DisMo : un annotateur multi-niveaux pour les corpus oraux

George Christodoulides[1], Giulia Barreca[2], Mathieu Avanzi[3]
(1) Centre Valibel, Institut Langue & Communication, Université de Louvain
Place Blaise Pascal 1, B-1348 Louvain-la-Neuve, Belgique
(2) Laboratoire MoDyCo, CNRS, Université Paris Ouest Nanterre La Défense
200, Avenue de la République, FR-92001 Nanterre, France
Université Catholique de Milan
1, Largo A. Gemelli, 20123 Milan, Italie
(3) DTAL, Faculty of Modern & Medieval Languages, University of Cambridge
Sidgwick Avenue, CB3 9DA Cambridge, Royaume-Uni
george@mycontent.gr; giulia.barreca@gmail.com; mathieu.avanzi@gmail.com

Résumé. Dans cette démonstration, nous présentons l'annotateur multi-niveaux *DisMo*, un outil conçu pour faire face aux spécificités des corpus oraux. Il fournit une annotation morphosyntaxique, une lemmatisation, une détection des unités poly-lexicales, une détection des phénomènes de disfluence et des marqueurs de discours.

Abstract.

DisMo: a multi-level annotator for spoken language
In this demonstration we present the multi-level automatic annotator *DisMo* which is specifically designed for the challenges posed by spoken language corpora. Its output comprises of part-of-speech tagging, lemmatization, multi-word unit detection, detection of disfluency phenomena and discourse markers.

Mots-clés : annotation morphosyntaxique, corpus oraux, disfluences, unités poly-lexicales
Keywords: part-of-speech tagging, spoken corpora, disfluencies, multi-word expressions

1 Introduction

L'annotation des corpus oraux présente des défis particuliers, liés aux caractéristiques de la langue parlée et sa transcription, notamment : l'absence de ponctuation, les unités de segmentation multiples, les disfluences et la syntaxe souvent non-canonique. Si la méthodologie d'analyse et les outils d'annotation automatique doivent être adaptés, il est toutefois souhaitable de pouvoir 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. 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 soit un apprentissage automatique à partir de corpus oraux corrigés manuellement (Eshkol et al. 2010), soit l'utilisation de fichiers de paramétrage spécifiques pour les données orales (Benzitoun et al. 2012).

Cet article de démonstration présente *DisMo* (Christodoulides et al. 2014), un outil d'annotation automatique spécifiquement conçu pour les corpus oraux, et 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. Actuellement, 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 million tokens), la collection des corpus du centre VALIBEL (Simon et al. 2014) (environ 6 million tokens), et le Corpus Oral de français de Suisse Romande (OFROM) (0,5 million 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 l'annotation de 57 mille tokens. 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, la précision du système est de l'ordre de 97% pour le jeu d'étiquettes complet, et 98% pour un jeu d'étiquettes simplifié.

2 Architecture du système

L'annotateur 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 (p.ex. au niveau des mots ou même des syllabes) peut permettre 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, et les annotations sont stockées séparément. Nous avons d'ailleurs développé des scripts qui facilitent le traitement des transcriptions selon certaines conventions (indices de locuteurs, conventions de transcription 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* ; Christodoulides 2014). *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 :

- Prétraitement et découpage en unités lexicales (tokenisation) ;
- Application de ressources linguistiques: les unités non-ambiguës sont annotées, la liste des étiquettes possibles est établie pour les autres. Certaines disfluences et unités poly-lexicales sont reconnues à ce stade, ainsi que les marqueurs de discours et les unités poly-lexicales potentiels ;
- Annotation morphosyntaxique (en partie du discours) préliminaire, à l'aide d'un modèle statistique CRF ;
- Détection des disfluences et de la segmentation, à l'aide de règles et d'un modèle CRF ;
- Annotation morphosyntaxique finale, combinée avec la détection des unités poly-lexicales, à l'aide d'un modèle statistique CRF.
- Post-traitement des annotations, à l'aide des règles de cohérence.

DisMo est écrit en C++ et utilise plusieurs bibliothèque de source ouverte, notamment *OpenFST*, *Helsinki Finite-State Transducer Technology (HFST)*[1] et *CRF++ toolkit*[2]. Les ressources lexicales sont basées sur les dictionnaires DELA (Courtois et al., 1997), GLÀFF (Sajous et al., 2013) et des dictionnaires des unités nommés crées par les auteurs.

3 Niveaux d'annotation

L'articulation de l'annotation sur plusieurs niveaux est présentée dans la Figure 1. L'étiquetage morphosyntaxique attribue une des 64 catégories d'étiquettes (12 catégories principales), ainsi que des informations supplémentaires (genre, nombre, lemme) à chaque unité lexicale minimale et aussi à chaque unité poly-lexicale identifiée. Le jeu d'étiquettes complet, ainsi qu'une comparaison avec d'autres annotateurs est disponible sur le site web du logiciel (www.corpusannotation.org/dismo/tagset). Le schéma d'annotation pour les phénomènes de disfluence (hésitations, amorces, allongements, répétitions, insertions, substitutions, interruptions, etc.), ainsi que les algorithmes utilisés pour leur détection, sont détaillés dans (Christodoulides & Avanzi 2015).

FIGURE 1 : *Les différents niveaux d'annotation sous forme de TextGrid*.

_	un	mois	dans	une	agence	Saint	Cloud	donc	là	travaillait	euh	_	la	main	_	tok-min (518)
_		NOM:com	PR		NOM:com	NOM:prop		ADV	AD	VER:impf	ITJ	_	DE	NOM:com	_	pos-min (518)
SIL:1											FIL	SIL:1			SIL:1	disfluency (518)
_	un	mois	dans	une	agence	Saint Cloud		donc	là	travaillait	euh	_	la	main	_	tok-mwu (497)
_		NOM:com	PR		NOM:com	NOM:pro		ADV	AD	VER:impf	ITJ	_	DE	NOM:com	_	pos-mwu (497)
SIL:1								CON				SIL:1			SIL:1	discourse (102/497)
_	un mois dans une agence à Saint Cloud donc là on travaillait euh											_	à la main		_	ortho (119)

[1] http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/
[2] http://crfpp.googlecode.com/svn/trunk/doc/index.html

4 Conclusion

Dans cet article, nous avons présenté une synthèse des travaux antérieurs portant sur l'annotation des corpus oraux, qui a permis d'aboutir à un nouveau système, conçu spécifiquement pour traiter les phénomènes propres aux corpus oraux. Ce système nous a permis d'annoter un grand ensemble de donnés de corpus oraux. Grâce à cette campagne d'annotation, plusieurs études ont pu être conduites (notamment dans sur la question de la variation régionale, l'étude de phénomènes phonotactiques, p. ex. la liaison, la chute des liquides post-obstruantes), les phénomènes syntaxiques propres à l'oral, la fluence et la disfluence, etc. Des interfaces web pour faciliter l'accès à ces données, ainsi qu'un service web pour l'utilisation de l'annotateur, sont en cours de conception et seront bientôt mises à la disposition de la communauté. *DisMo* est disponible (licence GPL3) sur le site web www.corpusannotation.org, en version autonome (plateformes Windows, Mac et Linux) ou comme plug-in pour le logiciel de gestion et d'annotation de corpus *Praaline*.

Références

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

AVANZI M., BEGUELIN M.-J., DIEMOZ F. (2012). *Présentation du corpus OFROM – corpus oral de français de Suisse romande*, Université de Neuchâtel, http://www.unine.ch/ofrom

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, 99-112.

BLANC O., CONSTANT M., DISTER A., WATRIN P. (2008). Corpus oraux et chunking. Actes de Journées d'étude sur la parole (JEP), Avignon, France.

CHRISTODOULIDES G., AVANZI M. (2015). Automatic Detection and Annotation of Disfluencies in Spoken French Corpora, *Proceedings of Interspeech 2015,* Dresde, Allemagne, 6-10 septembre 2015, 5 pp.

CHRISTODOULIDES G., AVANZI M., GOLDMAN, J-PH. (2014). DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech, *Proceedings of LREC 2014*, Reykjavik, Islande, 3902-3907.

CHRISTODOULIDES, G. (2014). Praaline: Integrating tools for speech corpus research. Actes de *IX Language Resources and Evaluation Conference (LREC 2014)*, 26-31 mai 2014, Reykjiavic, Islande, 31-34.

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.

DURAND J., LAKS B., LYCHE C., (EDS) (2009). *Phonologie, variation et accents du français*, Paris, Hermès.

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.*

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 International Conference on New Methods in Language Processing*, Manchester, UK.

SIMON A.C., FRANCARD M., HAMBYE P. (2014). "The VALIBEL Speech Database", In: Durand J., Gut U., Kristoffersen G., *The Oxford Handbook of Corpus Phonology*, DOI: 10.1093/oxfordhb/9780199571932.013.017.

VALLI A., VERONIS J. (1999). Étiquetage automatique de corpus oraux : problèmes et perspectives. *Revue française de* linguistique appliquée, 4(2), 113-133.