projet ORFEO. Ceci a suscité un intérêt à poursuivre l'étude des corrélats prosodiques des particules de discours sur ces corpus de parole. Ce papier présente donc l'état d'avancement des annotations de

particules de discours sur cette large variété de corpus de parole. La section 2 présente les corpus de parole et les annotations effectuées, ainsi que quelques statistiques sur les mots considérés. La section 3 détaille les annotations des fonctions discursives de ces mots.

# 2 Corpus et annotations

L'utilisation des corpus de parole du projet ORFEO[iii] et du corpus de parole de la campagne d'évaluation ESTER2 (Galliano et al., 2009) permet de couvrir une large gamme de types de parole correspondant à différents degrés de spontanéité, que l'on a regroupés en quatre catégories :

**Contes** : FRE (FRENCH ORAL NARRATIVE[iv]) est un corpus de contes oraux.

**Infos** : EST (ESTER2), est un corpus de journaux d'information radiodiffusés (Galliano et al., 2009). Il contient essentiellement de la parole préparée (e.g. journalistes), plus quelques interviews.

**Interviews, dialogues et conversations** : CFP (CFPP2000, Corpus de Français Parlé Parisien[v]), est composé d'interviews sur les quartiers de Paris et de la proche banlieue (Branca-Rosoff et al., 2000). COR, partie française du projet C-ORAL-ROM[vi] de parole spontanée dans quatre langues romanes, contient des enregistrements de dialogues et de conversations, et quelques parties plus formelles (Cresti et al., 2004). CRF (CRFP[vii], Corpus de Référence du Français Parlé), contient des enregistrements de parole correspondant à différentes situations de parole et niveaux d'études des locuteurs (Delic, 2004). TUF et VAL désignent respectivement la partie française du corpus TUFS[viii], et la partie du corpus VALIBEL[ix] mise à disposition du projet ORFEO.

**Interactions** : CLA et FLE désignent respectivement les parties mises à disposition du projet ORFEO des corpus CLAPI[x] (Corpus de LAngue Parlée en Interaction), et FLEURON[xi] ; ce dernier corpus correspond à des situations auxquelles les étudiants sont confrontés lors de leur arrivée dans une université française (interactions avec personnels de l'université, avec enseignants, …). TCO (TCOF[xii], Traitement de Corpus Oraux en Français), est un corpus d'interactions entre locuteurs. OFR (OFROM[xiii], Corpus Oral de français de Suisse Romande) contient des enregistrements d'interactions et d'interviews (Avanzi et al., 2012). DEC (DECODA) contient des dialogues anonymisés enregistrés lors d'appels au service clients de la RATP (Bechet et al., 2012). Finalement, HUS est un corpus d'enregistrements de réunions de travail.

Tous les corpus ont été enregistrés en France, excepté VAL (enregistré en Belgique) et OFR (enregistré en Suisse). Pour tous les corpus du projet ORFEO (i.e., tous les corpus ci-dessus, sauf ESTER2), nous avons utilisé les alignements parole-texte réalisés dans le cadre du projet ORFEO. Au total, plus de cinq millions de mots alignés sont disponibles, et le tableau 1 indique le nombre total de mots alignés pour chacun des corpus.

Un sous-ensemble d'occurrences des mots *alors*, *bon*, *donc* et *quoi* a été sélectionné aléatoirement dans chaque corpus (typiquement 50 à 100 occurrences dans chaque corpus), et chaque occurrence a été manuellement annotée suite à l'écoute d'un segment de parole couvrant l'occurrence du mot, plus une quinzaine de mots avant et autant après. Une interface a été développée pour faciliter l'écoute des différents extraits, la visualisation avec *Praat* du signal et de la segmentation, ainsi que pour la pose des étiquettes. L'annotation a consisté à indiquer si l'occurrence du mot correspond à une fonction discursive (DP) ou non. En cas de fonction discursive, une annotation détaillée de la valeur pragmatique est également indiquée (e.g., introduction, ré-introduction, conclusion, émotion, …). Quelques occurrences trop mal alignées ont été considérées comme incorrectes, et exclues de l'annotation manuelle. En vue d'une étude des corrélats prosodiques des DPs, la segmentation du mot aux niveaux segmental et supra-segmental a été corrigée lorsque nécessaire. Pour les quatre

mots considérés, le tableau 1 indique le nombre d'occurrences annotées de chacun de ces mots dans les corpus, ainsi que le nombre d'occurrences déjà annotées et la proportion d'utilisation en tant que DP. L'annotation de *quoi* a commencé récemment, et la sélection des étiquettes pragmatiques fines est en cours de finalisation.

Une annotation prosodique des données est également effectuée. En ce qui concerne la présence de pause avant ou après le mot, cela résulte de l'analyse de la segmentation en mots obtenue par l'alignement automatique parole-transcription, éventuellement après correction manuelle. Une segmentation du flux de parole en groupes intonatifs est réalisée avec le logiciel *ProsoTree* (Bartkova & Jouvet, 2013), qui exploite les inversions de pentes de la fréquence fondamentale (F0) décrites dans (Martin, 1986) et détermine les frontières des groupes intonatifs en se basant sur les pentes de F0, le niveau du F0 et la durée des voyelles. Diverses normalisations sont mises en œuvre. Pour la durée des voyelles, cela prend en compte les voyelles présentes dans les mots précédents et suivants. Pour le niveau du F0, la plage de F0 du locuteur concerné est estimée sur le fichier audio complet. Comme indiqué précédemment, lors de l'annotation manuelle, les segmentations erronées des mots étudiés sont corrigées. Une première analyse des corrélats acoustiques de quelques particules de discours a été présentée dans (Dargnat et al., 2015), mais sur de la parole préparée uniquement (e.g., ESTER2). D'où l'annotation en cours d'une large variété de corpus pour poursuivre l'étude sur divers types de parole, incluant divers niveaux de spontanéité.

| Corpus | | Contes | Infos | Interviews, conversations, … | | | | | Interactions, … | | | | | | *total* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FRE | EST | CFP | COR | CRF | TUF | VAL | CLA | FLE | TCO | OFR | DEC | HUS | |
| Nb. mots. (millions) | | 0.14 | 1.82 | 0.41 | 0.22 | 0.38 | 0.58 | 0.25 | 0.02 | 0.03 | 0.36 | 0.28 | 0.65 | 0.17 | *5.32* |
| ***alors*** | Fréq. occ. | 0,6% | 0,2% | 0,4% | 0,4% | 0,4% | 0,2% | 0,3% | 0,2% | 0,5% | 0,4% | 0,5% | 0,8% | 0,4% | --- |
| | Nb. annot. | 100 | 178 | 88 | 87 | 92 | 87 | 82 | 38 | 74 | 77 | 91 | 66 | 81 | *1144* |
| | **Fréq. DP** | 23% | 56% | 78% | 76% | 68% | 68% | 71% | 58% | 93% | 77% | 84% | 80% | 88% | *69%* |
| ***bon*** | Fréq. occ. | 0,1% | 0,1% | 0,4% | 0,3% | 0,5% | 0,5% | 0,4% | 0,5% | 0,3% | 0,5% | 0,2% | 0,4% | 0,5% | --- |
| | Nb. annot. | 91 | 186 | 75 | 90 | 84 | 84 | 79 | 82 | 72 | 81 | 93 | 83 | 71 | *1171* |
| | **Fréq. DP** | 60% | 59% | 87% | 80% | 90% | 75% | 90% | 41% | 61% | 94% | 86% | 84% | 82% | *75%* |
| ***donc*** | Fréq. occ. | 0,1% | 0,2% | 0,7% | 0,7% | 0,9% | 0,5% | 0,4% | 0,3% | 1,6% | 0,8% | 0,7% | 0,8% | 0,9% | --- |
| | Nb. annot. | 71 | 190 | 88 | 79 | 68 | 77 | 89 | 65 | 81 | 78 | 98 | 84 | 63 | *1131* |
| | **Fréq. DP** | 66% | 78% | 75% | 84% | 87% | 84% | 87% | 68% | 93% | 90% | 86% | 88% | 89% | *82%* |
| ***quoi*** | Fréq. occ. | 0,1% | 0,1% | 0,2% | 0,3% | 0,3% | 0,5% | 0,3% | 0,7% | 0,1% | 0,5% | 0,3% | 0,2% | *0,3%* | --- |
| | Nb. annot. | 50 | --- | 40 | 45 | 45 | 46 | 44 | 44 | 21 | 44 | 50 | 10 | 40 | *479* |
| | **Fréq. DP** | 20% | --- | 70% | 76% | 89% | 74% | 77% | 57% | 29% | 82% | 72% | 40% | 72% | *66%* |

TABLEAU 1 : Nombre de mots alignés, fréquence d'occurrence des mots sélectionnés, nombre d'occurrences annotées, et fréquence d'utilisation en tant que particule de discours (DPs).

# 3 Annotations de particules de discours

L'annotation des fonctions discursives s'est faite en deux temps. D'abord une distinction entre l'emploi du mot comme DP et non-DP, a priori assez facile (ex. pour *bon* et *quoi*), mais qui a nécessité l'explicitation de critères plus poussés pour *alors* et *donc*, fonctionnant également comme

connecteurs en français. Ensuite, une liste d'étiquettes plus fines correspondant à différentes acceptions des emplois comme DP a été mise en place. Cette deuxième étape pourrait paraître simple dans la mesure où des études sur chaque item ont déjà été faites et où il aurait suffit d'appliquer leurs classifications à nos données. En pratique, cela est plus difficile : la labellisation a consisté en un va-et-vient entre étiquettes préétablies dans la littérature et données concrètement analysées. La difficulté est de trouver un équilibre entre des étiquettes suffisamment générales pour s'appliquer à différents corpus, et suffisamment spécifiques pour respecter les propriétés distinctives des particules entre elles (ex. *bon* vs. *quoi*, vs. *alors*, vs. *donc*), et des valeurs pour chaque particule (*quoi* conclusif, *quoi* reformulateur, etc.).

Par manque de place, nous ne pouvons détailler ici l'ensemble des étiquettes retenues pour chaque item. Dans le tableau 3, nous résumons les différents emplois répertoriés pour *bon*, *donc* et *quoi* en français et développons par la suite la réflexion sur *alors* en illustrant les valeurs les plus fréquentes comme particule (tableau 2) : suspension d'une activité en cours ou mise en attente ; y compris hésitation (DP-susp) ; introduction ou réintroduction d'un thème (DP-intro) ; conclusion ou acceptation d'une idée (équivalent de *bon*) (DP-conclu) ; gestion de l'interaction (DP-interact).

| ***Item*** | ***Fonction*** | ***Annotation*** | ***Exemples*** |
|---|---|---|---|
| ***alors*** | Adv. temp. anaph. | Non-DP | … commencent à diviser les générations dans les années 50 et un peu plus tard les voisins on est **alors** en 1988 … `[Ester 20010726 0900 1000 rfi]` |
| | Connecteur | | … et si l'irak refuse **alors** réfléchissons euh ou euh plus que réfléchissons euh agissons pour euh utiliser d' autres moyens éventuellement la force … `[Ester 20001012 0930 1030 rfi]` |
| | Particule (DP) | DP-susp | L1-… non ce n'est pas celui-ci / L2- il faut que j'aille sur quoi alors ? / L1- sur le site RAPT / L2- **alors** je suis sur le site RATP / L1- RATP / L2- RATP OK / L1- alors ensuite vous cliquez … `[Decoda 20091112 RATP SCD 0711]` |
| | | DP-intro | … effet que maintenant euh la les forces régulières les forces loyalistes vont mettre le paquet sur bouaké *(pause)* **alors** la question qui qui se pose à la mi journée c'est de savoir qui … `[Ester 20020919_1400_1500_rfi]`<br>… mettez-le xxx vous allez voir vous allez bien entendre *(pause + changement d'interlocuteur)* Stanislas Nordey **alors** le rôle de Saint-François euh r- rôle écrasant a été créé par euh … `[Ester 20041007 0800 0900 INTER DGA]` |
| | | DP-conclu | … en achetant tout simplement des produits vous savez étiquetés satisfait ou remboursé **alors** c'est une gestion mais ça marche il l'a prouvé il a rempli son frigo … `[Ester 19991026 0700 0800 inter]` |
| | | DP-interact | *Réponse*<br>L1- … et vous pensez l'avoir perdu où madame ? / L2- **alors** euh j'ai deux endroits possibles alors je sais que je l'ai passé au à le au métro … `[Decoda 20091112_RATP_SCD_0989]`<br>*Relance (question)*<br>L1- j'ai 10 ans / L2- est-ce que tu as préparé tes affaires / L3. Oui (chuchoté) / L3- **alors** qu'est-ce que tu as préparé comme affaires ? `[Ester 19981207_0700_ 0800_inter_fm_dga]`<br>*Reprise thématique*<br>L1- … un document conjoint au moins pour le l'arrêt de la violence aujourd'hui / L2- **alors** l'arrêt de la violence et ce que cela suppose … `[RATP SCD 0989]` |

TABLEAU 2 : Etiquettes retenues et exemples pour l'annotation du mot *alors.*

| *Item* | *Fonction* | *Annotation* | *Exemples* |
|---|---|---|---|
| ***bon*** | Adjectif | Non-DP | … ça devait pas être euh grand-chose hein parce qu'il y avait euh vraiment un **bon** départ dans le dos me semble-t-il mais euh ça doit se jouer … [CORALROM fmedsp01] |
| | Particule (DP) | DP | … c'est pour savoir combien il y a de voyageurs qui montent dans l'autobus **bon** et là aujourd'hui je passe il me dit validez votre coupon alors moi … [DECODA 20101206 RATP SCD 0104] |
| ***donc*** | Adverbe (coord.)[xiv] | Non-DP | … ne sera pas atteint par ces frappes au contraire on peut penser qu'il va renforcer son pouvoir **donc** les américains ont agi un petit peu pour montrer qu'ils pouvaient agir … [Ester 19981217 0700 0800 inter fm dga] |
| | Particule (DP) | DP | … le travail de l'acteur soutenu le plus souvent par une ingénieuse scénographie Lev Dodine a reçu **donc** à Taormina le huitième prix européen du théâtre sa réaction … [Ester 20000414 0930 1030 rfi fm dga] |
| ***quoi*** | Pronom | Non-DP | … c'était **quoi** le CEG ? on allait au CEG après … [CFPP Mathilde Lelong F 85 Marie Louise Orsin F 64 11e] |
| | Particule (DP) | DP | … euh la diversité euh tout ce qui fait qu'une ville peut euh peut s'enrichir **quoi** … [CFPP_Andre_Morange_H_58_Mo] |

TABLEAU 3 : Annotations et exemples pour les mots *bon, donc et quoi.*

# 4 Conclusion

Ce papier a présenté les annotations en cours de quelques mots fréquemment utilisés en tant que particules de discours en français. Une large variété de corpus de parole correspondant à différents niveaux de spontanéité sont considérés, allant de la parole préparée (contes et journaux d'information radiodiffusés) à la parole spontanée (interviews et interactions entre personnes). Deux niveaux d'annotation sont considérés, d'abord l'utilisation du mot en tant que particule de discours ou pas, et ensuite la fonction discursive détaillée dans le cas des occurrences utilisées en tant que particules de discours. En vue d'une étude des corrélats prosodiques, les segmentations des mots aux niveaux segmental et supra-segmental ont été corrigées lorsque nécessaire, et des paramètres prosodiques ont été calculés, de même que la position du mot dans son groupe intonatif. Ces corpus serviront à l'analyse des corrélats prosodiques et à l'étude de la détection / classification automatique des fonctions discursives. Il est prévu de diffuser, pour les données annotées, les annotations des fonctions discursives, les annotations prosodiques effectuées, et aussi les extraits de signal de parole lorsque cela sera possible.

# Remerciements

Ce travail est mené dans le cadre de l'opération ProsodCorpus du CPER LCHN (Contrat Plan Etat Région "Langues, Connaissances et Humanités Numériques"). Les annotations automatiques ont été effectuées sur Grid5000, qui bénéficie du support d'un groupe d'intérêt scientifique hébergé par Inria, et incluant le CNRS, RENATER et quelques universités et organisations (voir https://www.grid5000.fr).

# Références


AIJMER K. (2006), *Understanding Pragmatic Markers. A variational pragmatic approach*, Edinburgh: Edinburgh UP.

AVANZI M., BEGUELIN M.-J., DIEMOZ F. (2012-2015), *Présentation du corpus OFROM – corpus oral de français de Suisse romande*. Université de Neuchâtel, Switzerland.

AUCHLIN A. (1981), « *Mais heu, pis bon, ben alors voilà, quoi* ! Marqueurs de structuration de la conversation et complétude », *Cahiers de linguistique française* 2, pp.141-160.

BARTKOVA K., JOUVET D. (2013), "Automatic detection of the prosodic structures of speech utterances", *SPECOM 2013, 15th, Int. Conf. on Speech and Computer*. pp. 1-8. Springer, Berlin.

BASTIEN A. (2015), *Prosodie et fonctions pragmatiques des marqueurs discursifs à base verbale.* Master Dissertation, Université of Lorraine.

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., DE MORI R., ARBILLOT E. (2012), "DECODA: a call-centre human-human spoken conversation corpus", *LREC 2012, 8th Int. Conf. on Language Resources and Evaluation*, Istanbul, Turkey.

BOUCHARD D. (2002), « *Alors, donc, mais...*, particules énonciatives et/ou connecteurs ? Quelques considérations sur leur emploi et leur acquisition », *Syntaxe et sémantique,* 1/3, 63-73.

BRANCA-ROSOFF S., FLEURY S., LEFEUVRE F., PIRES M., *Discours sur la ville. Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000)*.

BRÉMOND C. (2002), *Les petites marques du discours. Le cas du marqueur métadiscursif* bon *en français*, Ph.D Dissertation, Université d'Aix–Marseille I.

BRUXELLES S., TRAVERSO V. (2006), « Usages de la particule 'voilà' dans une réunion de travail : analyse multimodale ». In M. Drescher, B. Job (eds), *Les marqueurs discursifs dans les langues romanes: approches théoriques et méthodologique*s, Peter Lang, pp. 71-92.

CASTADOT F. (2014). *Le pragmatème* alors *et ses quasi-synonymes dans l'oral contemporain*. Master Dissertation, University of Lorraine.

CHANET C. (2001), « 1700 occurrences de la particule *quoi* en français parlé contemporain: approches de la "distribution" et des fonctions en discours », *Marges Linguistiques* 2, pp. 52-80.

COL G., DANINO C., RAULT J. (2015), « Éléments de cartographie des emplois de *voilà* en vue d'une analyse instructionnelle », *Revue de Sémantique et Pragmatique* 37, pp. 37-60.

CRESTI E., DO NASCIMENTO F. B., MORENO-SANDOVAL A., VERONIS J., MARTIN P., CHOUKRI K. (2004), "The C-ORAL-ROM CORPUS. A Multilingual Resource of Spontaneous Speech for Romance Languages", *LREC 2004, 4th Int. Conf. on Language Resources and Evaluation*, Lisbon, Portugal

DARGNAT M. (à par.), « Particules et interjections ». In A. Abeillé et D. Godard (eds), *Grande Grammaire du Français*.

DARGNAT M., BARTKOVA K., JOUVET D. (Nov. 2015), "Discourse Particles In French: Prosodic Parameters Extraction and Analysis", *SLSP 2015, Int. Conf. on Statistical Language and Speech Processing*, Budapest, Hungary.

DEGAND L., FAGARD B. (2011), "*Alors* between Discourse and Grammar: the Role of Syntactic Position", *Function of Language* 18, pp. 19-56.

DEGAND L., CORNILLIE B., PIETRANDREA A. (eds) (2013), *Discourse Markers and Modal Particles: Categorization and description*, Amsterdam/Philadelphia: Benjamins.

Équipe DELIC (2004), « Autour du Corpus de référence du français parlé », *Recherches sur le français parlé*, n° 18, Publications de l'université de Provence, 265 p.

DENTURCK E. (2008), *Étude des marqueurs discursifs : l'exemple de "quoi"*. Master Dissertation, Gent University.

DOSTIE G. (2004), *Pragmaticalisation et marqueurs discursifs*, Liège : De Boeck/Duculot.

FERNANDEZ-VEST J. (1994), *Les particules énonciatives dans la construction du discours*. PUF.

FISCHER K. (ed.) (2006), *Approaches to Discourse Particles,* Elsevier: Amsterdam.

GADET F. (2003). *La variation sociale en français,* Paris : Ophrys.

GALLIANO S., GRAVIER G., CHAUBARD L. (2009). "The ESTER 2 evaluation campaign for rich transcription of French broadcasts", *INTERSPEECH 2009, 10th Annual Conf. of the Int. Speech Communication Association*, Brighton, UK, pp. 2583-2586.

HANSEN M.-B. M. (1998), *The Function of Discourse Particles*, Amsterdam: Benjamins.

KOCH P., OESTERREICHER W. (2001), « Langage oral et langage écrit ». In *Lexikon der romanistischen Linguistik*, Tübingen: Niemeyer. Tome 1-2, pp. 584-627.

MARTIN P. (1987), "Prosodic and rhythmic structures in French", *Linguistics*, vol.25, pp. 925-949.

MÉTRICH R., FAUCHER E., COURDIER G. (2002), *« Invariables Difficiles », Dictionnaire allemand-français des particules, connecteurs, interjections et autres « mots de la communication »*, ATILF.

NODA H. (2011), *Intersubjectivité : modulation et ajustement. Cas des marqueurs discursifs* hein*,* quoi*,* n'est-ce pas *en français et* darô*,* yo*,* yone *en* japonais. Thèse de Doctorat de l'Université de Franche-Comté.

PAILLARD D. (1998), « Les mots du discours comme mots de langue », *Le Gré des langues* 14, pp.10-41.

TESTON-BONNARD S. (2006), *Propriétés topologiques et distributionnelles des constituants non régis. Application à une description syntaxique des particules discursives*, Ph.D. Dissertation, University of Provence.

---

[i] Même s'il est très intéressant du point de vue lexical (processus de *grammaticalisation* et de *pragmaticalisation*), nous laissons de côté cet aspect pour le moment. Cf. Degand, L., Evers-Vermeul, J. (2015), « Grammaticalization or pragmaticalization of discourse markers ? More than a terminological issue », *Journal of Historical Pragmatics* 16, pp. 59-85.

[ii] En général, les DPs sont des ajouts à l'énoncé tout entier, ou, pour certaines, peuvent s'employer seules (les interjections et les adverbes comme *d'accord, hélas, oui, bon*, etc.)

[iii] ORFEO project: http://www.projet-orfeo.fr/

[iv] French oral narrative: http://frenchoralnarrative.qub.ac.uk

[v] CFPP2000 : http://cfpp2000.univ-paris3.fr/

[vi] C-ORAL-ROM: http://lablita.dit.unifi.it/corpora/descriptions/coralrom/

[vii] CRFP : http://www.up.univ-mrs.fr/delic/corpus/index.html

[viii] TUFS : http://www.tufs.ac.jp/ts/personal/ykawa/art/2014_Waseda_Corpus_TUFS.pdf

[ix] Valibel: http://www.uclouvain.be/81834.html

[x] CLAPI. http://clapi.ish-lyon.cnrs.fr/

[xi] FLEURON : https://apps.atilf.fr/fleuron2/

[xii] TCOF: http://www.cnrtl.fr/corpus/tcof/

[xiii] OFROM: http://www.unine.ch/ofrom

[xiv] *Donc* est classé comme conjonction de coordination dans les grammaires scolaires, mais fonctionne en réalité comme un adverbe.