Analyse des données IPFC : développement de l'approche par codage

Isabelle Racine[1], Sylvain Detey[2] et Julien Eychenne[3]

[1]ELCF, Université de Genève
[2]SILS, Université Waseda
[3]Hankuk University of Foreign Studies

« Interphonologie du français contemporain : corpus oraux en L2 et évaluation »

Journées IPFC2013 - 9 et 10 décembre 2013 – Maison de Norvège, Cité Internationale, Paris

FNSNF FONDS NATIONAL SUISSE DE LA RECHERCHE SCIENTIFIQUE

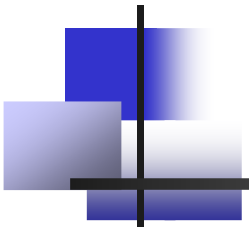
UNIVERSITÉ DE GENÈVE

JSPS

WASEDA University

Objectifs

- Présenter la méthode d'analyse des données développée pour IPFC
- Présenter le développement des outils dédiés à cette cette analyse
- Illustrer cette méthode de manière concrète à travers le cas de la liaison chez les apprenants de deux populations du projet IPFC

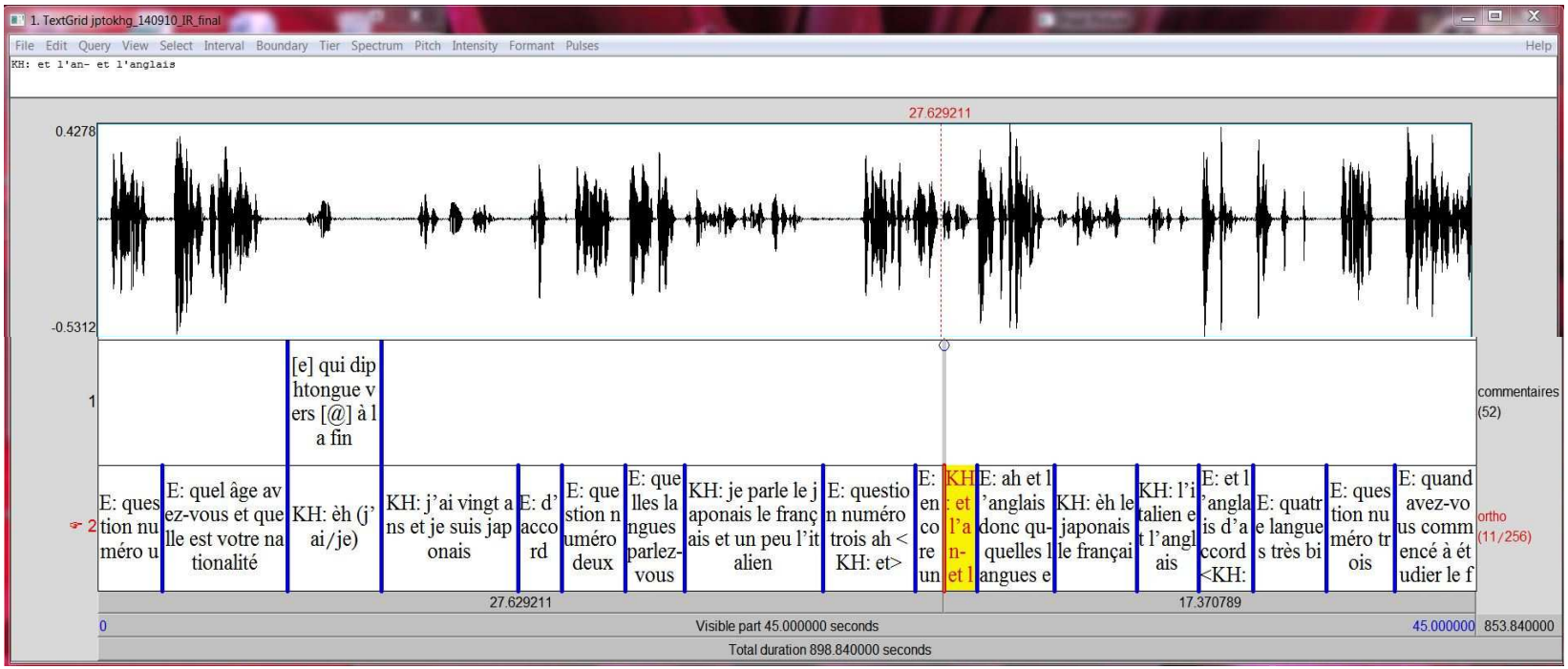


Plan

- Rappel: l'approche par codage pour les données IPFC
- Les différents codes
- Dolmen et ses différentes interfaces IPFC: illustration avec la liaison
- La comparaison inter-codeurs/codeuses
- Conclusion
- Perspectives

L’approche par codage

- Stade 1: Transcription orthographique des données sous Praat (Boersma & Weenink, 2009) avec alignement son-texte et conventions de transcription adaptées à la parole en L2 (Racine *et al.*, 2011)



L'approche par codage

- Après la transcription orthographique, quel traitement?
- Stade 2: le codage
- Dans PFC (corpus de locuteurs natifs, Durand *et al.*, 2009) ⇨ codage du schwa et de la liaison avec insertion de symboles alphanumériques pour coder la réalisation, le contexte environnant, etc.
- Coder un élément permet de combiner:
  1. Éléments descriptifs (p. ex. cible, contexte gauche/droit)
  2. Évaluation perceptive par codeur (p. ex. degré de conformité de la réalisation par rapport à la cible, degré de nasalité d'une voyelle, présence/absence de la liaison/schwa, etc.)

  ⇨ Traitement automatique et comparable des données et obtention de statistiques descriptives
- Approche par codage ⇨ approche perceptive à mi-chemin entre analyse phonologique grossière (ex. substitution / effacement / insertion) et analyse phonétique (acoustique) fine du signal (pour un développement plus détaillé, cf. Detey, 2012)

Les différents codes

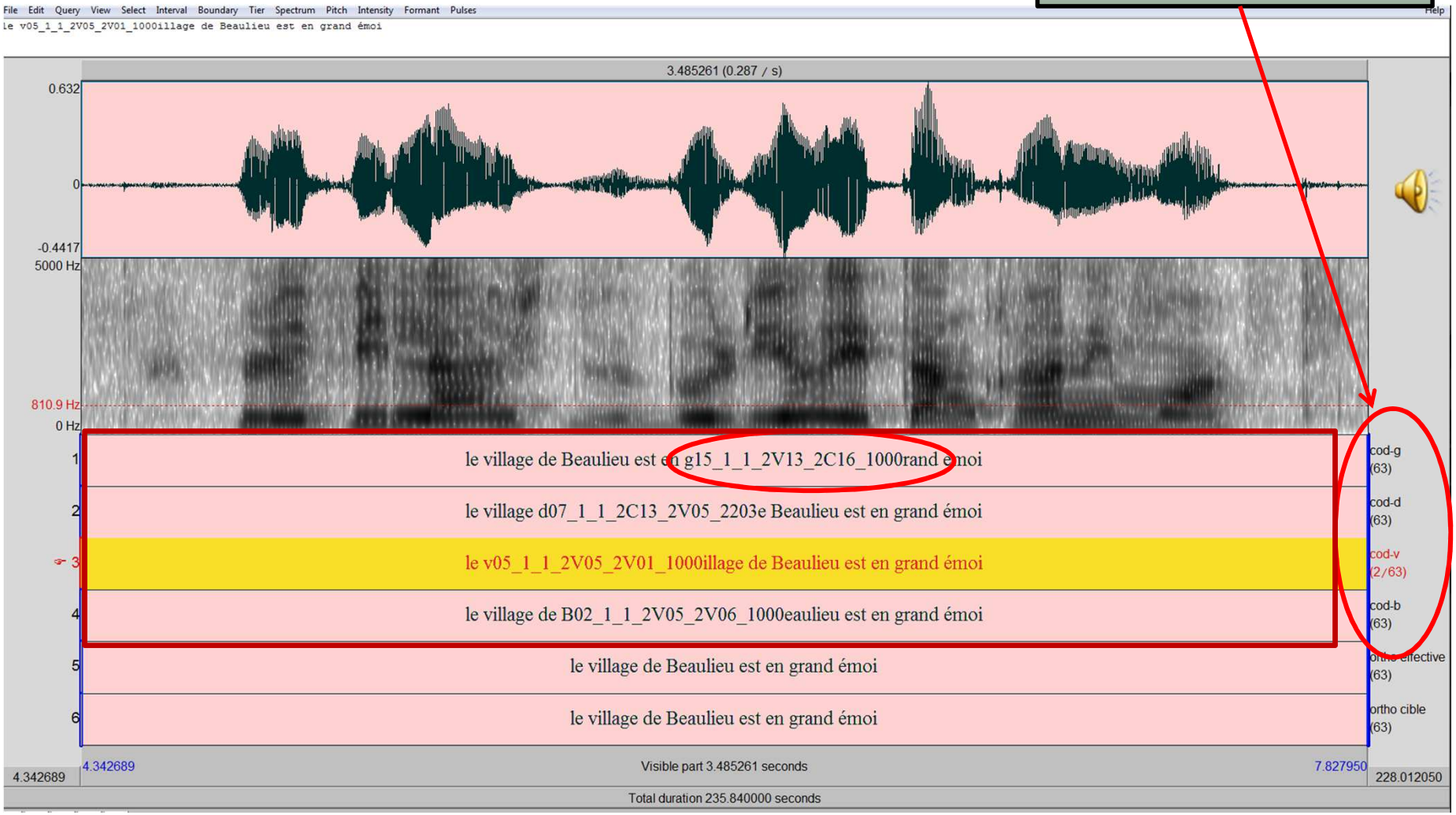
- Dans cette optique, développement successifs de différents codes alphanumériques pour:
  - Les voyelles nasales (cf. Detey, Racine & Kawaguchi, à paraître)
  - Les voyelles orales (cf. Detey & Racine, 2013)
  - La liaison (cf. Racine & Detey, 2012; Detey *et al*., to appear)
  - Les consonnes (en préparation)
- Deux points à souligner:
  - Importance d'avoir une cohérence globale entre les différents codes IPFC
    - Pour les codeurs/codeuses, 1 seule «logique»
    - Pour le développement de l'outil d'extraction, 1 seule «logique»
    - Recoupements entre certains champs appartenant à deux codes différents (p. ex. voyelles nasales et liaison)
  - Comme il s'agit d'une analyse perceptive, importance cruciale de garder un lien permanent entre son et transcription-codage

    ↳ codage sous Praat, dans des tires dédiées (idem PFC)

Les différents codes

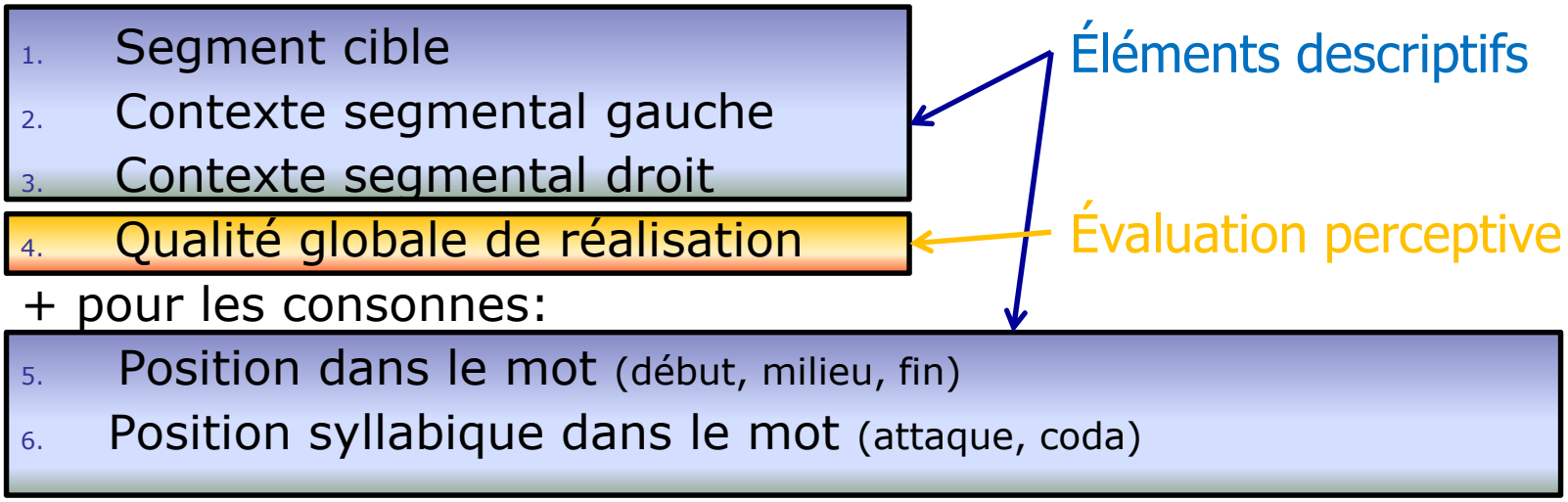
Importance de la nomenclature des différentes tires ⇨ utilisée pour l'analyse dans Dolmen

- Le codage des consonnes:

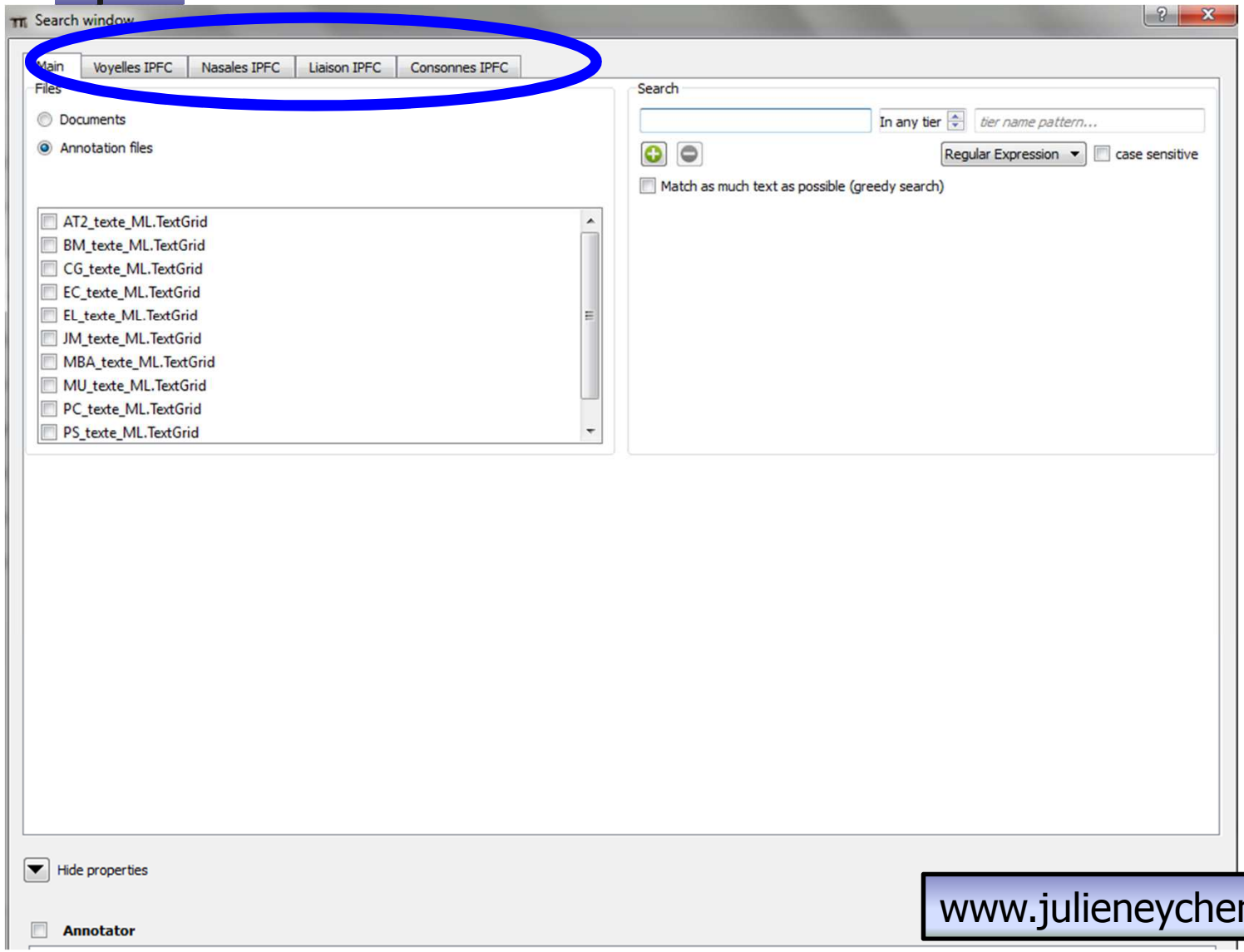


Les différents codes

- Structure de base du code:
  Une tire par segment (sauf v. moyennes, nasales et liaison):



Dolmen-IPFC: la liaison



Dolmen-IPFC (développé par J. Eychenne):

- Outil permettant d'effectuer des requêtes sur la base du codage

- Interaction avec Praat (Lecture/écriture de fichiers TextGrid + ouverture de fichiers dans Praat)

www.julieneychenne.info/dolmen

Dolmen-IPFC: la liaison

Statistiques descriptives de la liaison dans 20 textes IPFC (10 apprenants espagnols et 10 apprenants grecs chypriotes, même codeuse):

	Corpus ESP	Corpus GREC
Nombre :		
Nb d'occurrences	357	366
Nb de liaisons réalisées	214 (59.94%)	190 (51.91%)
Liaison avec enchaînement :		
Liaisons enchaînées	149 (69.62%)	164 (87.89%)
Liaisons non enchaînées	65 (30.38%)	26 (12.11%)
Nature de la consonne :		
Conforme à la cible	112 (52.34%)	172 (90.52%)
Partiellement conforme	66 (30.84%)	5 (2.64%)
Non conforme avec consonne présente	35 (16.36%)	13 (6.84%)
Non conforme avec consonne épenthétique	1 (0.46%)	0 (0%)
Consonne de liaison :		
/z/	85 (39.72%)	77 (40.53%)
/n/	71 (33.18%)	65 (34.21%)
/t/	55 (25.70%)	47 (24.74%)
/R/	3 (1.40%)	1 (0.52%)

Dolmen-IPFC: la liaison

Exportation des données en format csv:

	A	B	C	D	E	F	G	H
1	Fichier	Début	Fin	Contexte gauche	Cible	Catégorie du mot liaisonnant	Catégorie du mot suivant	Nb. de syll. et V. du mot l
2	PC_texte_ML.TextGrid	179.22	183.552	Ministre ¤ quelques fanatiques	11	NOM	AUX	
3	AT2_texte_ML.TextGrid	150.186	152.798	teur ¤ indiqueraient ¤ que des	11	DET	NOM	
4	MU_texte_ML.TextGrid	122.865	125.975	ge entier ¤ de plus ¤ quelques	11	DET	NOM	
5	Constantinou_4_texte_ML.TextGrid	57.1059	62.3031	ne cesse de baisser depuis les	11	DET	NOM	
6	Antoniou_4_texte_ML.TextGrid	157.763	159.572	impasse stupide ¤ il il s'est	30	AUX	PRP	
7	EL_texte_ML.TextGrid	57.802	60.1369	Ministre ¤ lassé des circuits	11	NOM	ADJ	
8	PC_texte_ML.TextGrid	79.0154	82.2341	se de baisser ¤ depuis les les	11	DET	NOM	
9	AT2_texte_ML.TextGrid	61.2264	63.7903	RP_100_11_1_0 en revanche très	10	ADV	ADJ	
10	MU_texte_ML.TextGrid	42.4888	48.8823	ituels qui tournaient toujours	10	ADV	PRP	
11	PS_texte_ML.TextGrid	170.792	174.268	son village ¤ était vraiment	30	ADV	DET	
12	Antoniou_4_texte_ML.TextGrid	69.9008	77.0851	s barrages chaque fois que les	11	DES	NOM	
13	Constantinou_4_texte_ML.TextGrid	139.224	143.321	Ministre ¤ quelques fanatiques	11	NOM	AUX	
14	Ioannou_4_texte_ML.TextGrid	173.445	176.91	on village était vri- vraiment	30	ADV	DET	
15	Georgiou_4_texte_ML.TextGrid	171.155	174.62	ire de Beaulieu ¤ ne sait plus	10	ADV	PRP	
16	PS_texte_ML.TextGrid	84.0105	85.8797	manifestent leur colère ¤ d'un	20	DET	ADJ	1VO
17	Antoniou_4_texte_ML.TextGrid	6.7246	10.3751	moi ¤ le Premier Ministre a en	20	EXF	EXF	1VN
18	Ioannou_4_texte_ML.TextGrid	80.783	82.8558	barrages ¤ chaque fois que les	11	DET	NOM	
19	Epifaniou_4_texte_ML.TextGrid	67.6083	70.2069	r les manifestations ¤ qui ont	31	AUX	PPA	
20	Georgiou_4_texte_ML.TextGrid	70.4544	75.7445	ne cesse de baisser depuis les	11	DET	NOM	
21	PS_texte_ML.TextGrid	22.0416	23.5819	n vin blanc sec ¤ ses chemises	11	NOM	PRP	
22	MBA_texte_ML.TextGrid	127.072	132.71	identité risquent de provoquer	40	INF	DET	
23	BM_texte_ML.TextGrid	145.381	149.407	e sentiment de se trouver dans	10	PRP	DET	
24	Nikolaou_4_texte_ML.TextGrid	184.745	190.746	ue se plutôt que de se trouver	40	INF	PDE	
25	Panteli_4_texte_ML.TextGrid	187.626	188.615	une impasse stupide ¤ il s'est	30	AUX	PRP	
26	Makris_4_texte_ML.TextGrid	111.311	116.103	n jeune membre de l'opposition	20	NOM	AUX	2VO
27	Epifaniou_4_texte_ML.TextGrid	152.211	156.604	e sentiment de se trouver dans	10	PRP	DET	

Tableau_liaison_GE_GR_liaison_M

⇨ Format qui permet de faire des recherches ainsi que des analyses statistiques sur les données

Comparaison inter-codeurs

- Cette fonction permet de comparer les évaluations des différents codeurs
- Possibilité de sélectionner les champs à comparer
- Calcul automatique d'un pourcentage d'accord entre les codeurs
- Possibilité d'écouter les occurrences divergentes, de les ouvrir sous Praat et, si nécessaire, de les corriger directement et de sauver le grid corrigé

↳ **Démonstration!**

Conclusion

- Très rapidement, diffusion au sein des équipes IPFC qui le souhaitent:
  - Des différents codes établis
  - Des interfaces Dolmen IPFC (avec tutoriel)
  - Des références à indiquer pour l'usage de ces outils

↳ L'usage de cette méthode et des outils dédiés permettra de comparer les données entre apprenants de différentes L1 (ex. japonais-espagnol, espagnol-grec chypriote, etc.) et de procéder à des analyses à grande échelle (p. ex. sur la liaison).

Perspectives

Chantiers:

- Automatisation du traitement des données
  - Calcul automatique de pourcentages
  - Etablissement automatique de profils d'apprenants
  - Développement des fonctions de comparaison (inter-populations, inter-tâches, etc.)
- Application à plus large échelle
- Développement du code (structure syllabique, etc.)

Merci de votre attention!

Ce projet a bénéficié et bénéficie du soutien:

- En Suisse:
-du Fonds national suisse de la recherche scientifique (132144/1)
- de la Faculté des Lettres de l'Université de Genève
- de la Société Académique de Genève (Fonds Ch. Bally)

- Au Japon:
- Japanese Society for the Promotion of Science (JSPS) – Grant-in-Aid for Scientific Research B n°23320121
- Waseda University (Special Research Grant, 2011B-297)

- Merci à Laurie Buscail, Yuji Kawaguchi et Françoise Zay
- Un merci particulier aux étudiant-e-s (Cécile Mollet, Tanjema Majeed, Nathalie Bühler, Marie-Laure Sandoz, Marion Didelot, Romain Isely, Marta Osorio et Judith Pérez Santos) pour leur travail minutieux de transcription orthographique et de codage des données!

FNSNF
FONDS NATIONAL SUISSE
DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ DE GENÈVE

JSPS

WASEDA University

Analyse des données IPFC : développement de l'approche par codage

Isabelle Racine[1], Sylvain Detey[2] et Julien Eychenne[3]

[1]ELCF, Université de Genève
[2]SILS, Université Waseda
[3]Hankuk University of Foreign Studies

« Interphonologie du français contemporain : corpus oraux en L2 et évaluation »

Journées IPFC2013 - 9 et 10 décembre 2013 – Maison de Norvège, Cité Internationale, Paris

FNSNF FONDS NATIONAL SUISSE DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ DE GENÈVE

JSPS

WASEDA University

Références

- Boersma, P. & Weenink, D. (2009). *Praat: doing phonetics by computer,* www.praat.org.
- Detey, S. (2012). Coding an L2 phonological corpus: from perceptual assessment to non-native speech models – an illustration with French nasal vowels. In Tono, Y., Kawaguchi, Y. & Minegishi, M. (eds.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam/Philadelphia: John Benjamins, 229-250.
- Detey, S. & Racine, I. (2013). L2 oral corpus data processing at the segmental level: methodological challenges. Workshop "*Cross Cultural Research on Speech Communication & Second Language learning processing*", Université de Bordeaux III, 15 mars 2013.
- Detey, S., Racine, I. & Kawaguchi, Y. (à paraître). Des modèles prescriptifs à la variabilités des performances non-natives : les voyelles nasales des apprenants japonais et espagnols dans le projet IPFC. Dans J. Durand, G. Kristoffersen & B. Laks (éds). *La phonologie du français : des normes aux périphéries* (Festschrift pour Chantal Lyche). Paris : Presses Universitaires de Paris Ouest.
- Detey, S., Racine, I., Kawaguchi, Y., Schwab, S. & Zay, F. (to appear). Variation among non-native speakers: Japanese and Spanish learners of French. In: Detey, S., Durand, J., Laks, B. & Lyche, C. *Varieties of Spoken French: a source book*. With DVD. Oxford: Oxford University Press.
- Durand, J., Laks, B. & Lyche, C. (éds) (2009). *Phonologie, variation et accents du français*, Paris, Hermès.
- Racine, I., F. Zay, S. Detey & Y. Kawaguchi (2011) De la transcription de corpus à l'analyse interphonologique: enjeux méthodologiques en FLE. In G. Col & S.N. Osu, (eds.), *Transcrire, écrire, Formaliser (1).* Rennes: PUR. *Travaux Linguistiques du CerLiCO* 24, 13-30.