Au-delà du token

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

George Christodoulides
Centre Valibel, IL&C, Université catholique de Louvain
Giulia Barreca
Université Paris Ouest Nanterre La Défense & Université catholique de Milan

8es Journées Internationales de Linguistique de Corpus
2-4 septembre 2015, Orléans, France

Introduction

- Défis particuliers lies aux caractéristiques de la langue parlée et sa transcription:
  - absence de ponctuation
  - unités de segmentation multiples
  - disfluences
  - syntaxe souvent non-canonique
- Par conséquent : méthodologie d'analyse adaptée
- Pourtant il est *souhaitable* de pouvoir comparer les résultats de l'analyse d'un corpus oral avec celles d'un corpus écrit : dénominateur commun

Présentation des outils

DisMo : Un annotateur automatique multi-niveaux pour des corpus oraux

Annotation :

- morphosyntaxique (POS)
- disfluences
- phénomènes discursifs

Conçu pour les défis spécifiques à l'oral:

- absence de ponctuation
- disfluences (phonétiques et syntaxiques)
- importance des « ponctuants » des marqueurs du discours et des frontières prosodiques pour le traitement

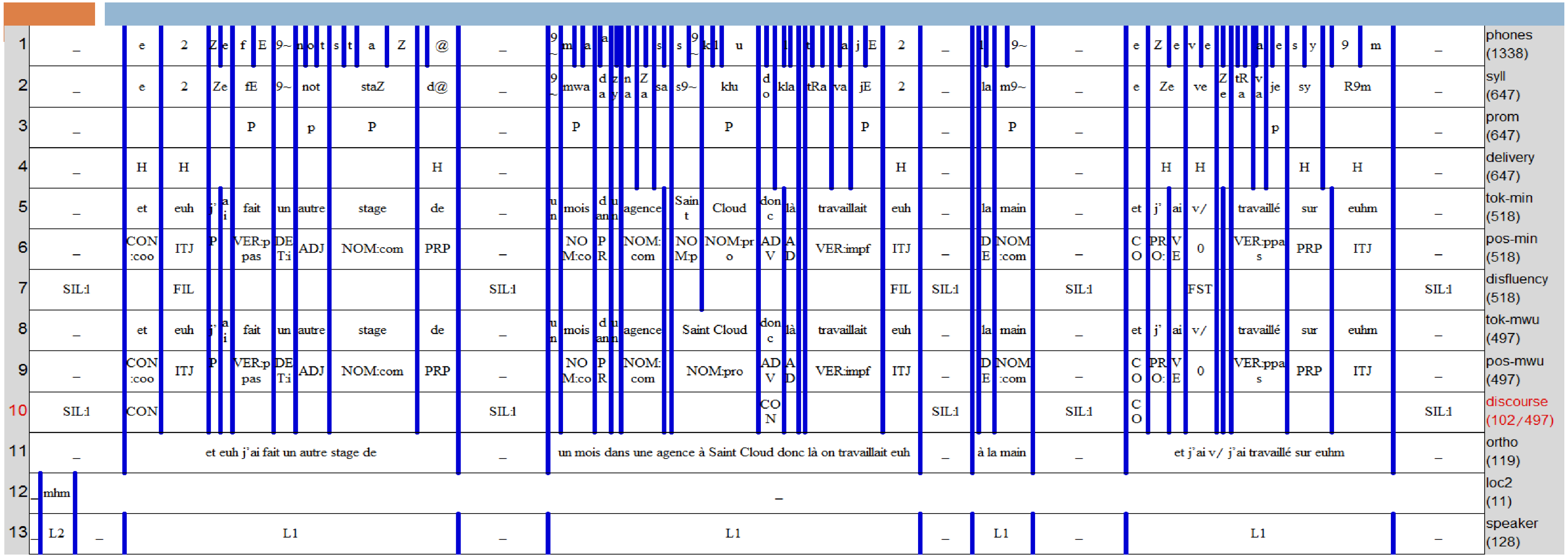
(l'annotation de l'écrit est possible aussi)

Langues : anglais, français (en préparation: néerlandais, portugais)

DisMo: entrée et sortie

- Entrée :
  - transcription orthographique
  - (optionnellement) alignée avec le son
- Sortie:
  - annotation morphosyntaxique (POS) à 2 niveaux:
    - token minimal
    - unités polylexicales (Multi-Word Units = MWU)
  - disfluences
  - marqueurs du discours et connecteurs

DisMo: exemple d'annotation

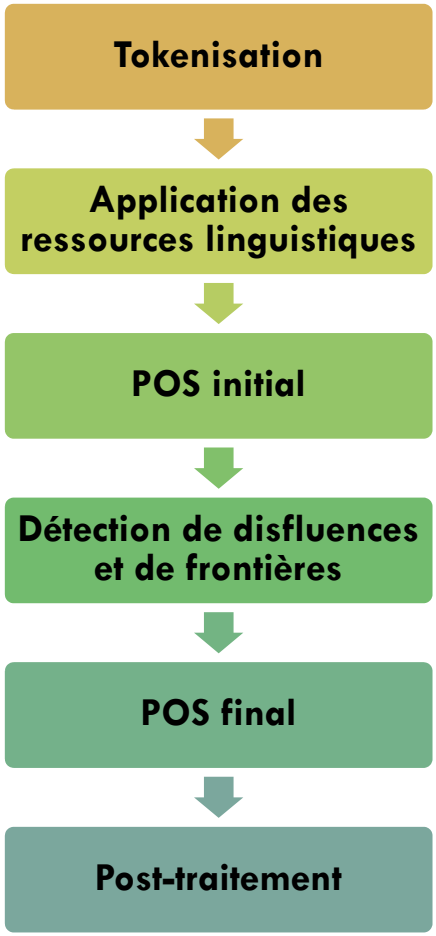


- 3 niveaux d'annotation:
  - tok-min / pos-min / disfluency
  - tok-mwu / pos-mwu
  - discourse
- Possibilité d'annotation multi-locuteurs (dialogue)

DisMo: architecture du système

Application en **cascade**: chaque étape affine les résultats des étapes précédents.

DisMo est un système **hybride**: modèles statistiques + règles + ressources linguistiques



Restriction des tags possibles / Détection des non-ambigus
Dictionnaire DELA (Courtois et al. 1997) +
Unitex (Paumier 2002) + GLàFF (Sajous et al. 2013)

Modèle statistique Champs Aléatoires Conditionnels
(CRF – Conditional Random Fields)

Modèle statistique CRF + règles

Modèle statistique CRF
Décision sur le regroupement des unités polylexicales

Règles linguistiques.
Un aller-retour entre le POS et les phénomènes discursifs pour améliorer leur détection.

DisMo : évaluation et fiche technique

- Évaluation LREC
  - précision 97% recall 96.5% f-measure: 96.8%
- Évaluation PFC
  - précision 97% sur échantillon de test
- Fiche technique
  - Système et ressources linguistiques en source libre (GPL / LGPLLR)
  - Intègre d'autres projets en source libre: CRF++, Open FST, HFST
  - C++ avec la bibliothèque Qt, Plateforme: Windows Mac Linux
  - Environnement: logiciel indépendant avec interface graphique ligne de commande plugin Praaline
- En entrée: Praat TextGrid, Transcriber, ELAN,Exmaralda Partitur, fichier texte
- En sortie: (les formats d'entrée) et aussi: tableau: tab-separated (Excel, OSD), XML (TEI), base de données (format Praaline)

Structure du jeu d'étiquettes

- Catégories principales:
  `ADJ ADV CON DET NUM FRG`
  `ITJ NOM PFX PRO PRP VER`
- Étiquettes à 2-3 niveaux – exemple:
  `DET:def DET:dem DET:ind DET:int`
  `DET:exc DET:par DET:pos`
- Information supplémentaire (pos-ext) morphologique etc.

Jeu d'étiquettes complet

Part of Speech tag-set					
Adjectives	ADJ	Adjective	Pronouns	PRO:dem	Pronoun, demonstrative
Adverbs	ADV	Adverb		PRO:ind	Pronoun, indefinite
	ADV:acr	Adverb		PRO:int	Pronoun, interrogative
	ADV:int	Adverb, interrogative		PRO:per:sjt	Pronoun, personal, subject
	ADV:neg	Adverb, negative		PRO:per:objd	Pronoun, personal, direct object
Conjunctions	CON:coo	Conjunction, co-ordinating		PRO:per:obji	Pronoun, personal, indirect object
	CON:sub	Conjunction, subordinating		PRO:pos	Pronoun, posessive
Determiners	DET:def	Determiner, definite article		PRO:rel	Pronoun, relative
	DET:dem	Determiner, demonstrative		PRO:ref	Pronoun, reflexive
	DET:ind	Determiner, indefinite		PRO:per:ton	Pronoun, personal, clitic (stressable)
	DET:int	Determiner, interrogative		Concatenated forms are annotated with tag1|tag2	
	DET:exc	Determiner, exclamative	Verbs	VER:cond	Verb, conditional perfect
	DET:par	Determiner, partitive		VER:cond:aux	Verb, conditional perfect, auxiliary
	DET:pos	Determiner, posessive		VER:fut	Verb, future
Numerals	NUM:crd:det	Cardinal number, determiner		VER:fut:aux	Verb, future, auxiliary
	NUM:crd:adj	Cardinal number, adjective		VER:impe	Verb, imperative
	NUM:crd:pro	Cardinal number, pronoun		VER:impf	Verb, imperfect
	NUM:crd:nom	Cardinal number, noun		VER:impf:aux	Verb, imperfect, auxiliary
	NUM:ord:adj	Ordinal number, adjective		VER:inf	Verb, infinitive
	NUM:ord:pro	Ordinal number, pronoun		VER:inf:aux	Verb, infinitive, auxiliary
	NUM:ord:nom	Ordinal number, noun		VER:ppas	Verb, past participle
Foreign	FRG	Foreign word		VER:ppre	Verb, perfect participle
Interjections	ITJ	Interjection		VER:ger	Verb, gerundive (only on MWU tier)
	ITJ:(category)	…specifying original POS		VER:pres	Verb, present
	ITJ:ono	Onomatopoeia		VER:pres:aux	Verb, present, auxiliary
Nouns	NOM:acr	Noun, acronym		VER:pres:entatif	Verb, existential (*voilà, voici*)
	NOM:com	Noun, common		VER:simp	Verb, simple past
	NOM:pro	Noun, proper		VER:simp:aux	Verb, simple past, auxiliary
	NOM:pro:acr	Noun, proper		VER:subi	Verb, subjunctive, imperfect
Prefixes	PFX	Prefix		VER:subi:aux	Verb, subjunctive, imperfect, aux.
Prepositions	PRP	Preposition		VER:subp	Verb, subjunctive, present
	PRP:det	Preposition + Determinant		VER:subp:aux	Verb, subjunctive, present, auxiliary

Annotation multi-niveaux

L'annotation par couches offre certaines possibilités :

- Garder une annotation 'classique' en pos-min susceptible d'être précisée dans une autre couche
- Annoter les unités polylexicales (locutions mots composés formes figés…) en 'tok-mwu' / 'pos-mwu'
- Ne pas mettre sur le même plan des informations d'ordre catégoriel et fonctionnel
- Noter les fonctions pragmatiques dans une couche dédiée ('discourse') tout en gardant les catégories originelles en 'pos-min'

Avantages de l'annotation multi-niveaux

- **Morphosyntaxe**: distinction informations d'ordre catégoriel et fonctionnel:

  « je connaissais déjà des gens […] je les (PRO:per:objd) rencontrais pendant les vacances donc dans l'Aveyron »
  (PFC Aveyronnais à Paris; 75xmb1gg)

- **Pragmatique**: distinction fonction pragmatique (couche discourse) / catégorie morphosyntaxique originelle (couche pos-min):

  « bon (pos-min: ITJ; discourse: MD) déjà j'ai dû passer un un euh un concours un test 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)

- **Lexique**: traitement des unités poly-lexicales (mots, adverbes composés, chiffres, locutions)

  « je n'ai pas pu obtenir de poste à Lyon tout de suite (pos-mwu ADV) donc j'ai été exilé » (PFC Lyon; 69aag1gg)

Avantages de l'annotation multi-niveaux

L'exemple des Introducteurs: *il y (avoir) / c'(être)*

« Ils vendent tous les objets euh que des objets euh russes ça c'est fabrication russe. Tous ce qui était objet russe oh il y en avait certains c'était euh donc des fournisseurs de Paris »
(PFC, Aveyronnais à Paris, 75xlv11g)

DisMo: deux niveaux d'annotation:

pos-min:

il (PRO:per:sjt); *y* (PRO:per:obji); *en* (PRO:per:obji) *avait* (VER:pres)

c' (PRO:dem); était (VER:impf)

pos-mwu:

il y en avait (INTROD)

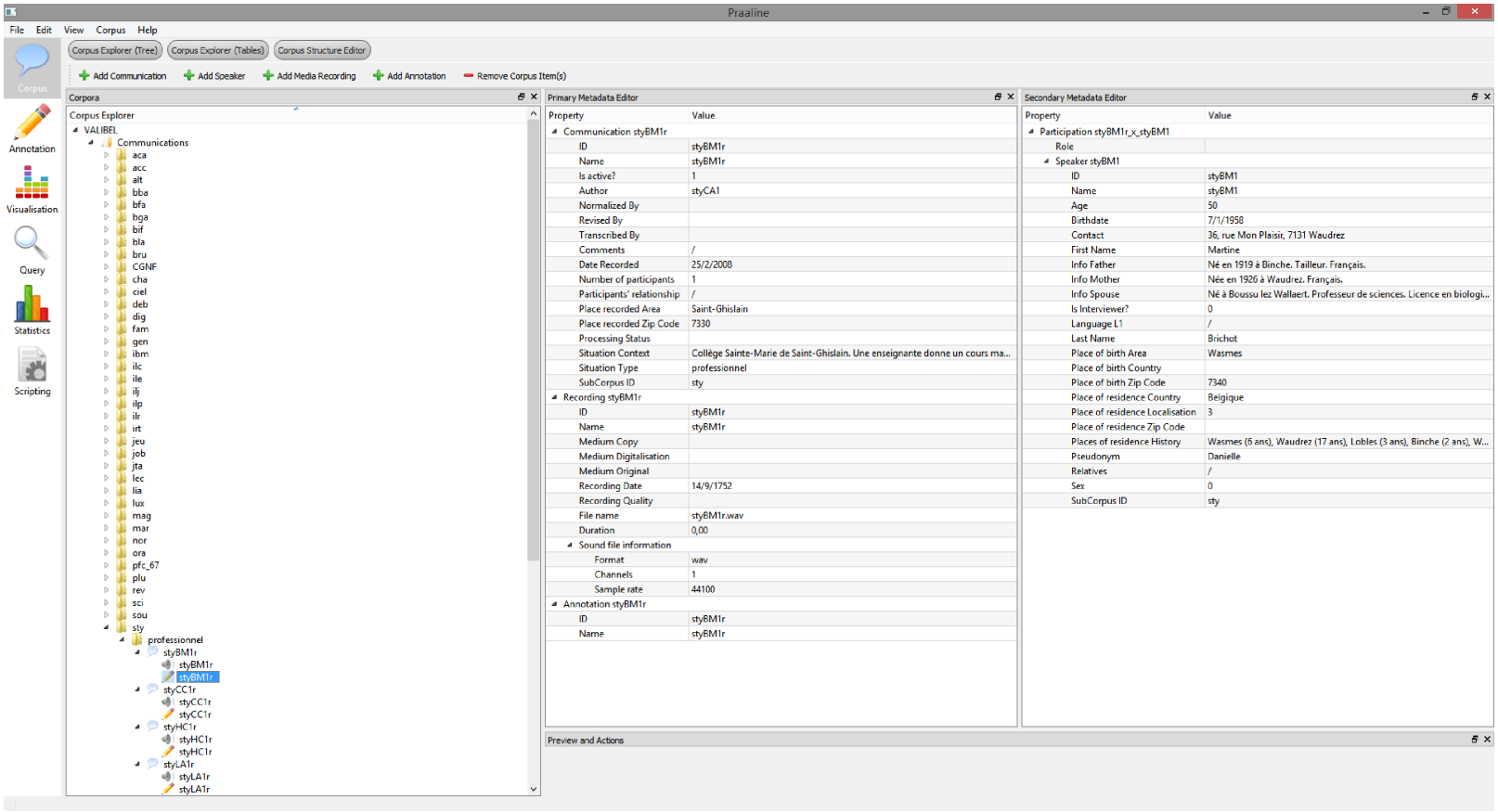
c'était (INTROD)

vs. autres annotateurs POS (p.ex. TreeTagger):

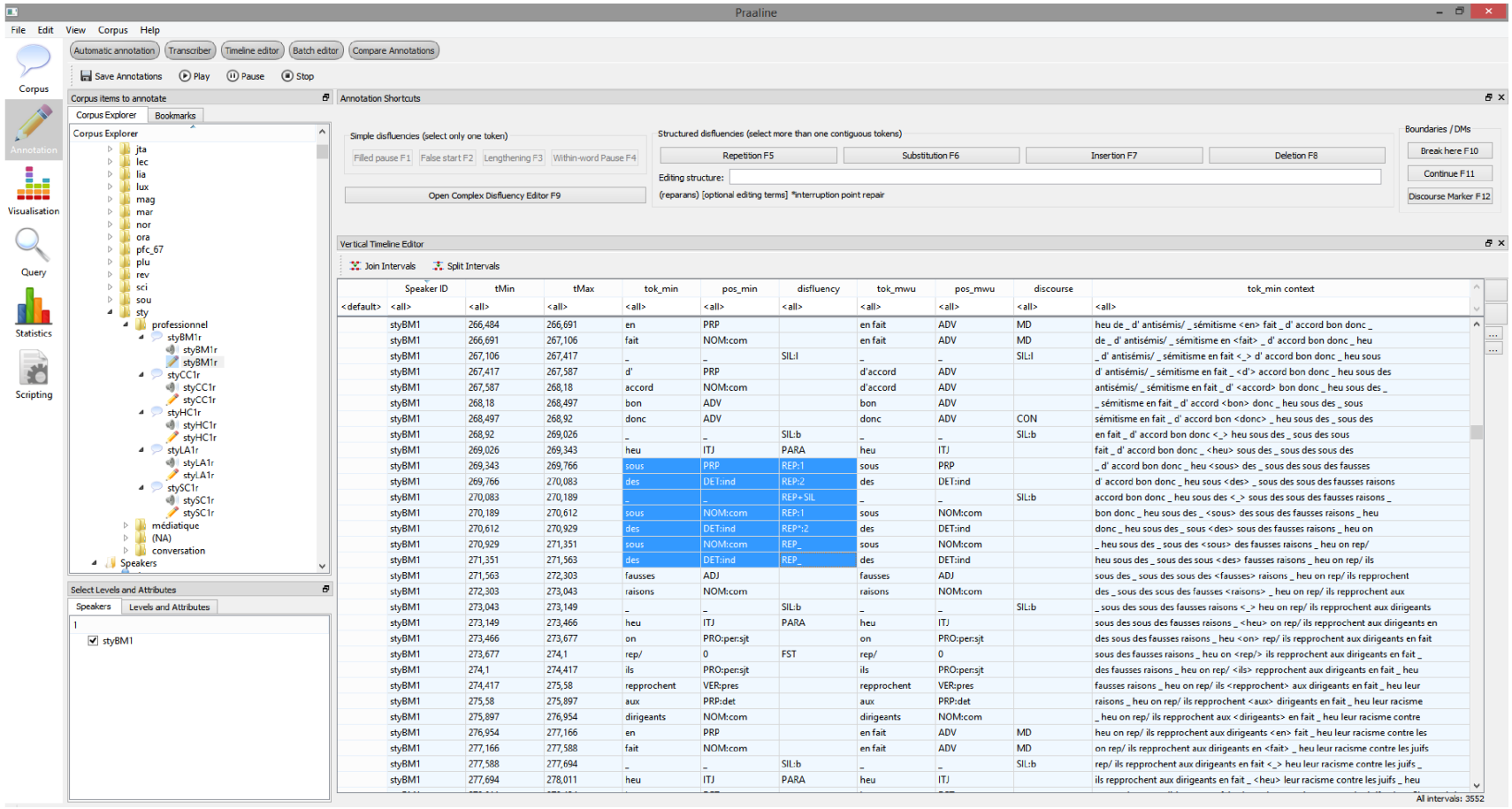
il (PRO:PER); *y* (PRO:PER); en (PRO:PER) *avait* (VER)

c' (PRO:PER); était (VER:IMPF)

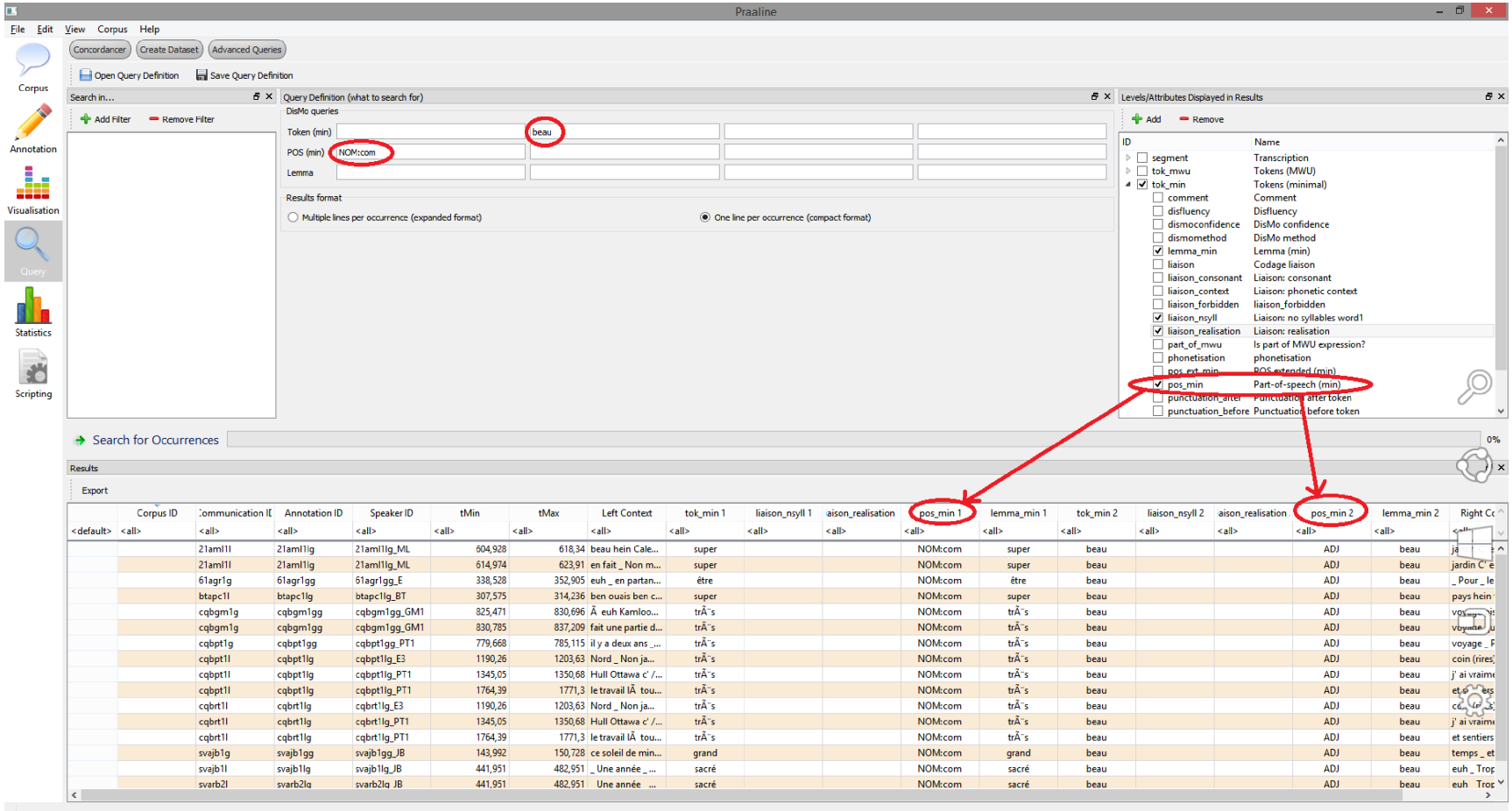
Praaline : édition de métadonnées



Praaline : consultation et annotation



Praaline : requêtes



Corpus annotés

- Actuellement trois grands corpus de référence du français parlé ont été annotés à l'aide de DisMo:
  - Corpus du projet Phonologie du Français Contemporain (PFC) (Durand et al. 2009) 1,4 million tokens
  - Collection des corpus du centre VALIBEL (Simon et al. 2014) environ 6 million tokens
  - Corpus Oral de français de Suisse Romande (OFROM) (Avanzi et al. 2012) 0,5 million tokens
  - Promis ☺ ESLO
- Plusieurs autres (C-Phonogenre, LOCAS, Rhapsodie…)

Phénomènes de disfluence

Détection et annotation des disfluences

- Amélioration de la performance du traitement automatique (analyse morphosyntaxique etc)
- Mais aussi, intérêt pour étudier le phénomène en soi

- Proposition d'un schéma d'annotation
- Appliqué sur un corpus de 7 heures de parole spontanée
- Utilisé pour entraîner un système automatique

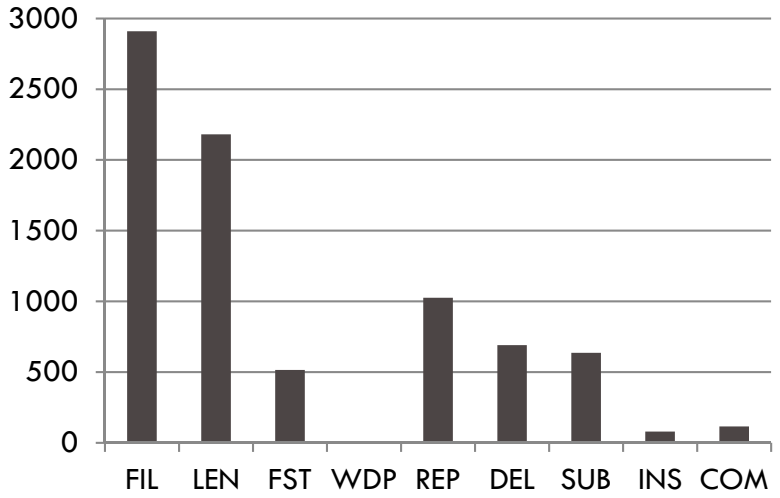
Schéma d'annotation des disfluences DisMo

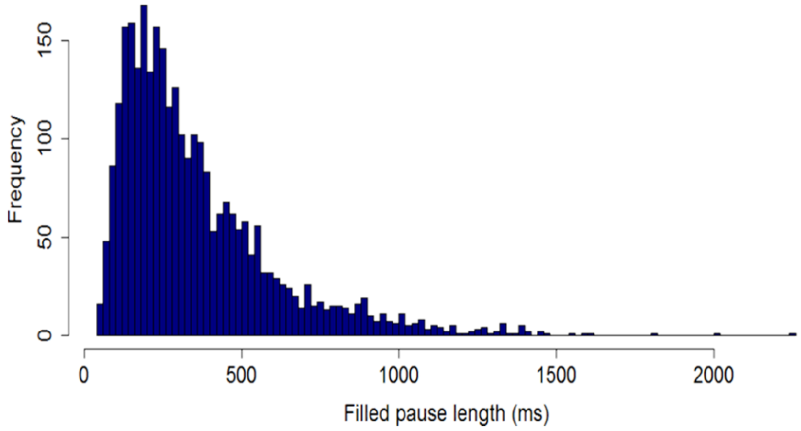
Niveau 1: Disfluences **simples** = affectent un seul token		
FIL	Pauses remplies	c' est pour ça que j' hésite euh un peu en parler FIL
LEN	Allongement lié à une hésitation	au cercle d'oenologie de= Bruxelles LEN
FST	Amorce lexicale	comme infirmière so/ sociale FST
WDP	Pause intra-mot	il m' a dit ça su+ _ +ffit WDP
Niveau 2: Répétitions où un ou plusieurs tokens sont répétés (exactement)		
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: Disfluences **structurées** (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 ça aurait INS* INS+FIL INS_ INS_ INS_
Niveau 4: Disfluencies **complexes** qui combinent plusieurs disfluences structurées (« backtracking table »)		
COM	Complexe	les ac/ les actions enfin les activités enfin professionnelles COM COM COM COM COM COM COM COM COM

Analyse des disfluences dans le corpus CPROM-PFC

- Corpus CPROM-PFC, sous-corpus interview sociolinguistique (Avanzi 2014)

- **Fréquence**: pauses remplies (autonomes et voyelles épenthétiques),allongements
- répétition > suppression > substitution
- **82%** des amorces lexicaux se trouvent au point d'interruption des disfluences structurées
- La durée des pauses remplies suit la distribution log-normale, uni-modale
- Indices prosodiques repair – reparandum +débit (art. rate), +f_0, +intensité





Analyse de séquences de disfluences dans le corpus CPROM-PFC

Séquence	Description	Occurrences
FIL SIL:l	Pause remplie, suivie d'une pause silencieuse longue	446
REP* REP_	Répétition d'un token	186
LEN SIL:l	Allongement suivi par une pause silencieuse longue	121
FIL SIL:b	Pause remplie, suivie d'une pause silencieuse brève	100
REP* REP+SIL REP_	Répétition d'un token, avec un pause brève entre les deux	90
LEN LEN	Allongement consécutive de deux tokens	74
LEN FIL	Allongement suivi par une pause remplie	68
REP:1 REP*:2 REP_ REP_	Répétition de deux tokens (A B A B)	58
REP:1 REP*:2 REP+SIL REP_ REP_	Répétition de deux tokens (A B A B), avec une pause silencieuse entre eux	39
DEL DEL*	Suppression de deux tokens	38
REP* REP+FIL REP+SIL REP_	Répétition d'un token, avec une pause remplie et une pause silencieuse	34
SUB* SUB_	Substitution d'un token avec un autre	33
REP:1 REP*:1 REP_	Double répétition d'un token	33
DEL*	Suppression d'un token	31
LEN SIL:b	Allongement suivi par une pause brève	30
LEN FIL SIL:l	Allongement suivi par une pause remplie et une pause longue	30
DEL DEL* SIL:l	Suppression de deux tokens, suivi par une pause longue	26

Détection automatique des disfluecnes

- FIL, FST, WDP selon conventions de transcription
- Allongements (LEN) : classification SVM (support vector machines), mesures prosodiques
- Répétitions : reconnaissance de patrons + modèles CRF
- Disfluences d'édition : modèles CRF (conditional random fields), formes + annotation morphosyntaxique

Disfluency type / Method	Prec	Recall	F-meas
LEN – SVN classifier	78.2%	87.4%	82.5%
REP – CRF model	84.3%	75.8%	79.8%
IP – Interruption point hypotheses	76.7%	52.0%	62.0%
SUB, INS, DEL – CRF models	(see Table 2)		

	Reparandum / Editing terms	Repair region	Prec	Recall	F-meas
Gold standard Interruption Points (upper limit)					
1	Separate	Predict	77.6%	51.4%	61.9%
2	Merged		74.7%	44.7%	55.9%
3	Separate	Ignore	82.4%	62.8%	71.3%
4	Merged		76.9%	53.2%	62.9%
Predicted Interruption Points (actual performance)					
1	Separate	Predict	54.3%	36.5%	43.7%
2	Merged		48.6%	31.3%	38.0%
3	Separate	Ignore	62.6%	42.1%	50.3%
4	Merged		59.2%	36.2%	44.9%

Christodoulides & Avanzi (2015) Automatic Detection and Annotation of Disfluencies in Spoken French Corpora, *Proc. Interspeech 2015*, Dresden, Germany, 6-10 September 2015, 5 pp.

Annotation macro-syntaxique

Chunking

- Les **chunks** sont des constituants continus et non-récursifs (Abney 1991)
- Types de chunks dans le *French Treebank* (Abeillé et al. 2003)
  - AP syntagme adjectival
  - AdP syntagme adverbial
  - NP syntagme nominal
  - PP syntagme prépositionnel
  - Srel proposition rélative
  - Ssub proposition subordonnée
  - Sint autres subordonnées
  - VPinf infinitifs
  - VPpart participiales
  - VN noyau verbal
  - COORD syntagme coordonné
- Application sur l'oral (Tellier et al. 2012)

Annotation syntaxique dans LOCAS-F

- LOCAS-F Corpus (Degand, Martin & Simon, 2014), 3.5 heures, 43.000 tokens, équilibré pour genre de discours, 14 genres
- Motivation : confronter la segmentation prosodique et la segmentation syntaxique
- Références
  - Blanche-Benveniste et al. (1990), « Le français parlé. Études grammaticales »
  - Bilger & Campione (2002), « Propositions pour un étiquetage en 'séquences fonctionnelles' »
  - Guide de codage: Tanguy, Van Damme, Degand & Simon (2012)
- Deux niveaux d'analyse syntaxique:
  - Unité de rection
  - Séquence fonctionnelle

Annotation syntaxique dans LOCAS-F

Unités de rection

- une UR est organisée autour d'un noyau (en général un verbe tensé) et se compose de l'élément recteur et de ses dépendants
- des éléments non régis peuvent entrer linéairement dans la composition d'une UR (inserts, marqueurs de discours…)

Séquences fonctionnelles

- séquences maximales « qui représentent les constituants fonctionnels rencontrés dans les textes sans entrer dans le détail de leur composition » (Bilger & Campione 2002: 118)
- toute séquence peut faire l'objet
  - d'expansions: liste, coordination, reformulations, etc.
  - de mise en dispositif (extraction, binarisation, pseudoclivée…)
- toute unité (rection, séquence) peut être inachevée

Annotation syntaxique dans LOCAS-F

- **Unités de rection**

urv	unité de rection verbale (complète)
ure	unité de rection elliptique
ura	unité de rection averbale (nominale, adverbiale, etc.)

- **Séquences fonctionnelles**

SS	séquence sujet
SV	séquence verbe
SO	séquence objet
SR	séquence régie
SA	adjoint
insert	unité de rection qui interrompt une unité en cours

Exemple d'annotation syntaxique dans LOCAS-F

- <alors>md <bon>md <pour ce qui est de notre euh représentation de l'intonation>ag <en fait>md [(on reprend)SV (le terme de profil mélodique)SO]urv <c'est-à-dire qu'>md [(on considère)SV (que tout énoncé a un profil mélodique qui lui est associé)SO]urv [(c'est)SV (un terme qu'avait utilisé par Dell quatre-vingt-quatre)SO]urv (cnf-1)

- [(visibilité)SN]ura <première caractéristique>ag [(c'est)SV (la ///S visibilité de la langue)SO]urv ///T (cnf-4)

- $<alors>_{md}$ $<pour\ finir>_{ag}$ $[(on\ se\ dit)_{SV}]^{urv}$ $<ben>_{md}$ $[(on\ va\ aller\ voir)_{SV}$ (un autre $film)_{SO}]^{urv}$ $[(on\ n'allait\ pas)_{SV}]^{urv\text{-}l}$ $<enfin>_{md}$ $[(on\ va\ aller\ voir)_{SV}$ (un autre $film)_{SO}]^{urv}$ (conv-inf-1)

Annotation syntaxique dans Rhapsodie

- 33.000 tokens, annotation prosodique et syntaxique (Lacheret et al. 2014)
- Notions: **unités illocutoires** et **piles** (Gerdes & Kahane 2009, Kahane 2012)
- Annotation semi-automatique à l'aide de FRMG (de la Clergerie et al. 2009), corrigée manuellement
- 7 relations de dépendance principales
  - root: elements that are not governed by another element
  - sub: grammatical subjects of verbs
  - obj: direct objects of verbs.
  - obl: oblique complements of verbs, including indirect objects
  - ad: adjuncts to the verb
  - pred: all elements that form a complex predicate with a verb (past participles, verbal complements of modals, predicative adjectives …).
  - dep: all the dependents of non-verbal forms
  - junc: to link elements to junctors (= coordinating conjunctions), giving the asymmetrical annotation of coordination put forward by Mel'čuk (1988).

Annotation syntaxique dans Rhapsodie

- Relations paradigmatiques (piles)
  - Coordination
  - Intensification — *des dizaines et des dizaines d'années*
  - Disfluency — *c'était un un un un enfin une super expérience*
  - Double formulation — *c'est à dire*
  - Reformulation — *X je veux dire Y*
  - Hypéronyme — *et tout ça, et cetera*
  - Négociation — *des des Français enfin des Français*

Annotation syntaxique dans Rhapsodie

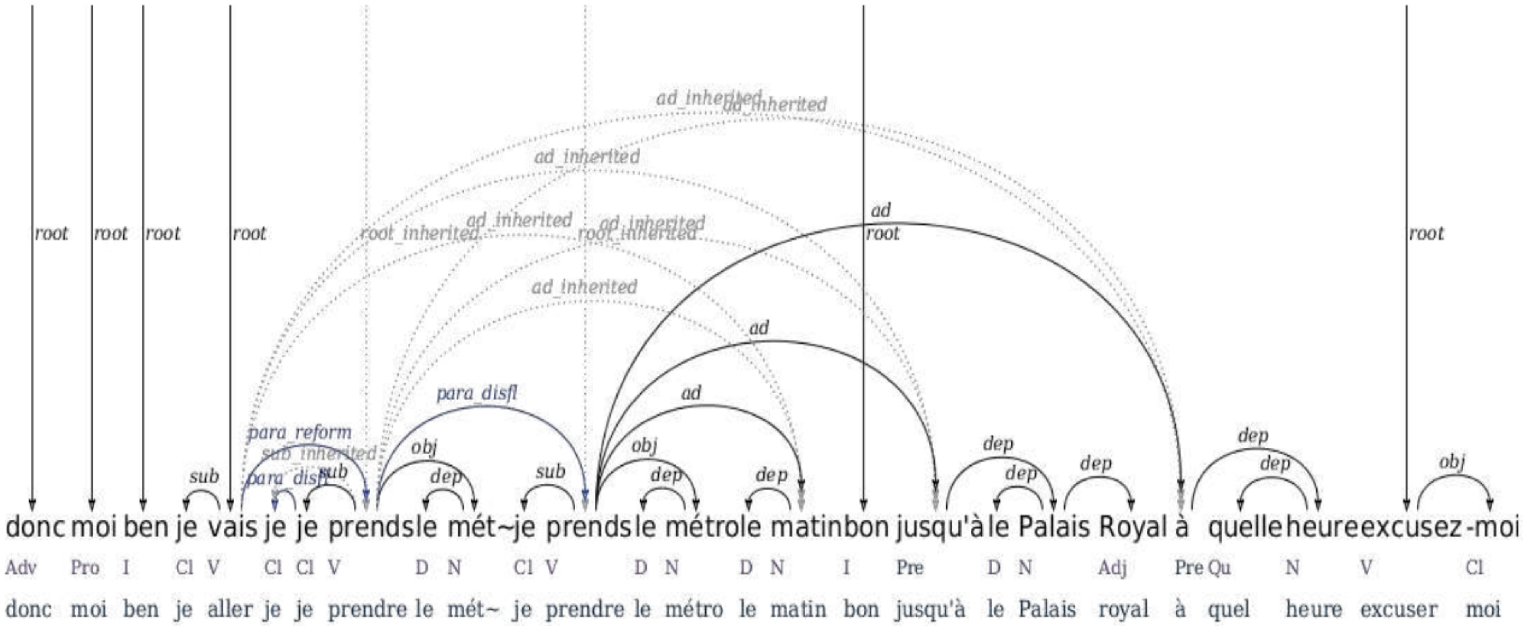


Figure 1. Dependency tree of the two turns: *$L1 donc < moi < "ben" { je vais | { je | je } prends le mét~ | je prends } le métro le matin "bon" jusqu' au Palais Royal //+ $L2 à quelle heure "excusez-moi"* // [Rhap-D0001, CFPP2000 corpus]
'$L1 so < me < "well" { I go | { I | I } take the met~ | I take the metro } in the morning "well" until Palais-Royal //+ $L2 at what time "excuse me" //'

Lacheret A., Kahane S., Beliao J., Dister A., Gerdes K., Goldman J.-P., Obin N., Pietrandrea P., Tchobanov A. (2014) Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French, *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, Reykjavik, Islande.

Quelques réflexions

- Les « séquences » de LOCAS sont parfois trop longues – information micro-syntaxique perdue
- L'annotation dans Rhapsodie et à la fois une annotation syntaxique et une annotation de phénomènes discursifs (speech acts)
- Motivation et ancrage des choix d'annotation ?...
- Lien avec des études en psycholinguistique ?... (sentence processing, eye tracking data)
- Possibilité de créer des outils automatiques ?...
- Comparabilité avec des analyses en dépendance pour l'écrit ?

Conclusion et perspectives

- Annotation morphosyntaxique robuste pour l'oral
- Avances sur la détection et l'annotation automatique des disfluences
- Proposition pour l'annotation macro-syntaxique:
  - Combiner une analyse en « chunks » motivée par des études psycholinguistiques + en cognition
  - … avec une analyse en dépendance de base (cf. Universal Dependency Tagset project)
  - Bien distinguer des différents niveaux d'annotation pour des différents types de phénomène
- Interopérabilité qui dépasse les frontières géographiques et de groupe de recherche

Références bibliographiques

Abeillé A., Clément, L., Toussenel, F. (2003). Building a treebank for French, in 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

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

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

Blanche-Benveniste C. (1990) "Un modèle d'analyse syntaxique 'en grilles' pour les productions orales", Anuario de Psicologia 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 2015, Dresde, Allemagne, 6-10 septembre 2015, 1849-1853.

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

Christodoulides G., Avanzi M. & J.-Ph. Goldman (2014). DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech, International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, 26-31 May 2014 : 3902-3907, www.corpusannotation.org/dismo.

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

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. Disponible 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", The Fourth Linguistic Annotation Workshop (LAW IV), Uppsala, Suède, 8 p.

Durand J., Laks B., Lyche C. (Eds.) (2009). Phonologie, variation et accents du français, Paris, Hermès

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

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, 101-116. icon

Simon A. C., Francard M., Hambye Ph. (2014, sous presse). The VALIBEL Speech Database, in J. Durand, U. Gut & G. Kristoffersen (éd.), The Oxford Handbook of Corpus Phonology. Oxford, Oxford University Press, 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, 431–438

Merci de votre attention

Au-delà du token
Expériences sur l'analyse syntaxique des corpus oraux avec l'annotateur multi-niveaux DisMo

George Christodoulides
Centre Valibel, IL&C, Université catholique de Louvain
Giulia Barreca
Université Paris Ouest Nanterre La Défense & Université catholique de Milan

george@mycontent.gr, giulia.barreca@gmail.com