Corela
Cognition, représentation, langage

HS-21 | 2017
Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage

Expériences sur l'analyse morphosyntaxique des corpus oraux avec l'annotateur multi-niveaux DisMo

George Christodoulides et Giulia Barreca

Éditeur
Cercle linguistique du Centre et de l'Ouest - CerLICO

Édition électronique
URL : http://corela.revues.org/4867
ISSN : 1638-573X

Référence électronique
George Christodoulides et Giulia Barreca, « Expériences sur l'analyse morphosyntaxique des corpus oraux avec l'annotateur multi-niveaux DisMo », *Corela* [En ligne], HS-21 | 2017, mis en ligne le 16 février 2017, consulté le 21 février 2017. URL : http://corela.revues.org/4867

Ce document a été généré automatiquement le 21 février 2017.

Expériences sur l'analyse morphosyntaxique des corpus oraux avec l'annotateur multi-niveaux DisMo

George Christodoulides et Giulia Barreca

Introduction

1 L'annotation des corpus oraux présente des défis particuliers, liés aux caractéristiques de la langue parlée et sa transcription. Tandis qu'un système d'annotation automatique conçu pour des corpus écrits peut se baser sur la ponctuation pour segmenter le texte en entrée, et procéder à l'analyse syntaxique sur la base des phrases graphiques, il n'est pas possible d'appliquer cette démarche à l'oral. L'absence de ponctuation et par conséquent les unités de segmentation multiples, ainsi que les disfluences et une syntaxe souvent « non-canonique » (avec des reformulations, des unités averbales ou elliptiques etc.) rendent la tâche de l'analyse morphosyntaxique de l'oral plus complexe. Si la méthodologie d'analyse et les outils d'annotation automatique doivent être adaptés à cette réalité, il est toutefois souhaitable de garder la possibilité de comparer un corpus oral avec un corpus écrit, sur base d'un « dénominateur commun », et d'enrichir l'annotation avec des couches supplémentaires pour décrire les phénomènes propres à l'oral.

2 C'est dans cette logique que nous avons proposé dans (Christodoulides et al. 2014) l'outil *DisMo*, un annotateur automatique conçu spécifiquement pour les corpus oraux, qui fournit une analyse multi-niveaux : étiquetage morphosyntaxique, lemmatisation, détection des unités poly- lexicales, détection et annotation des phénomènes de disfluence et des marqueurs de discours, et découpage en unités syntaxiques minimales (*chunks* : Abney, 1991). Dans cet article nous présenterons nos travaux sur le corpus

Phonologie du Français Contemporain (PFC) (Durand et al. 2009), travaux qui ont permis de réviser l'outil et d'améliorer sa performance. Nous présenterons en détail les choix théoriques et pratiques quant aux niveaux d'annotation, aux phénomènes détectés, aux jeux d'étiquettes, et nous terminerons par une évaluation de la performance.

3 Des études antérieures sur l'annotation morphosyntaxique des corpus oraux ont eu recours à des étiqueteurs conçus pour l'écrit, et adaptés à la suite d'un prétraitement des transcriptions des données orales (Blanc et al. 2008), ou d'un ajustement du corpus (Valli et Véronis, 1999). Certains auteurs ont également opté pour des solutions qui comportent un apprentissage automatique à partir de corpus oraux corrigés manuellement (Eshkol et al. 2010) et l'utilisation de fichiers de paramétrage spécifiques pour les données orales (Benzitoun et al. 2012).

4 Trois grands corpus de référence du français parlé ont été annotés à l'aide de *DisMo* : le corpus Phonologie du Français Contemporain (PFC) (Durand et al. 2009) (1,4 millions de tokens), la collection des corpus du centre VALIBEL (Simon et al. 2014) (environ 6 millions de tokens), et le Corpus Oral de français de Suisse Romande (OFROM) (0,5 millions de tokens) (Avanzi et al. 2012). Les modèles statistiques de *DisMo* ont été entraînés d'abord sur le corpus CPROM-PFC (Avanzi 2014), lui-même contenant des échantillons du corpus PFC. Deux annotateurs experts ont corrigé chacun une partie de l'annotation de 57 mille tokens issus de ces deux corpus. Ces premiers modèles ont été utilisés pour annoter 127 mille tokens du corpus PFC, et un annotateur expert a alors corrigé manuellement cet échantillon. L'ensemble de ces données a enfin permis d'entraîner des modèles statistiques. Après ces campagnes d'annotation et de correction, l'exactitude (anglais : *accuracy*) du système est de l'ordre de 97 % pour le jeu d'étiquettes complet, et 98 % pour un jeu d'étiquettes simplifié. Il faut toutefois noter que cette exactitude représente la meilleure performance calculée selon la méthode de validation croisée sur les échantillons du corpus PFC, qui représentent le même genre discursif (enquête sociolinguistique). L'annotation d'un autre genre discursif, comme par exemple les dialogues « map task », pourrait présenter des défis supplémentaires qui influenceront la performance : cette question dépasse la portée de la présente étude.

Annotation multi-niveaux

5 L'annotateur *DisMo* accepte plusieurs types d'entrées : l'analyse complète se base sur une transcription orthographique alignée au signal de la parole, au moins au niveau de l'énoncé. Dans ce cas, le système prend en compte dans son analyse des paramètres prosodiques calculés automatiquement (pauses silencieuses, débit de parole et mouvements mélodiques), notamment pour l'identification des disfluences et des unités de segmentation. Un alignement d'unités à un niveau plus fin (par exemple au niveau des mots ou même des syllabes) permet d'améliorer la performance. Il est aussi possible d'annoter une transcription non alignée, ou même un texte écrit. Le système est capable d'annoter des interactions impliquant plusieurs locuteurs : les tours de parole sont considérés comme des indices de segmentation. Des scripts facilitent le traitement des transcriptions selon les conventions de transcription (par exemple le symbole pour indiquer une amorce lexicale, etc.). Les formats que le système accepte en entrée sont des fichiers *Praat TextGrid*, *TranscriberAG*, *ELAN*, *Exmaralda Partitur*, ou des fichiers texte. *DisMo* peut ajouter des tiers avec les résultats d'annotation et stocker des fichiers dans les formats mentionnés, produire en sortie des fichiers XML et OpenDocument, ou effectuer

des modifications dans une base de données relationnelle (SQL, selon le schéma du logiciel *Praaline*, cf. Christodoulides, 2014). Les annotations sont structurées sur trois niveaux principaux, auxquels s'ajoutent des propriétés :

Niveau d'analyse		Propriétés
`tok-min`	Unités lexicales (tokens) minimales	`pos-min` : étiquette morphosyntaxique `pos-ext-min` : information morphosyntaxique supplémentaire `lemma-min` : lemme `disfluency` : annotation des phénomènes de disfluence
`tok-mwu`	Unités polylexicales	`pos-mwu` : étiquette morphosyntaxique `pos-ext-mwu` : information morphosyntaxique supplémentaire `lemma-mwu` : lemme
`chunks`	Unités syntaxiques minimales	`chunk` : étiquette d'unité syntaxique minimale `discourse` : indique les connecteurs et les marqueurs du discours potentiels

Tableau 1 : Niveaux d'annotation et propriétés

6 L'articulation de l'annotation sur plusieurs niveaux (couches) présente des avantages (par rapport à l'analyse qu'on obtient avec un annotateur comme par exemple *TreeTagger*, Schmid, 1994), que nous allons illustrer avec des exemples extraits du corpus PFC. Il est possible de représenter correctement le fait qu'une unité polylexicale, prise comme un ensemble, ait une fonction morphosyntaxique différente de la catégorie des unités qui la constituent :

> « je n'ai pas pu obtenir de poste à Lyon tout de suite (pos-mwu ADV) donc j'ai été exilé » (PFC Lyon ; 69aag1gg)
> « Ils vendent tous les objets euh que des objets euh russes ça c'est (pos-mwu INTROD) fabrication russe. Tous ce qui était objet russe oh il y en avait (pos-mwu INTROD) certains c'était (pos-mwu INTROD) euh donc des fournisseurs de Paris » (PFC, Aveyronnais à Paris, 75xlv1lg)

7 Dans l'exemple (1), la locution adverbiale « tout de suite » sera annotée comme ADV sur le niveau des unités polylexicales tandis que les unités minimales correspondantes sont annotées « tout/ADV de/PRP suite/NOM :com » sur le niveau pos-min. De même, dans l'exemple (2), « c'est », « il y en avait » et « c'était » sont annotés comme des introducteurs sur le niveau pos-mwu, tout en gardant l'annotation morphosyntaxique de leurs constituants dans la couche pos-min (par exemple « c'/PRO :per :sjt était/ VER :impf », « il/PRO :per :sjt y/PRO :per :obji a/VER :pres »). Le même traitement en deux niveaux s'applique à plusieurs types d'unités polylexicales : mots composés, locutions adverbiales, numéraux, etc. L'annotation multi-niveaux permet aussi de bien faire la distinction entre la fonction pragmatique et la catégorie morphosyntaxique originale d'un token ou d'une suite de tokens :

> 1. « d'entrée un concours c'est-à-dire qu'on a été plusieurs et il y en a qu' ont pas été pris quoi (pos-min ITJ ; discourse : MD) on a été genre une dizaine ils en ont pris genre six tu vois (pos-mwu : VER :pres ; discourse : MD) un truc comme ça » (PFC Paris ; 75cab1gg)

8 L'exemple ci-dessous montre que « bon », « quoi » et « tu vois » seront annotés comme des marqueurs du discours sur le niveau d'annotation approprié (discourse/chunks), tout en gardant leur catégorie morphosyntaxique originale. Ce choix facilitera par exemple

une étude statistique sur les catégories morphosyntaxiques et les contextes syntaxiques des marqueurs de discours. Dans les sections suivantes, nous présentons l'étiquetage proposé pour la morphosyntaxe, les disfluences et les chunks.

Annotation morphosyntaxique – Comparaison des jeux d'étiquettes

9 Le jeu d'étiquettes morphosyntaxiques de *DisMo* est composé de 12 catégories principales : verbes (VER), noms (NOM), adjectifs (ADJ), adverbes (ADV), conjonctions (CON), déterminants (DET), pronoms (PRO), numéraux (NUM), prépositions (PRP), interjections (ITJ), préfixes (PFX) et mots étrangers (FRG). La plupart des étiquettes s'articulent à deux ou trois niveaux. Par exemple, la classe des déterminants comporte les déterminants définitifs (DET :def), démonstratifs (DET :dem), indéfinis (DET :ind), interrogatifs (DET :int), exclamatifs (DET :exc), partitif (DET :par) et possessifs (DET :pos). Le temps et le mode des verbes sont indiqués au deuxième niveau des étiquettes.

10 Le tableau 2 suivant présente une comparaison des jeux d'étiquettes proposés par DisMo et par d'autres étiqueteurs, afin d'évaluer les choix adoptés par chaque système, mais aussi afin d'établir des correspondances permettant le croisement de deux ou plusieurs annotations sur les mêmes données.

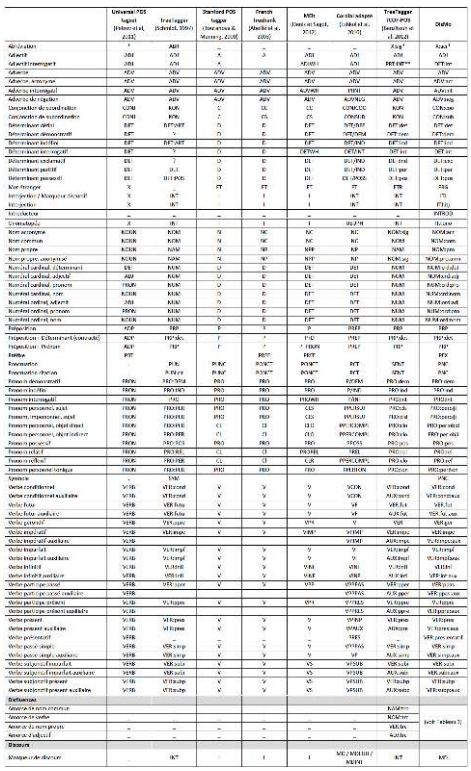

Tableau 2 : Comparaison des jeux d'étiquettes entre notre système (DisMo), le Univeral POS Tagset, le French Treebank et les annotateurs Treetagger, Stanford POS tagger, MElt, Cordial adapté et Treetagger TCOF-POS.

11 Ces étiquettes sont attribuées à chaque unité lexicale minimale (tok-min : pos-min) et chaque unité polylexicale (tok-mwu : pos-mwu). L'étiquette supplémentaire INTROD est utilisée pour les introducteurs et ne peut apparaître que sur le niveau des unités polylexicales. La lemmatisation est réalisée à l'aide des ressources linguistiques (un

dictionnaire), qui fournit aussi des informations morphosyntaxiques supplémentaires pour chaque forme : son genre, nombre et personne, selon le cas. Pendant l'annotation morphosyntaxique, les formes sont désambiguïsées et l'information supplémentaire est indiquée par DisMo dans les attributs « pos-ext ». Il faut noter que ce processus n'essaye pas d'appliquer des règles d'accord : toutes les possibilités sont données (par exemple, on retrouvera l'étiquette « pos-ext = ms :fs » pour une forme qui peut être soit du masculin, soit du féminin singulier).

12 On observe que les étiquettes complètes de DisMo sont les plus précises parmi les systèmes comparés dans le Tableau 2. En même temps, les catégories principales (premier niveau des étiquettes morphosyntaxiques) correspondent globalement aux catégories proposées par le projet *Universal Part-of-Speech Tagset* (Petrov et al, 2011) qui a cherché un jeu d'étiquettes suffisamment générique pour décrire 22 langues différentes et a mis à disposition plusieurs corpus annotés avec ces étiquettes.

Annotation des disfluences

13 Nous proposons également un protocole détaillé pour l'annotation des phénomènes de disfluence (Christodoulides & Avanzi, 2015), qui s'ajoute aux autres niveaux d'annotation de DisMo. Notre système d'annotation a été inspiré par les systèmes décrits dans (Shriberg, 1994), (Heeman et al, 2006) et (Brugos & Shattuck-Hufnagel, 2012), ainsi que le protocole d'annotation des disfluences proposé par le Linguistic Data Consortium et appliqué sur le corpus anglais *Switchboard* dans le cadre du projet MDE (Mateer & Taylor, 1995) et les propositions du projet français VoxForge (Clavel et al, 2013).

14 Notre système d'annotation regroupe les phénomènes de disfluence sous quatre catégories : disfluences simples (de nature essentiellement phonétique et prosodique), répétitions, disfluences structurées, et disfluences complexes. Les étiquettes pour l'annotation des disfluences sont rattachées aux unités lexicales minimales (tok-min), ou éventuellement sur les syllabes, si le corpus comporte une transcription phonétique et une syllabation.

15 On définit les disfluences **simples** comme étant les phénomènes de disfluence qui affectent seulement un token minimal : les **pauses remplies** (hésitations autonomes et voyelles épenthétiques), les **allongements** liés à une hésitation, les **amorces lexicales** où l'articulation d'un mot est interrompue par le locuteur avant la fin de ce mot, et les **pauses à l'intérieur** d'un token minimal (mot). Dans les grands corpus oraux du français, ils existent plusieurs conventions de transcription pour ces phénomènes. Des tokens spéciaux (« euh », « euhm » etc.) sont généralement utilisés pour la transcription des hésitations autonomes ; toutefois, il est rare de trouver un corpus où la distinction entre une hésitation autonome et un schwa allongé en position finale soit faite de manière systématique. Les amorces lexicales sont le plus souvent indiquées avec une transcription orthographique approximative, indiquant l'interprétation de la part du transcripteur du mot qui aurait été prononcé, suivie d'un caractère spécial (comme le tiret ou la barre oblique).

16 Les **répétitions** sont des tokens ou des suites de tokens qui sont répétés par le locuteur sans aucun changement de forme, éventuellement en alternance avec des pauses silencieuses ou remplies, à condition que cette répétition n'ait pas une fonction grammaticale ou emphatique. Par exemple, dans l'énoncé « *le le le chien* », la répétition de

l'article défini « le » sera annoté comme une disfluence. Par contre, dans les énoncés « oui oui », « voilà voilà », ou « *il est très très joli* », les mots « oui », « voilà » et « très » sont répétés intentionnellement (étant des marqueurs du discours, ou pour insister) et ne seraient pas annotés comme des disfluences (cf. Dister, 2014). Il n'est pas toujours facile d'appliquer cette distinction de manière cohérente : pour cette raison, l'étiquetage automatique proposé par DisMo annote tous ces cas comme des répétitions (en d'autres termes, dans le cas des répétitions, nous avons privilégie le rappel plutôt que la précision).

17 Les disfluences **structurées** incluent :

18 – les **substitutions**, où le locuteur revient sur une partie de son énoncé pour modifier (substituer) certains éléments (tokens), gardant la même structure syntaxique. Par exemple : *normalement je louais enfin je loue toujours un appart*,

19 – les **insertions**, où le locuteur revient sur une partie de son énoncé pour ajouter (insérer) certains éléments (tokens), gardant la même structure syntaxique. Par exemple : *c'est vrai que Béthune vivre à Béthune...*

20 – et les **suppressions / interruptions syntaxiques**, où le locuteur arrête son énoncé avant d'achever une structure syntaxique et recommence avec une autre, nouvelle structure syntaxique. Par exemple : *c'est vraiment en tout cas la parole...*

21 Les répétitions, ainsi que les trois types des disfluences structurées, rentrent dans le schéma suivant (proposé par Shriberg, 1994) :

> *(reparandum) * point d'interruption (interregnum, marqueurs d'édition éventuels) (reparans).*

22 Le **reparandum** est la partie de l'énoncé qui est répété ou qui sera corrigé, modifié ou abandonné. Le **point d'interruption** est la frontière qui délimite la fin du reparandum et le début de l'interregnum. Ce point ne coïncide pas forcement avec l'instant où le locuteur a remarqué un souci ou son intention de modifier ses propos. L'**interregnum** est la partie entre le reparandum et le reparans : il consiste souvent des pauses (silencieuse, remplie, ou les deux en série) et, éventuellement, des marqueurs d'édition, c'est-à-dire des marqueurs du discours utilisés par le locuteur pour signaler son intention de modifier ce qui a été dit (par exemple « enfin », « ben » etc.). Le **reparans** est la suite du message du locuteur qui suit la disfluence, et selon (Shriberg, 2001), si on enlève les deux premières parties, le reste de l'énoncé serait fluide du point de vue syntaxique. Cette catégorisation des phénomènes de disfluence, ainsi que les étiquettes correspondantes utilisées dans le schéma d'annotation de DisMo, sont synthétisés dans le Tableau 3.

Niveau 1 : Les **disfluences simples** qui affectent un seul token		
FIL	Pauses remplies	c' est pour ça que j' hésite euh un peu en parler FIL
LEN	Allongement lié à une hésitation	au cercle d'oenologie de= Bruxelles LEN
FST	Amorce lexicale	comme infirmière so/ sociale FST
WDP	Pause intra-mot	il m' a dit ça su+ _ +ffit WDP
Niveau 2: Les **répétitions** d'un ou plusieurs tokens à l'identique		
REP	Répétition	les disques et et lancer les jingles REP* REP_
		il a il a il a dit que REP:1 REP:2 REP:1 REP*:2 REP_ REP_
		c' est pas c' est pas un système génial REP:1 REP:2 REP*:3 REP_ REP_ REP_
Niveau 3: Les disfluences **structurées**, dites d'édition		
DEL	Suppression	c' est vraiment un en tout cas la parole DEL DEL DEL DEL*
SUB	Substitution	cette personne était enfin c' est un ami de SUB* SUB:edt SUB_ SUB_
INS	Insertion	c' est vrai que Béthune euh vivre à Béthune a été INS* INS+FIL INS_ INS_ INS_
Niveau 4: Disfluences complexes qui sont une combinaison des disfluences simples et structurées (mais qui ne peuvent pas être décrites en combinant les niveaux 1 à 3)		
COM	Complexe	les ac/ les actions enfin les activités enfin professionnelles COM COM COM COM COM COM COM COM COM

Tableau 3 : Schéma d'annotation des disfluences dans *DisMo*

23 Dans le cas des répétitions, les tokens répétés sont numérotés en série : le premier token de la séquence répétée est annoté REP :1, le deuxième REP :2 et ainsi de suite. De cette manière, l'utilisateur de l'annotation peut facilement repérer la structure et la longueur d'une répétition. Dans le cas des répétitions et des disfluences structurés (codes REP, SUB, INS et DEL), on ajoute des suffixes aux étiquettes pour indiquer les trois régions de la structure de la disfluence : le point d'interruption est indiqué par un astérisque (*), les marqueurs d'édition sont indiqués par le suffixe « :edt » et les tokens qui font partie du reparans sont indiqués avec le caractère « _ ». On n'annote pas un reparans dans le cas des interruptions syntaxiques, car tout l'énoncé qui suit pourrait être considéré comme le reparans.

24 Le système d'annotation est **hiérarchique**, c'est-à-dire que les étiquettes du niveau 1 (disfluences simples) peuvent se combiner avec celles du niveau 2 et 3, ainsi que les étiquettes du niveau 2 (répétitions) peuvent se combiner avec celles du niveau 3 (disfluences structurées). Cela permet une annotation facile de certaines cooccurrences de phénomènes de disfluence qui sont très fréquents dans les corpus oraux : dans le Tableau 3, on peut observer par exemple que la production d'une pause remplie (hésitation) dans l'interregnum de l'insertion est annotée INS+FIL. Dans la Figure 1, on présente des exemples qui illustrent comment cette catégorisation hiérarchique des phénomènes de disfluence est appliquée selon le protocole d'annotation *DisMo*.

1

token	le	_	euhm	le
disf-lex		SIL	FIL	
disf-rep	REP*	REP	REP	REP_
disf-struct				

2

token	à	_	à	aux
disf-lex		SIL		
disf-rep	REP*	REP	REP_	
disf-struct	SUB	SUB	SUB*	SUB_

3

token	portez-	vous	toujou/	_	euh	_	portez-	vous	tous	les	vêtements
disf-lex			FST	SIL	FIL	SIL					
disf-rep											
disf-struct	SUB	SUB	SUB*	SUB	SUB	SUB	SUB_	SUB_	SUB_		

4

token	qui	pose	débat	_	qui	pose	pas	mal	de	débats
disf-lex				SIL						
disf-rep										
disf-struct	INS	INS	INS*	INS	INS_	INS_	INS_	INS_	INS_	INS_

5

token	y	aura	même	des	des	gens	qui	vont	_	enfin	un	néerlandophone	et	un	fran/	un	francophone	qui	vont	gagner	encore	plus
disf-lex															FST							
disf-rep				REP*	REP_									REP*	REP	REP_						
disf-struct				SUB	SUB	SUB	SUB	SUB*	SUB	SUB:edt	SUB_	SUB_	SUB_	SUB_	SUB_	SUB_	SUB_	SUB_	SUB_			

Figure 1 : Annotation hiérarchique des phénomènes de disfluence utilisant le protocole DisMo

25 On note qu'il est possible de stocker et analyser les trois niveaux d'annotation des disfluences à l'aide d'attributs d'annotation séparés : ce système (illustré sur la Figure 1) est souvent plus intuitif pour un annotateur humain. Dans l'exemple (1) une répétition est combinée avec une pause silencieuse et une pause remplie. Dans l'exemple (2), la répétition « à à » est enchâssée à l'intérieur du reparandum d'une substitution. Dans l'exemple (3), la substitution est combinée avec une amorce lexicale (au point d'interruption), suivie par une pause silencieuse et une pause pleine qui se trouvent à l'intérieur de l'interregnum. L'exemple (4) clarifie la notion d'une insertion : le locuteur ajoute l'élément « pas mal de » tout en gardant la même structure syntaxique globale. L'exemple (5), où une longue substitution se combine avec deux répétitions (à l'intérieur du reparandum et du reparans) montre que le protocole d'annotation permet un codage flexible avec un nombre restreint d'étiquettes.

26 Malgré la flexibilité du système de codage hiérarchique, il s'avère que certains phénomènes de disfluence syntaxique ne peuvent pas être décrits selon le schéma reparandum-interregnum-reparans : il s'agit des **disfluences complexes**. Pour l'annotation des disfluences complexes on adopte la proposition de (Heeman et al. 2006), selon laquelle les interruptions et les correspondances entre tokens sont présentés de manière tabulaire. Ce système est d'ailleurs similaire à la proposition des « piles » de (Gerdes & Kahane, 2009).

27 Afin de faciliter l'annotation manuelle et la correction de l'annotation automatique des phénomènes de disfluence, nous avons développé un outil informatique d'édition, présenté sur la Figure 2. L'outil permet à l'utilisateur de choisir une suite de tokens et indiquer le type de disfluence : l'outil insère le codage approprié selon le protocole et vérifie que toute correction reste dans le cadre du protocole. Cet outil d'édition est disponible sous le logiciel *Praaline*.

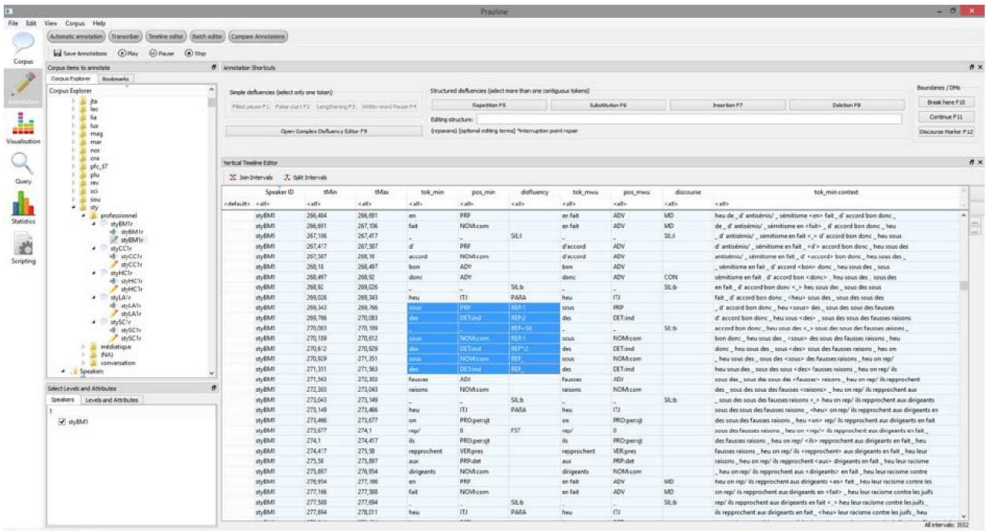

Figure 2 : Outil d'annotation des disfluences pour faciliter l'application du protocole DisMo

Annotation des unités syntaxiques minimales (chunks)

28 Les *chunks* sont des constituants continus et non-récursifs (Abney, 1991). Comme ils représentent une analyse syntaxique minimale des phrases, ils sont particulièrement adaptés pour analyser la syntaxe « non-canonique » de l'oral. Pour ce module d'annotation, nous sommes inspirés du travail de (Tellier et al, 2012) et nous avons utilisé essentiellement la même méthode. Le corpus *French Treebank* (Abeillé, 2003) contient une analyse syntaxique en constituants complète pour environ 200 mille tokens extraits des articles du journal *Le Monde*. Nous avons converti les étiquettes morphosyntaxiques de ce corpus aux étiquettes génériques (catégories principales) de *DisMo*, et nous avons entraîné un modèle statistique CRF (voir section 6) qui modélise la segmentation en *chunks* comme un problème d'annotation de séquence. Seulement certains types des *chunks* ont été retenus parmi ceux proposé dans le corpus *French Treebank* (voir tableau 4) :

AP	Groupe adjectival (incluant les éventuels adverbes qui modifient l'adjectif)
AdP	Groupe adverbial
NP	Groupe nominal (nom ou pronom et éventuels déterminants)
PP	Groupe propositionnel (groupe nominal introduit par une proposition)
VN	Groupe verbal (incluant les catégories VN, VPinf et VPpart du *French Treebank*)
CONC	Conjonctions de coordination
CONS	Conjonctions de subordination

Tableau 4 : Types de *chunks* détectés

Architecture du système et modules d'annotation

29 *DisMo* est structuré autour de six modules, chaque module ajoutant ou modifiant des annotations sur les différents niveaux. Les opérations suivantes sont appliquées en cascade (voir Figure 3) :

30 – Prétraitement et découpage en unités lexicales (tokenisation) ;

31 – Application de ressources linguistiques : les unités non-ambiguës sont annotées, la liste des étiquettes possibles est établie pour les autres. Certaines disfluences et unités polylexicales sont reconnues à ce stade, ainsi que les marqueurs de discours et les unités polylexicales potentielles ;

32 – Annotation morphosyntaxique (en partie du discours) préliminaire, à l'aide d'un modèle statistique de Champs Aléatoires Conditionnels (Conditional Random Fields, CRF) ;

33 – Détection des disfluences et de la segmentation, à l'aide de règles et d'un modèle CRF ;

34 – Annotation morphosyntaxique finale, combinée avec la détection des unités polylexicales, à l'aide d'un modèle statistique CRF.

35 – Post-traitement des annotations, à l'aide des règles de cohérence.

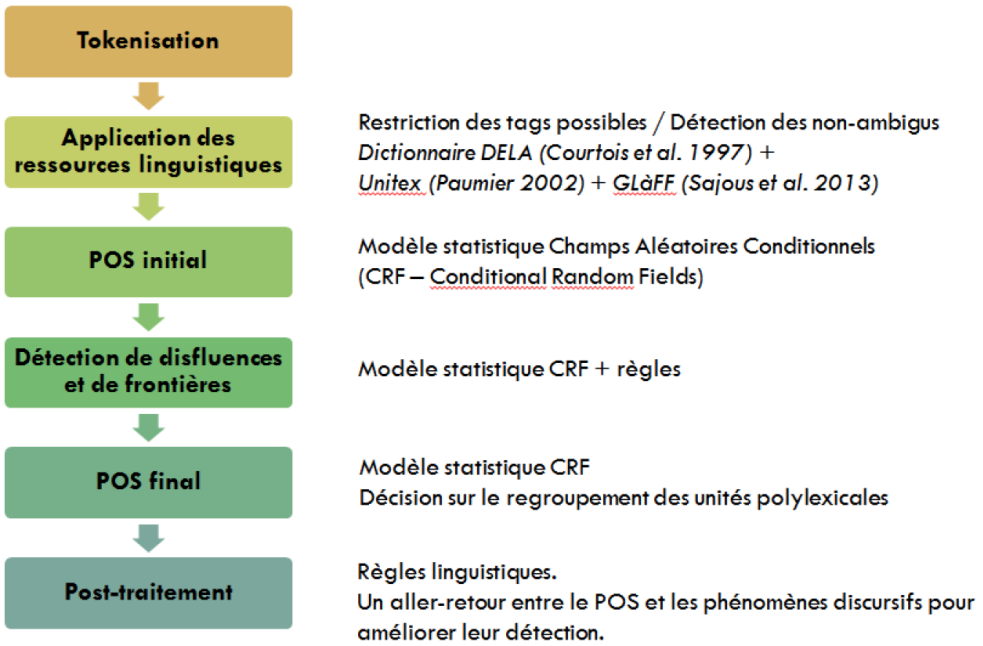

Figure 3 : Étapes de traitement automatique dans DisMo

36 L'annotation morphosyntaxique initiale se limite à l'attribution des catégories principales (voir évaluation de l'exactitude de l'annotation morphosyntaxique initiale, dans la section 7). Cet étiquetage est ensuite utilisé par les modules de détection des disfluences et du découpage en unités syntaxiques (*chunking*). Les méthodes de détection des disfluences sont exposées en détail dans (Christodoulides & Avanzi 2015). Quand une répétition est détectée, la suite de tokens répétée est simplifiée (par exemple « le » au lieu de « le le le ») pour améliorer l'exactitude de l'étape d'annotation morphosyntaxique finale (pendant laquelle les étiquettes détaillées sont attribuées). Cette simplification est invisible à l'utilisateur : il s'agit d'une opération interne, qui modifie la suite des tokens fournie aux

modules d'annotations suivants (notamment l'étape POS final). La répétition est « rétablie » dans la sortie finale et au niveau du post-traitement la catégorie morphosyntaxique attribué au token simplifié est recopié aux autres tokens qui font partie d'une répétition. De même, le phénomène de la répétition est annoté, selon le codage exposé dans la section 4.

Évaluation de la performance

37 Nous avons évalué la performance de l'étiquetage morphosyntaxique initial (en catégories principales), pour des différentes tailles du corpus d'entraînement (et du paramètre de lissage *c* du modèle CRF). Nous nous focalisons sur ce module car tous les autres se basent sur ces résultats. Le corpus d'entraînement a été construit à partir des échantillons de Paris et de Lyon du corpus PFC, corrigé manuellement (voir section 1). Le corpus de test contient 50 mille tokens du corpus CPROM-PFC (Avanzi 2014).

Taille du corpus d'entraînement			*c*	Temps (min) d'entraînement	Exactitude sans dictionnaire	Exactitude avec dictionnaire	Gain
Tokens	Séquences	Features					
50 k	4055	6761860	1	15,5	93,72%	94,98%	1,26%
			2	19,1	93,73%	95,00%	1,27%
			3	15,7	93,81%	95,05%	1,24%
60 k	4904	7798071	1	16,4	93,78%	95,00%	1,22%
			2	21,0	93,90%	95,06%	1,16%
			3	27,4	93,91%	95,07%	1,16%
70 k	5771	8730189	1	26,2	94,19%	95,28%	1,09%
			2	30,9	94,19%	95,29%	1,10%
			3	33,5	94,21%	95,31%	1,10%
90 k	7321	10269306	1	28,9	94,40%	95,44%	1,04%
			2	36,8	94,46%	95,47%	1,01%
			3	44,7	94,39%	95,44%	1,05%
100 k	7866	11266500	1	38,9	94,53%	95,52%	0,99%
			2	47,3	94,49%	95,50%	1,01%
			3	50,4	94,51%	95,51%	1,00%
120 k	9384	12962304	1	45,1	94,59%	95,58%	0,99%
			2	58,6	94,58%	95,57%	0,99%
			3	65,0	94,57%	95,57%	1,00%
140 k	**10859**	**14757462**	1	58,8	94,57%	95,58%	1,01%
			2	71,6	94,58%	95,57%	0,99%
			3	**73,2**	**94,59%**	**95,58%**	0,99%

Tableau 5 : Évaluation de la performance de l'annotation morphosyntaxique initiale. L'exactitude est calculée avant l'application des modules de disfluence, POS final et post-traitement.

38 Nous observons que l'application des ressources linguistiques (dictionnaire des formes fléchies) diminue progressivement mais reste intéressante. L'amélioration de la performance de l'annotation morphosyntaxique initiale se stabilise progressivement, ce qui nous amène à la conclusion que la taille du corpus d'entraînement est suffisante. Comme nous avons déjà mentionné, après le traitement des disfluences simples et notamment des répétitions, la précision de l'annotation morphosyntaxique finale atteint 97 % (pour le jeu d'étiquettes complet).

Conclusion et perspectives

39 Dans cet article nous avons présenté le résultat de notre collaboration pour le développement et l'amélioration de l'annotateur multi-niveaux pour les corpus oraux

DisMo, en utilisant le corpus PFC et avec une campagne de correction des annotations importante. La possibilité d'annoter automatiquement des grands corpus oraux avec une précision comparable à celle des systèmes conçus pour l'écrit, ouvre plusieurs perspectives de recherche (à titre indicatif, sur la question de la variation régionale, la phonétique, les phénomènes syntaxiques à l'oral, la fluence, etc.).

40 Dans la suite de cette étude, nous sommes en train d'ajouter une couche d'annotation supplémentaire, qui repère des relations de dépendance syntaxique entre les *chunks*. Nous avons comparé les systèmes d'annotation syntaxique proposés dans deux corpus oraux : le corpus *Rhapsodie* (Bawden et al.2014), qui propose sept relations et un système d'« empilement » pour décrire les disfluences d'édition ; et le corpus *LOCAS-F* (Martin *et al.* 2014) qui propose une double annotation en « séquences fonctionnelles » (Bilger & Campione 2002) et unités de rection. Nous avons confronté ces deux systèmes au jeu d'étiquettes pour l'annotation en relations de dépendance proposé dans le cadre du projet *Stanford Universal Dependencies* (de Marneffe *et al.* 2014), ainsi que les propositions de (Deulofeu *et al.* 2010). L'annotateur automatique DisMo est mis à jour régulièrement, selon les résultats du projet : les chercheurs intéressés peuvent consulter notre site web (www.corpusannotation.org) pour suivre ces développements.

BIBLIOGRAPHIE

Abeillé, A., Clément, L., Toussenel, F. (2003) "Building a treebank for French", A. Abeillé (ed.) *Treebanks*, Kluwer, Dordrecht

Abney S. (1991). Parsing by chunks, in Berwick, R., Abney, R., Tenny, C. (eds) *Principle-based Parsing*, Kluwer Academic Publisher

Avanzi M. (2014). A Corpus-Based Approach to French Regional Prosodic Variation, *Nouveaux cahiers de linguistique française*, vol. 31, pp. 309-323

Avanzi M., Béguelin, M.-J. & Diémoz, F. (2012-2015). Présentation du corpus OFROM – corpus oral de français de Suisse romande, Université de Neuchâtel, http://www.unine.ch/ofrom

Barreca, G., Christodoulides, G. (2014) "Un concordancier multi-niveaux pour des corpus oraux". *Actes de la 21ème Conférence Traitement Automatique du Langage Naturel (TALN)*, 1-4 juillet 2014, Marseille, France, pp. 499-504

Bawden, R., Botalla, M.-A., Gerdes, K., Kahane, S. (2014) "Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie". *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC)*, 26-31 mai 2014, Reykjavik, Islande, pp. 2320-2325

Benzitoun C., Fort K., Sagot B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. *Actes de JEP – TALN – RECITAL 2012*, vol. 2 : TALN, pp. 99-112

Bilger M., Campione C. (2002). Propositions pour un étiquetage en 'séquences fonctionnelles', *Recherches sur le français parlé*, 17, pp. 117-136

Blanc O., Constant M., Dister A., Watrin P. (2008). Corpus oraux et chunking. *Actes de Journées d'étude sur la parole (JEP)*, Avignon, France

Blanche-Benveniste C. (1990) Un modèle d'analyse syntaxique 'en grilles' pour les productions orales, *Anuario de Psicología*, Liliane Tolchinsky (coord.), vol. 47, Barcelona, pp. 11-28

Christodoulides G. & Avanzi M. (2015). Automatic detection and annotation of disfluencies in spoken French corpora, in ISCA (éds.), *Proceedings of Interspeech*, Dresde, Allemagne, 6-10 septembre 2015, pp. 1849-1853

Christodoulides, G. (2014). Praaline : Integrating tools for speech corpus research. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, 26-31 mai 2014, Reykjiavic, Islande, pp. 31-34

Christodoulides, G., Avanzi, M., Goldman, J.-Ph. (2014) DisMo : A Morphosyntactic, Disfluency and Multi-Word Unit Annotator, *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC)*, 26-31 mai 2014, Reykjavik, Islande, pp. 3902-3907

Clavel C., Adda G., Cailliau F., Garnier-Rizet M., Cavet A., Chapuis G., Cournicous S., Danesi C., Daquo A.-L., Deldossi M., Guillemin-Lanne S., Seizou M., Suignard P. (2013) Spontaneous Speech and Opinion Detection : Mining Call-centre Transcripts, *Language Resources and Evaluation*, vol. 1, pp. 40

Courtois B., Garrigues M., Gross G., Gross M., Jung R., Mathieu-Colas M., Monceaux A., Poncet-Montange A., Silberztein M., Vivès R. (1997). Dictionnaires électronique DELAC : les mots composés binaires. Rapport technique 56, LADL, Université Paris 7

de La Clergerie É., Sagot B., Nicolas L., Guénot M.-L. (2009) FRMG : évolutions d'un analyseur syntaxique TAG du français, Journée de l'ATALA « Quels analyseurs syntaxiques pour le français ? », Paris, France

de Marneffe, M.C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C.D (2014) Universal Stanford dependencies : A cross-linguistic typology, *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC)*, 26-31 mai 2014, Reykjavik, Islande, pp. 4585-4592

Degand L., Simon A.C. (2009). On identifying basic discourse units in speech : Theoretical and empirical issues. *Discours*, 4. En ligne : http://discours.revues.org/5852

Deulofeu J., Dufort L., Gerdes K., Kahane S., Pietrandrea P. (2010) "Depends on what the French say : Spoken corpus annotation with and beyond syntactic function", *Proceedings of the 4th Linguistic Annotation Workshop (LAW IV)*, Uppsala, Suède, 8 p.

Dister A., (2014) 'parler sans accent pour moi c'est sans sans sans bafouiller' Quelles répétitions de formes en français parlé ?, *Actes du 4e Congrès mondial de linguistique française*

Durand, J., Bernard L., Lyche C. (2009). Le projet PFC : une source de données primaires structurées. In J. Durand, B. Laks et C. Lyche (eds) (2009) *Phonologie, variation et accents du français*. Paris : Hermès. pp. 19-61

Eshkol I., Tellier I, Taalab S., Billot S. (2010). Étiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques. *Actes de 10es Journées Internationales d'analyse statistique des données textuelles (JADT)*

Gerdes K., Kahane S. (2009) Speaking in piles : Paradigmatic annotation of French spoken corpus, *Processing of the 5th Corpus Linguistics Conference*, Liverpool, 15 p.

Heeman P., McMillin A., Scott Yaruss J. (2006) An annotation scheme for complex disfluencies, *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP)*, pp. 1081-1084

Honnibal, M., Johnson, M. (2014) Joint Incremental Disfluency Detection and Dependency Parsing, *Transactions of the Association of Computational Linguistics*, 2 (2014), pp. 131-142

Kahane S. (2012) De l'analyse en grille à la modélisation des entassements, in S. Caddeo, M.-N. Roubaud, M. Rouquier, F. Sabio (éds.), *Penser les langues avec Claire Blanche-Benveniste*, Presses de l'université de Provence, pp. 101-116

Lafferty J., McCallum A., Pereira F. (2001) Conditional Random Fields : Probabilistic Models for Segmenting and Labelling Sequence Data, *Proceedings of the International Conference on Machine Learning*, pp. 282-289

Martin L., Degand L., Simon A. C. (2014) "Forme et fonction de la périphérie gauche dans un corpus oral multigenres annoté", *Corpus*, no. 13, pp. 243-265

Mateer M., Taylor A. (1995) *Disfluency annotation stylebook for the Switchboard corpus*, Manuscript, Department of Computer and Information Science, University of Pennsylvania

Sajous F., Hathout N., Calderone B. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. *Actes de TALN*

Schmid H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees, *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK

Shriberg E. (1994) *Preliminaries to a Theory of Speech Disfluencies*, Ph.D. thesis, University of California, Berkeley

Shriberg E. (2001) To 'errrr' is human : ecology and acoustics of speech disfluencies, *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 153-169

Simon A. C., Francard M., Hambye Ph. (2014, sous presse) The VALIBEL Speech Database, in J. Durand, U. Gut & G. Kristoffersen (éds.), *The Oxford Handbook of Corpus Phonology*, Oxford, Oxford University Press, pp. 552-561

Tellier I., Duchier D., Eshkol I., Courmet A., Martinet M. (2012). Apprentissage automatique d'un chunker pour le français, *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 2 : TALN, Grenoble, 4-8 juin 2012, pp. 431–438

Tsuruoka, Y., Tsujii, J., Ananiadou, S. (2009) Fast Full Parsing by Linear-Chain Conditional Random Fields, *Proceedings of the 12th Conference of the European Chapter of the ACL*, 30 mars – 3 avril 2009, Athènes, Grèce, pp. 790-798

Valli A., Véronis J. (1999). Étiquetage automatique de corpus oraux : problèmes et perspectives. *Revue française de linguistique appliquée*, vol. 4, no. 2, pp. 113-133

RÉSUMÉS

L'annotation des corpus oraux présente des défis particuliers, liés aux caractéristiques de la langue parlée et sa transcription. Si la méthodologie d'analyse et les outils d'annotation automatique doivent être adaptés à ces défis, il est toutefois souhaitable de garder la possibilité de comparer un corpus oral avec un corpus écrit, sur base d'un « dénominateur commun », et d'enrichir l'annotation avec des couches supplémentaires pour décrire les phénomènes propres à l'oral. Dans cet article nous présentons l'approche implémentée dans l'outil *DisMo*, un annotateur automatique conçu spécifiquement pour les corpus oraux, qui propose une analyse à plusieurs niveaux : étiquetage morphosyntaxique, lemmatisation, détection des unités poly-lexicales, détection et annotation des phénomènes de disfluence et des marqueurs de discours, et découpage en unités syntaxiques minimales. Nous présenterons nos travaux sur le corpus

Phonologie du Français Contemporain (PFC) qui ont permis de réviser l'outil et d'améliorer sa performance. Nous préciserons les choix théoriques et pratiques quant aux niveaux d'annotation, les phénomènes détectés, les jeux d'étiquettes, ainsi qu'une évaluation de la performance de l'annotation automatique.

Annotating spoken corpora poses unique challenges stemming from the particular characteristics of spontaneous speech and its transcription. Automatic annotation tools need to adapt to these challenges. At the same time, it is desirable to define a "least common denominator" of written and spoken language corpora, to allow for comparisons between these two modalities, and apply an enriched annotation scheme for phenomena specific to spoken language. In this article, we present the approach implemented in the *DisMo* automatic annotator, which is specifically designed for spoken corpora, and which generates a multi-level annotation, including : part-of-speech tagging, lemmatisation, multi-word unit detection, detection and annotation of disfluencies and discourse markers, and chunking. We present our work on the French corpus of the *Phonologie du Français Contemporain (PFC)* project ; this work allowed us to improve the tool. We discuss the theoretical and practical considerations that informed the choice of levels of annotation, types of phenomena detected, and tag sets, and we present a performance evaluation of the automatic annotation.

INDEX

Mots-clés : exploitation de corpus oraux, annotation multi-niveaux, annotation automatique
Keywords : exploitation of oral corpora, multilevel annotation, automatic annotation

AUTEURS

GEORGE CHRISTODOULIDES

GIULIA BARRECA