UMONS
Université de Mons

Faculté de Psychologie et des Sciences de l'Education

Alignement automatique et annotations supplémentaires du corpus PFC
Nouvelles pistes de recherche

George Christodoulides

Service de Métrologie et des Sciences du Langage, Université de Mons
george@mycontent.gr

langage
INSTITUT DE RECHERCHE EN SCIENCES ET TECHNOLOGIES DU LANGAGE DE L'UMONS

Journées FLORAL-(I)PFC 2018
Paris, 22-27 novembre 2018

Plan

- Introduction
- Traitement du corpus PFC
  - Alignement automatique
  - Annotation morphosyntaxique – syntaxique
  - Annotation prosodique
  - Présentation du corpus en ligne
- Pistes de recherche qu'on peut aborder avec les nouvelles annotations
- Diffusion des outils et du corpus
- Conclusion

Traitement automatique du corpus PFC

- Un projet pour enrichir les données PFC
  - Commencé en 2014 sous l'impulsion de Bernard Laks
  - Collaboration avec Giulia Barreca : annotation morphosyntaxique, corrections
  - Le développement des outils a avancé indépendamment du travail sur les données PFC (et sur d'autres pistes que celles qui étaient identifiées au début)
  - Traitement des données originales (non-anonymisées) du corpus
  - En parallèle : anonymisation => besoin de combiner ces deux projets
  - Développements dans la communauté : ORTOLANG, CORLI, Comité pilotage PFC
- Un projet pour enrichir les autres grands corpus du français diffusés actuellement sous une licence ouverte comme Creative Commons (p.ex. ESLO, ORFEO) est envisageable

Traitement automatique du corpus PFC

1. Pré-traitement des données
    1. Amélioration de la qualité des fichiers sons
    2. Corrections dans les fichiers de transcription
    3. Séparation des locuteurs
2. Annotation morphosyntaxique
3. Alignement automatique
    1. Phonétisation
    2. Alignement
4. Annotation prosodique
5. Annotation syntaxique

Les points d'enquête traités

Points d'enquête		Tokens (min)
FRANCE		
11a	Douzens	38.693
12a	Rodez	42.233
13a	Marseille Centre Ville	32.000
13b	Aix-Marseille	18.164
21a	Dijon	31.233
31a	Toulouse	48.533
38a	Grenoble	19.989
42a	Roanne	24.868
44a	Nantes	24.180
50a	Brécey	36.009
54b	Ogéviller	26.482
61a	Domfrontais	68.945
64a	Biarritz	36.519
69a	Lyon	23.093
75c	Paris Centre Ville	54.097
75x	Aveyronnais à Paris	49.711
80a	Amiens	
81a	Lacaune	22.115
85a	Vendée	24.610
91a	Brunoy	24.917
92a	Puteaux-Courbevoie	11.315
974	Île de la Réunion	27.939

Points d'enquête		Tokens (min)
AFRIQUE		
aba	Béjaia	24.607
aca	Chlef	23.141
bfa	Burkina Faso	25.451
cia	Abidjan	44.347
cya	Cameroun	14.075
maa	Bamako	76.577
rca	Bangui	42.448
sna	Sénégal Dakar	23.479
CANADA		
caa	Peace River	22.455
cqa	Québec ville	27.967
cqb	Saguenay	119.348
BELGIQUE		
bga	Gembloux	31.619
bla	Liège	28.153
bta	Tournai	26.611
SUISSE		
sca	Neuchâtel	60.547
sga	Genève	41.104
sva	Nyon	50.461
TOTAL CORPUS PFC		**1.368.035**

1. Pré-traitement des données

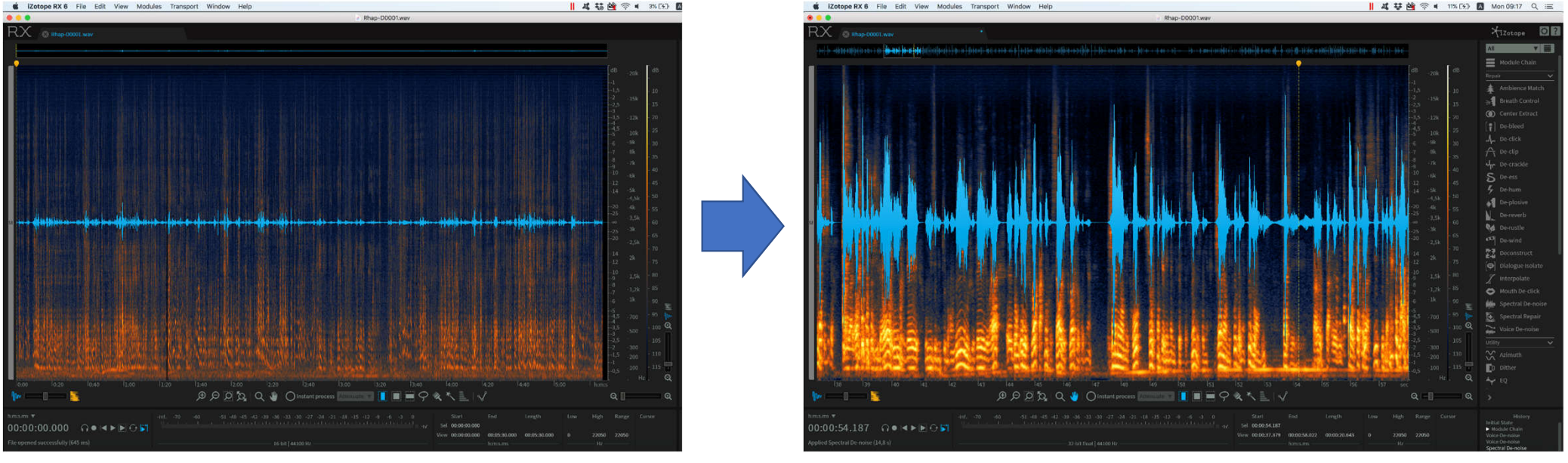
1.1 Amélioration de la qualité des fichiers sons

Avec le logiciel iZotope RX 6 Audio Editor (logiciel commercial)

- De-clip (restore clipped samples at high quality, new maximum level -1dB)
- De-click (remove random clicks)
- De-hum (remove 50 Hz power line noise and its harmonics)
- De-reverb (remove reverberation, in light vocal processing mode)
- Voice De-noise (adaptive noise reduction)
- De-plosive (light removal of hard microphone puffs)
- Phase correction (if needed)
- Equaliser Match (using the “full dialogue” preset)
- Leveler (normalisation of audio levels, respecting dialogue dynamics)

1. Pré-traitement des données

1.1 Amélioration de la qualité des fichiers sons



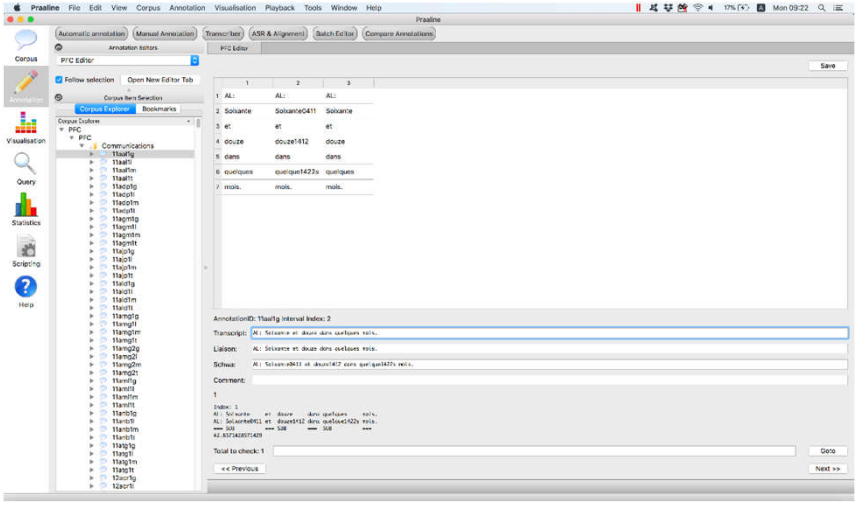
1. Pré-traitement des transcriptions

1.2 Prétraitement des TextGrids PFC

- Selon le protocole PFC :
  - La transcription orthographique indique toujours le locuteur
  - Les chevauchements sont transcrits entre chevrons
  - Il y a toujours 3 tiers (transcription, schwa, liaison) – avec le même nombre de tokens
  - On utilise des parenthèses pour inclure des commentaires ou évènements
  - Prononciation spéciale: SAMPA entre crochets
- Il est difficile pour un transcripteur de respecter ces règles systématiquement
- Par conséquent, on a besoin de scripts de contrôle de qualité
- Éventuellement, apporter des corrections à la main

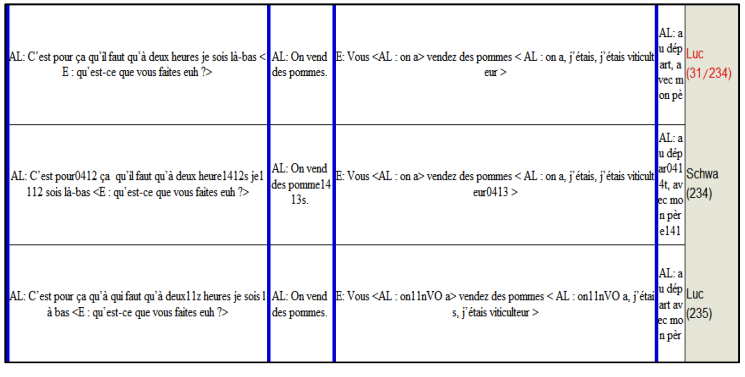
1. Pré-traitement des transcriptions

- Importation de tous les TextGrids dans une base de données *Praaline*
- Scripts: vérifier nombre de locuteurs, correspondance de tokens dans les trois tiers, conventions de transcription, ponctuation, etc.
- Correction automatique si possible
- De plus, 2700 corrections apportées manuellement
- Interface d'édition spécialisé dans *Praaline*

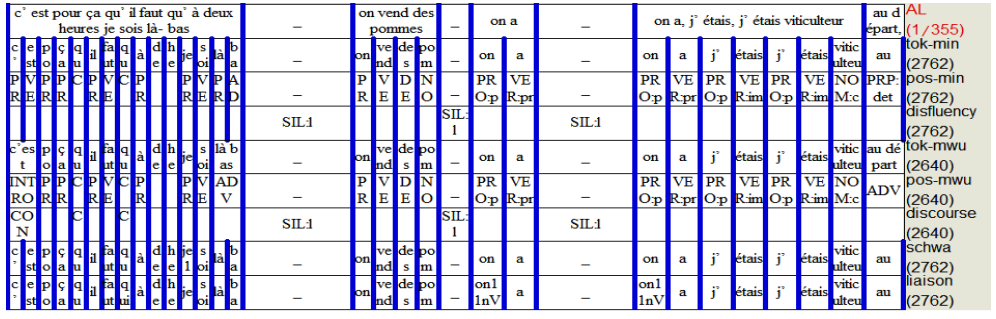


1. Pré-traitement des transcriptions

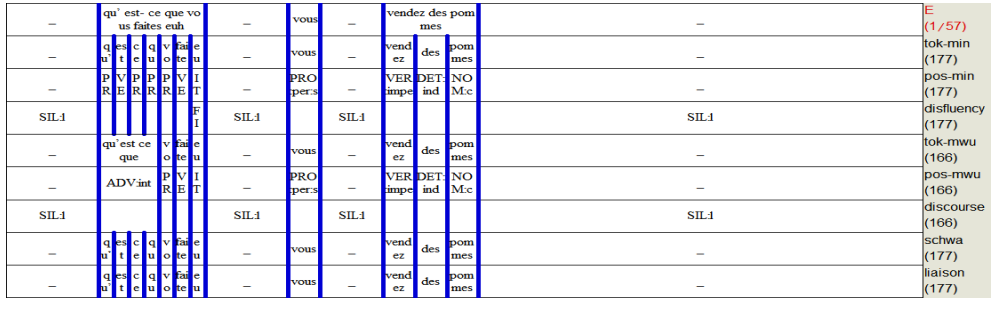
1.3 Séparation des locuteurs



TextGrid PFC



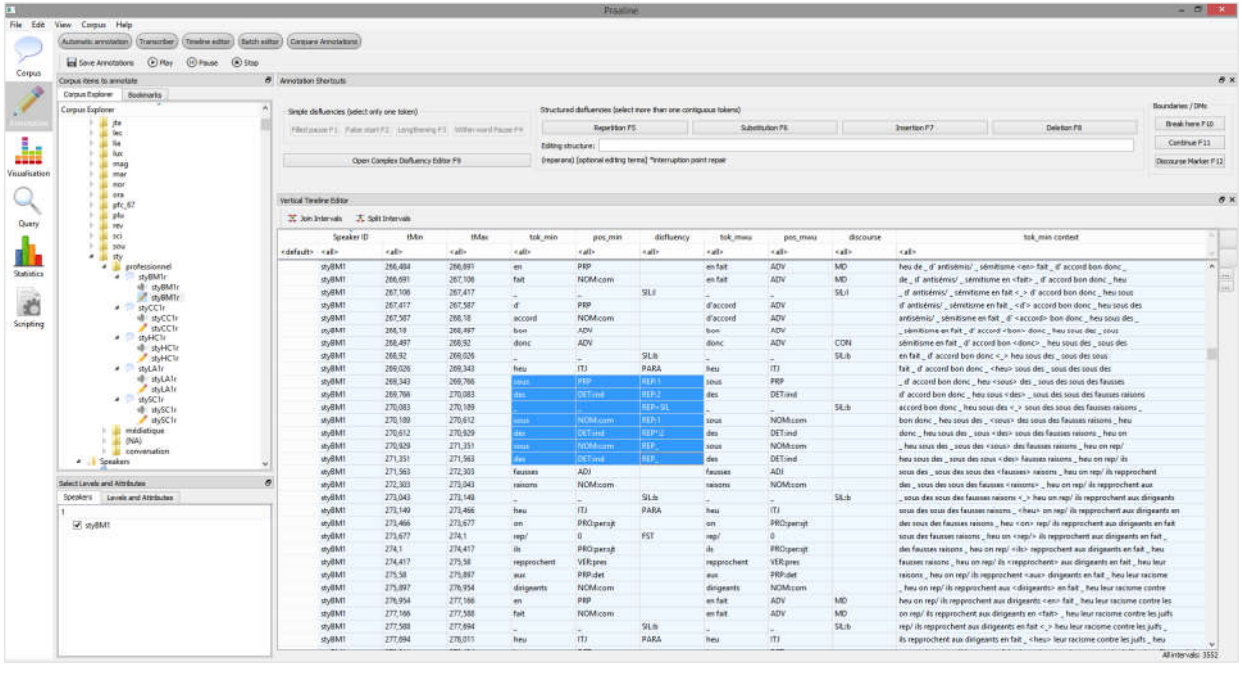
Locuteur AL (à aligner)



Enquêteur E (à aligner)

2. Annotation morphosyntaxique

- Annotateur DisMo : annotation morphosyntaxique (part of speech), détection des disfluences



Les conventions de transcription PFC aident à la détection des disfluences (p.ex. utilisation de la virgule pour noter les interruptions syntaxiques et les répétitions)

L'annotation morphosyntaxique contribuera à la phonétisation

3. Alignement Automatique

- Un outil d'alignement automatique texte/son est structuré en deux parties (phases de traitement de données):
  - en amont: pré-traitement des transcriptions, phonétisation, dictionnaire de prononciation, application des conventions de transcription
  - en aval: un système de reconnaissance automatique de la parole qui assure l'alignement texte/son, suivi par un système de syllabation
- Un système idéal
  - s'appuie sur les conventions de transcription
  - choisit la phonétisation correcte au cas où plusieurs prononciations sont possibles
  - propose des modèles acoustiques robustes
  - adapte son modèle acoustique au locuteur et aux conditions d'enregistrement
  - gère les cas de chevauchement des locuteurs
  - peut traiter les données automatiquement et localement (i.e. sur l'ordinateur de l'utilisateur, sans l'obliger de partager ses données avec un fournisseur de services externe)

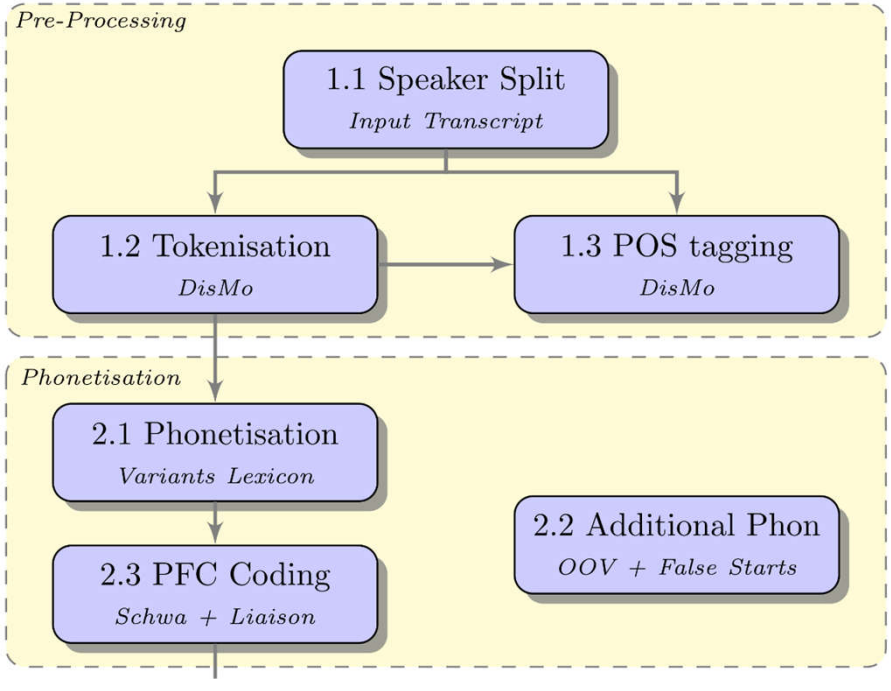
Outils d'alignement texte/son pour le français

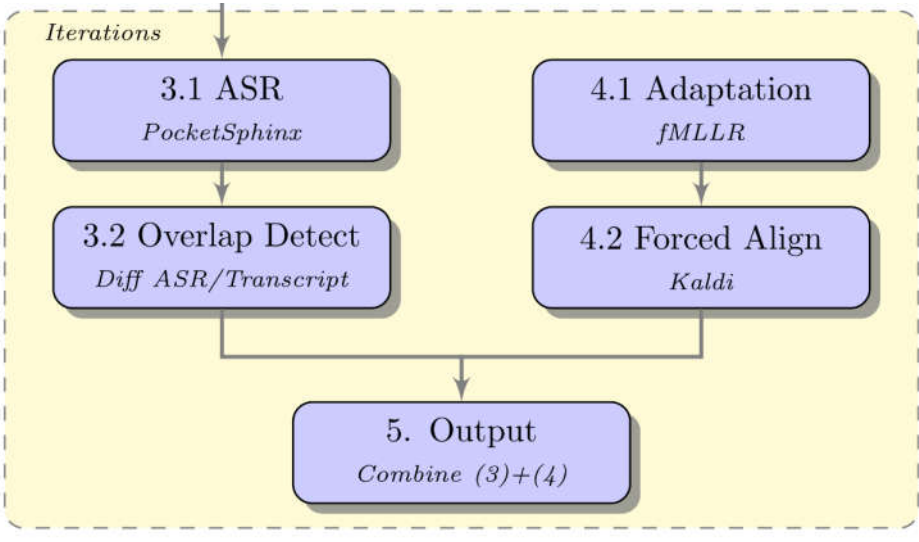
Outil	Traitement en amont	Traitement en aval	Avantages/Inconvenients
EasyAlign	Scripts sous Praat, utilise un phonétiseur pour DOS	HTK modèles 1-phone (Hvite)	Simple, largement utilisé Vieille technologie Pas d'adaptation
SPPAS	Interface graphique Python Lexique de prononciation	HTK modèles 3-phone (Julius ASR)	Meilleurs modèles en général, pas d'adaptation
Train&Align	Interface Web Pas de phonétisation	HTK (1-phone, 3-phone)	Adaptation au locuteur Seulement sur le web
Montreal Forced Aligner	Scripts Python, pas d'interface graphique, pas de phonétisation	Kaldi (1-phone, 3-phone, speaker-adapted)	Adaptation, moderne, pas de dictionnaire
SailAlign	Scripts Python et Perl, pas de phonétisation	HTK	Long Sound Alignment seulement

3. Alignement Automatique

- Nous proposons un système qui combine:
  - un dictionnaire des variantes de prononciation, basé sur GLÀFF (gros lexique à tout faire du Français, Hathout *et al.*, 2014) et d'autres sources
  - la sortie du système de reconnaissance de la parole Kaldi (Povey *et al.*, 2011) utilisant des modèles de 1-phones, 3-phones et 3-phones après adaptation
  - la sortie du système de reconnaissance de la parole PocketSphinx (Walker *et al.*, 2004) pour traiter les cas de chevauchement des locuteurs
- Les modèles acoustiques produits sont compatibles avec le MFA (McAuliffe *et al.*, 2017)
- Nous avons appliqué ce système sur la totalité des données du corpus PFC.

3. Alignement Automatique





3. Alignement Automatique

3.1 Phonétisation

- Sources d'information
  - Lexique GLÀFF : variantes de prononciation
  - Étiquette morphosyntaxique DisMo
  - Annotation PFC : liaison et schwa
- On utilise l'annotation PFC pour réduire les variantes possibles
  - Ceci implique que l'alignement automatique sera cohérent avec l'annotation PFC
  - Exemple: Soixante0411 et douze1412 dans quelque1422s mois
- Mots hors lexique (out-of-vocabulary words) : modèle de phonétisation statistique (G2P) avec l'outil *Phonetisaurus*
- Faux départs: script qui cherche dans le lexique toutes les phonétisations possibles de cette suite de caractères en position initiale

3. Alignement Automatique

- Système basé sur *Kaldi*
  - Entrainement d'un modèle de monophones
  - Entrainement d'un modèle de triphones
  - Adaptation du modèle au locuteur + caractéristiques de l'enregistrement
- Paramètres : MFCC + Deltas
- Cepstral mean and variance normalization (CMVN)
- Adaptation : Feature space Maximum Likelihood Linear Regression (fMLLR)
- Entrainement d'un modèle d'alignement *par région et style de parole*
- « Vérification croisée » : on utilise chaque modèle pour aligner toutes les données

3. Alignement Automatique

Alignement des chevauchements

- PocketSphinx: reconnaissance automatique (n-best) sur l'énoncé
- On essaye de trouver des mots qui peuvent servir comme points d'ancrage pour améliorer la précision des frontières qui délimitent la partie du signal qui correspond à chaque locuteur
- Détection des énoncés où les locuteurs prennent la parole successivement (pas des vrais chevauchements)
- Cette partie du processus pourrait être améliorée (p.ex. utilisation des paramètres prosodiques pour séparer les locuteurs)

3. Alignement - évaluation

Region to align	Text 11a	12a	13a
11a		84.8%	84.8%
12a	87.3%		89.7%
13a	89.9%	90.1%	

Region to align	Conversation 11a	12a	13a
11a		81.7%	81.8%
12a	83.5%		84.2%
13a	82.8%	83.3%	

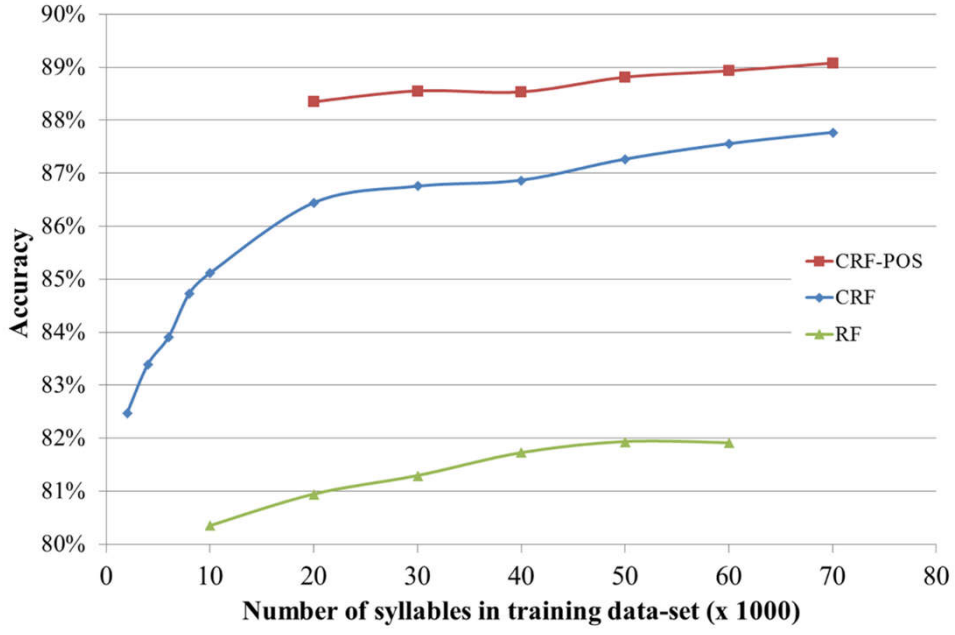
- On ne possède pas d'alignement de référence (étalon d'or)
- Chaque modèle, entrainé sur le point d'enquête X, est utilisé pour aligner tous les autres points d'enquête => créer des groupes ?
- Indiquer les énoncés qui posent problème ? (potentiellement mal transcrits)

4. Annotation prosodique

- Un système pour l'annotation automatique de la proéminence et des frontières prosodiques
- Méthode
  - Extraction des paramètres prosodiques
    Durée de la syllabe / du noyau (ms), pitch minimun, maximum, moyen (f0), mouvement mélodique (Δf0) intrasyllabique et intersyllabique, intensité (pics dans noyau syllabique), présence et durée de pause après la syllabe, structure syllabique, présence du schwa final, token (mot) correspondant à la syllabe, étiquette morphosyntaxique de ce mot, position de la syllabe dans le mot
  - Annotation sur la base des modèles CRF (annotation des séquences)

4. Annotation prosodique

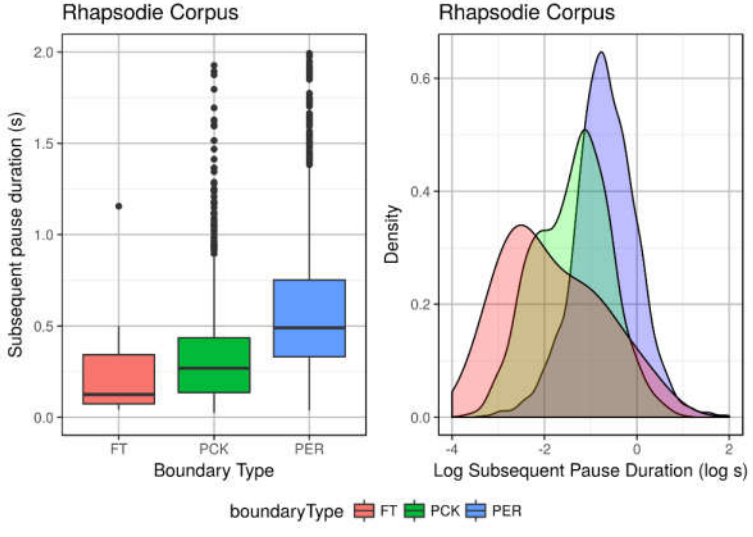
- Proéminence (cf. Christodoulides & Avanzi 2014, *An evaluation of machine learning methods for prominence detection*, Proc. Interspeech, 116-119)



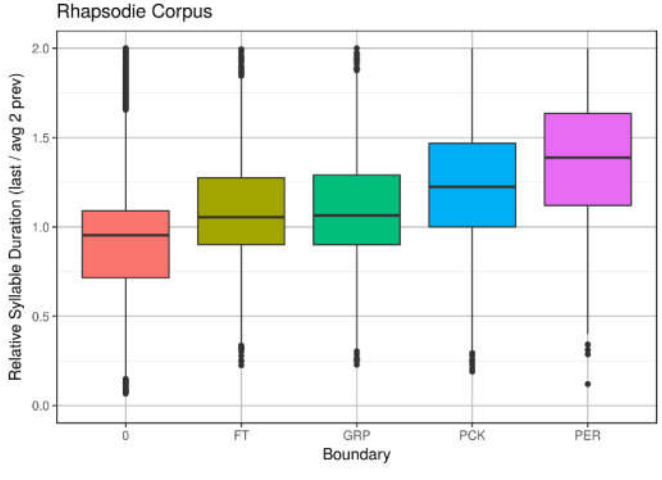
4. Annotation prosodique

Frontières prosodiques

- Entraînement d'un modèle statistique CRF sur le corpus Rhapsodie (Lacheret *et al.*, 2014) – le seul corpus avec une licence suffisamment libre pour diffuser l'outil!

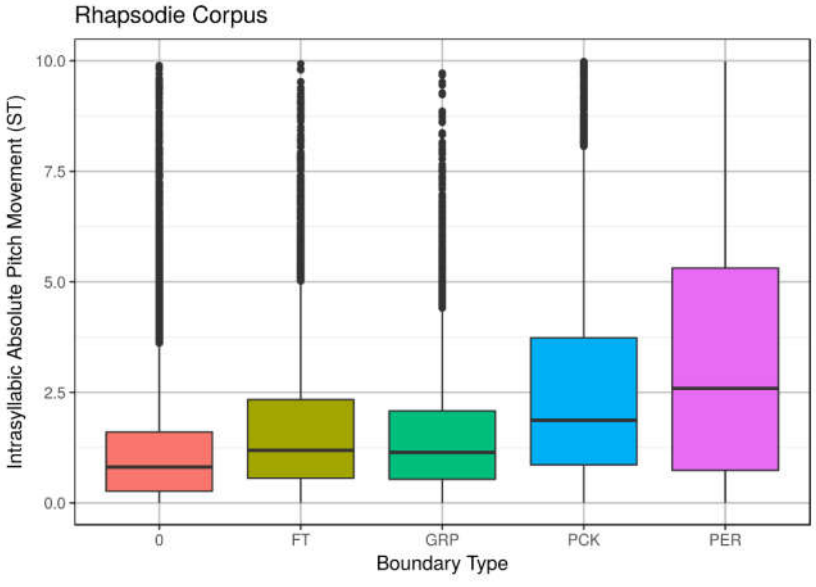


Durée de pause après la syllabe de frontière

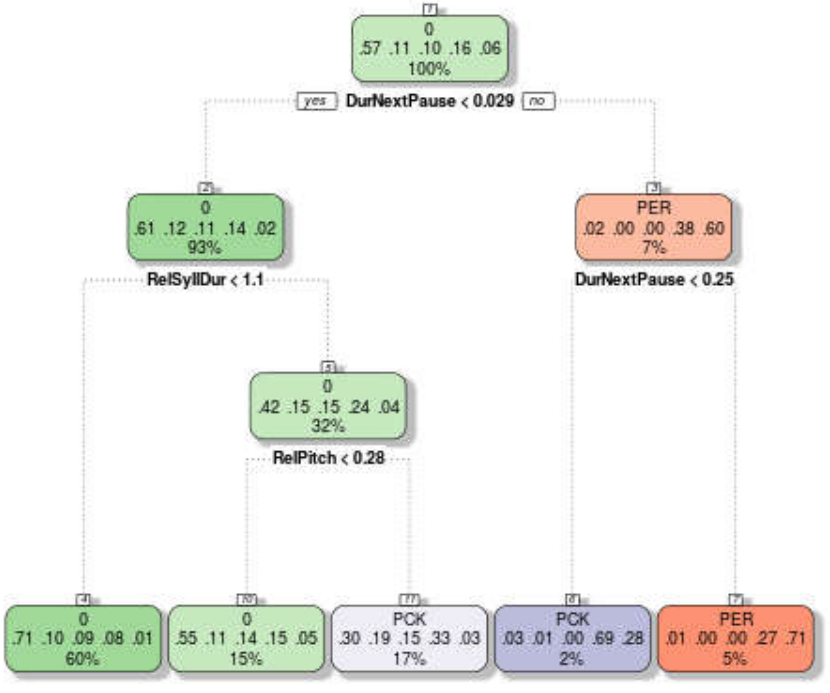


Durée relative de la syllabe de frontière

4. Annotation prosodique



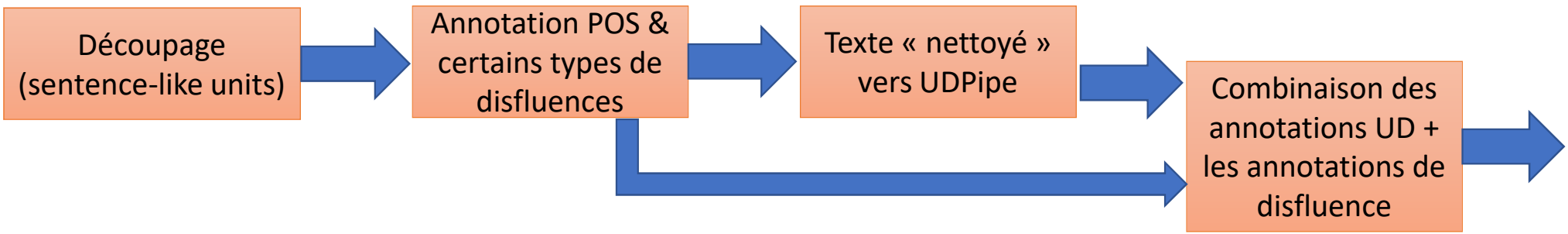
Mouvement intra-syllabique



Classification Tree for Rhapsodie Boundary Type

5. Annotation syntaxique et morphosyntaxique

- Nous avons poursuivi le développement de l'annotateur morphosyntaxique *DisMo* (Christodoulides & Barreca, 2017), en profitant des corpus annotés publiés par le projet Universal Dependencies (de Marneffe *et al.*, 2014)
- Méthode:

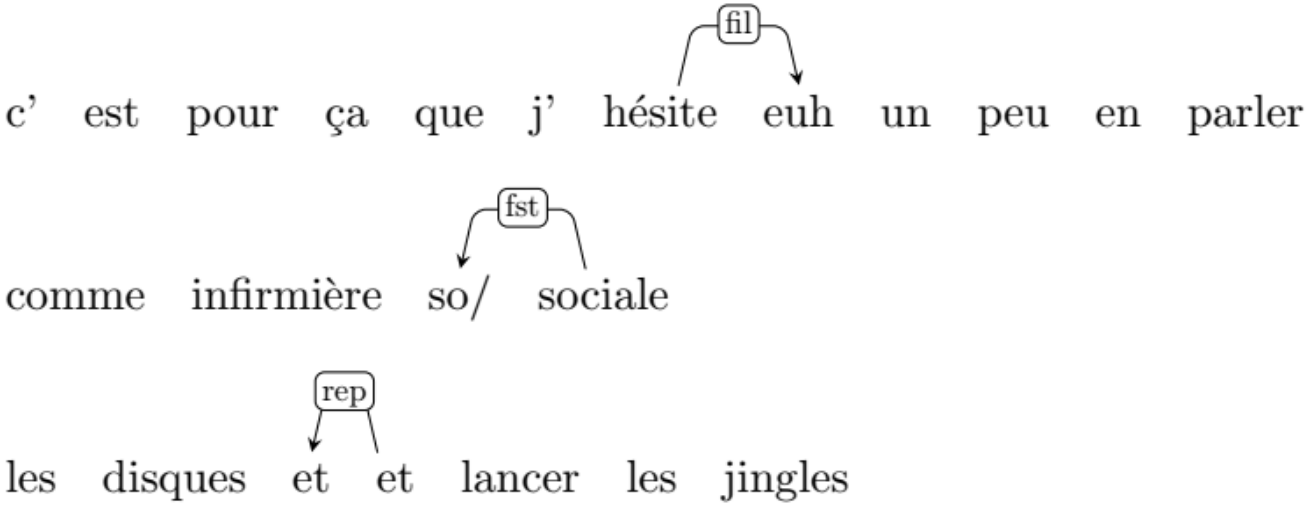


5. Annotation syntaxique et morphosyntaxique

Niveau 1: Disfluences **simples** = affectent un seul token		
FIL	Pause remplie	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: Disfluences **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

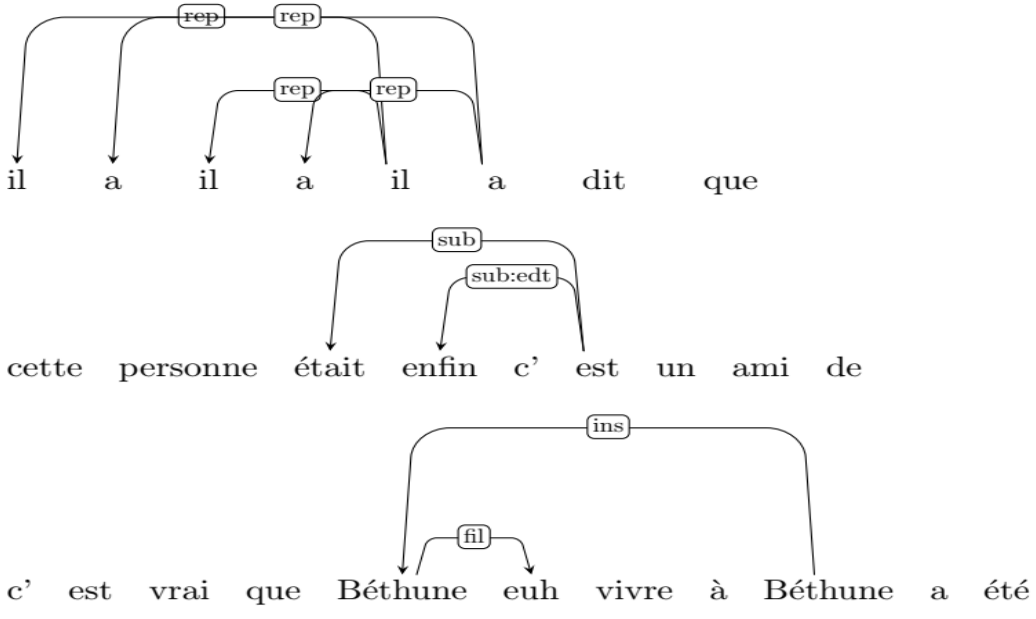
5. Annotation syntaxique et morphosyntaxique

- Représentation de cette annotation en dépendances, afin de les combiner avec le système UD – aller-retour entre les deux systèmes



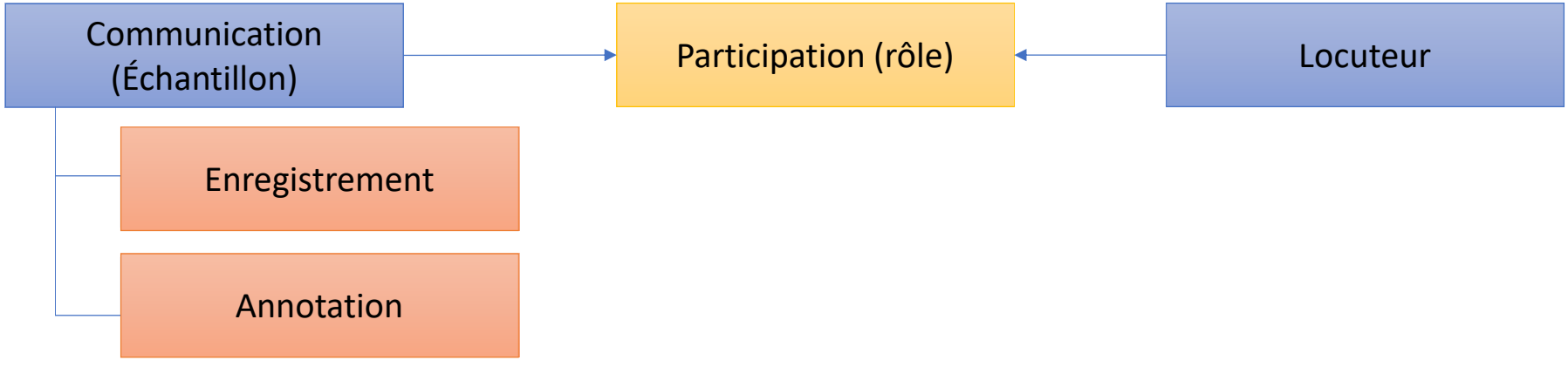
5. Annotation syntaxique et morphosyntaxique

- Représentation de cette annotation en dépendances, afin de les combiner avec le système UD – aller-retour entre les deux systèmes



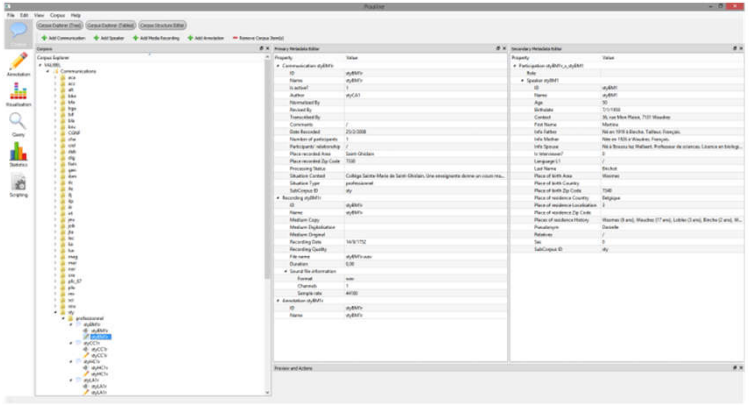
6. Diffusion du corpus annoté en ligne

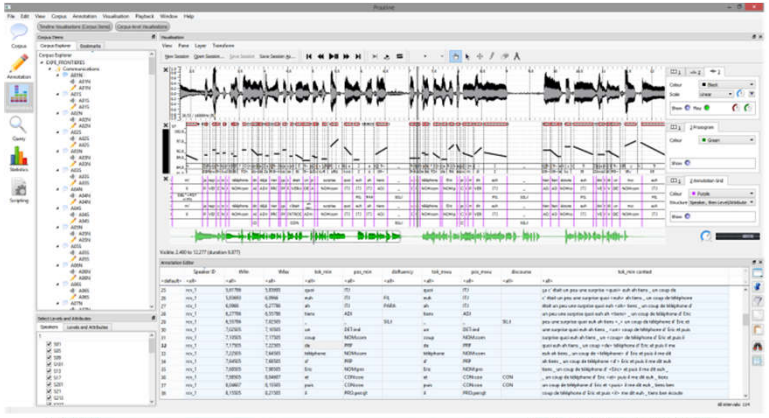
- Générer un site web pour un corpus via *Praaline*. Le site web généré utilise le système de gestion de contenu *Django*
- Présentation des métadonnées

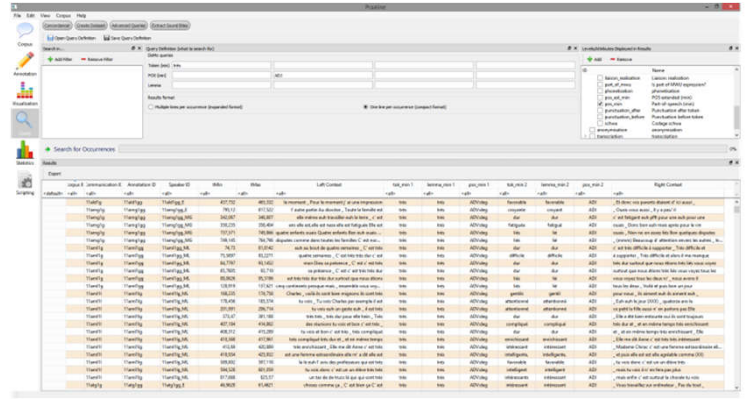


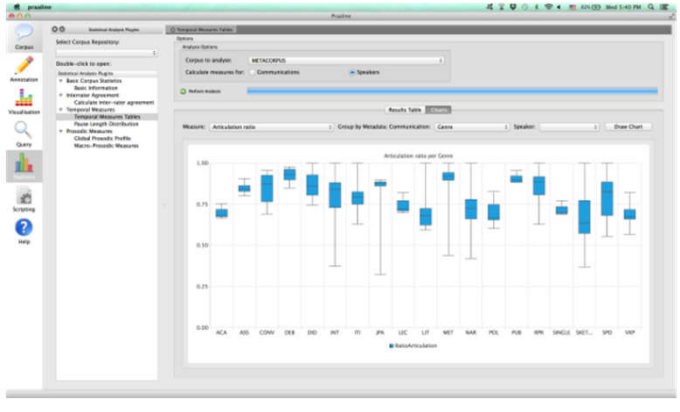
- Une page par annotation (cf. projet « Reciprosody »)

Comment accéder à la boîte à outils ? www.praaline.org (et sur Ortolang)









Perspectives: études

(sujet discuté aux Journées PFC-FLORAL 2018)

Conclusion et perspectives

- Diffusion du Corpus PFC (partie anonymisée vs non-anonymisée) enrichi

- Diffusion des outils

- Standards XML pour le stockage pérenne

UMONS
Université de Mons

Faculté de Psychologie et des Sciences de l'Education

Merci de votre attention

Alignement automatique et annotations supplémentaires du corpus PFC : nouvelles pistes de recherche

George Christodoulides
Service de Métrologie et des Sciences du langage, Université de Mons
george@mycontent.gr

langage
INSTITUT DE RECHERCHE EN SCIENCES ET TECHNOLOGIES DU LANGAGE DE L'UMONS

Journées FLORAL-(I)PFC 2018
Paris, 22-27 novembre 2018