

Transcription de corpus d'apprenants multilingues de FLE et analyse interphonologique: enjeux méthodologiques

Isabelle Racine, Sylvain Detey, Françoise Zay & Yuji Kawaguchi

ELCF, Université de Genève,
SILS, Waseda University & LiDiFra, Université de Rouen,
Tokyo University of Foreign Studies

24^{ème} colloque international du CerLiCO

« Transcrire, Écrire, Formaliser »

4-5 juin 2010



UNIVERSITÉ DE GENÈVE



WASEDA University



東京外国語大学 Tokyo University of Foreign Studies



Plan

I. Introduction:

Enjeux et objectifs du projet « Interphonologie du français contemporain » (IPFC)

II. La transcription du corpus:

a) Problèmes et décisions au niveau orthographique

b) Problèmes et décisions au niveau phonétique

III. Conclusion



Le projet IPFC (Detey & Kawaguchi, 2008; Racine et al., à paraître)

- Quelques projets récents basés sur des corpus dans le champ de la phonétique et de la phonologie d'une langue seconde, avec des apprenants de:
 - néerlandais (Neri, Cucchiarini & Strik, 2006)
 - polonais (Cylwik, Wagner & Demenko, 2009)
 - allemand (Gut, 2009)
 - anglais (Gut, 2009; Visceglia, Tseng, Kondo, Meng & Sagisaka, 2009)
- ⇒ Avec ciblage sur des aspects tant segmentaux que suprasegmentaux (Trouvain & Gut, 2007 ; Meng, Tseng, Kondo, Harrison & Visceglia, 2009)
- Mais, très peu de données de ce type pour le français (voir p. ex. Delais-Roussarie & Yoo, 2010; Pillot-Loiseau, Amelot & Fredet, 2010)

 **Projet « Interphonologie du français contemporain »**

www.projet-pfc.net



Le projet IPFC (Detey & Kawaguchi, 2008; Racine et al., à paraître)

- 1^{er} objectif: constituer un large corpus de données orales issues de locuteurs de différentes L1s en utilisant un protocole et des outils identiques.
- Ce protocole a été adapté à partir de celui du projet PFC (« Phonologie du français contemporain », Durand, Laks & Lyche 2002, 2005, 2009, <http://www.projet-pfc.net>) et inclut 5 tâches différentes :
 - Répétition d'une liste spécifique de mots lue par un natif
 - Lecture de 2 listes de mots (liste PFC + liste spécifique)
 - Lecture du texte PFC
 - Entretien guidé
 - Interaction entre deux apprenants
- 2^{ème} objectif: examiner des questions méthodologiques au niveau de l'articulation entre linguistique de corpus et méthodologie utilisée par la psycholinguistique.



Le projet IPFC (Detey & Kawaguchi, 2008; Racine et al., à paraître)

- Les 2 premiers corpus du projet IPFC:
 - IPFC – Japon:
 - Une centaine d'apprenants de français à TUFs (japonophones de Tokyo) – 3 tâches (listes de mots et texte)
 - Enregistrement des conversations en cours
 - Niveaux CECR variés
 - IPFC – Espagne
 - 14 apprenants de français à Genève (hispanophones d'Espagne)
 - 12 apprenants de français à Madrid (hispanophones d'Espagne)
 - Niveau B2-C1 du CECR.
 - Protocole complet
- ⇒ Centrage commun sur les voyelles nasales (Detey et al., à paraître; Racine et al., 2010)
- ⇒ Centrage spécifique pour IPFC-Espagne sur les occlusives sonores (Racine et al., à paraître)



**Quelle(s) transcription(s) pour le corpus
IPFC?**



Les enjeux de la transcription

- Selon Delais-Roussarie (2009):
«Transcrire des données sonores consiste à fournir une représentation symbolique du signal. Cette représentation n'est pas équivalente au signal, dans la mesure où elle est le résultat d'une analyse, ou plutôt d'une abstraction, des données réelles».
 - Elle distingue 4 grandes difficultés liées à cette analyse du signal:
 - 1) Les problèmes liés à la difficulté d'écoute
 - 2) Les difficultés résultant de la reconstruction perceptive effectuée obligatoirement par le transcripneur
 - 3) Le poids des préjugés linguistiques (ex. «ne» de négation)
 - 4) Les ambiguïtés liées au code oral (ex. «il mange» / «ils mangent»)
- ⇒ Nécessite l'établissement de conventions de transcriptions



Les données d'apprenants

- La difficulté liée à la **reconstruction perceptive** effectuée par le transcripteur devient prépondérante lorsqu'il s'agit de transcrire des données d'apprenants.
- Le taux de désaccord entre transcripteurs augmente considérablement lorsqu'il s'agit de données non natives (10 à 34%) vs données natives (5%) (Zechner, 2009).
 - ⇒ Impossible d'éviter un certain degré d'interprétation!
- 2 impératifs:
 - Nécessité de lisibilité des données ⇒ besoin d'un niveau orthographique simple
 - Nécessité de fidélité ⇒ dans le domaine de la phonologie, besoin d'un système permettant de rendre compte des réalisations des apprenants
- Questions:
 - Comment rendre compte des formes « déviantes »?
 - API suffisant?

Transcription orthographique

1) Lecture vs spontané :

morphologie encodée vs morphologie à décoder

[le]

- «**le** maire» : tâche de lecture, le locuteur encode un singulier. Cible phonologique = /l $\text{\textcircled{e}}$ /
- «c'était **le\les** choix» : conversation, le locuteur généralise [e] à la place de [$\text{\textcircled{e}}$], on se sait pas quelle est la forme visée, l'auditeur peut décoder soit un singulier soit un pluriel. 2 cibles phonologiques possibles : /l $\text{\textcircled{e}}$ / - /le/

Le contexte ne désambiguïse pas :

Il y avait le français et l'anglais c'était le/les choix on pouvait choisir

Transcription orthographique

- « le premier Ministre » : en lecture, pas d'hésitation sur le fait qu'il s'agit d'un masculin avec dans la réalisation une mauvaise correspondance graphie-phonie.
- La transcription orthographique des conversations doit-elle refléter la morphologie du transcripteur ou celle du locuteur?
 - « [ilaBevenu] » : - généralisation de l'auxiliaire « avoir »?
- « être » avec réalisation déviante?
- « je l'ai revue et je lui ai donné les (transcriptions) »
 - « je l'a revue et je l'ai donné ... »
 - « je la revue et je l'ai donné ... »
- « les gens (...) ils habitent (...)
ils ne connaient pas \ il ne connaît pas

Transcription orthographique

2) Mots existants, mots inexistantes et emprunts à la L1

Dans l'interlangue, pas facile de déterminer quelle est la « cible » :

[lɛnɔ̃bʁɛdɛnɔ̃nbudist]  

a) « le nom de nonne bouddhiste c'est [...] Min Tchi »

b) « le nombre (= nom) de nonne bouddhiste... »

⇒ La transcription orthographique b) reflète l'encodage du locuteur, pas le problème d'interprétation du transcripteur.

⇒ Distinction à faire dans la transcription phonétique entre **code-switch** (retour à la L1) et **emprunt** (forme lexicale de la L1 mais adaptation phonétique à la L2)



Transcription orthographique : conclusion

- L'exigence de lisibilité ainsi que la compréhension globale du message passent obligatoirement par une forme d'interprétation des données orales.
- Cette interprétation est un donné dans les tâches de lecture. En spontané, elle est le plus souvent (re)construite par le contexte.
- Dans la plupart des cas, une transcription orthographique peut être fournie, ce qui permet d'éviter de:
 - multiplier les segments inaudibles
 - mêler API et orthographe standard

mais ce qui oblige à:

- indiquer les interprétations multiples ou peu fiables
- décider si c'est le plan morphologique ou phonologique qui est en jeu

Transcription phonétique fine

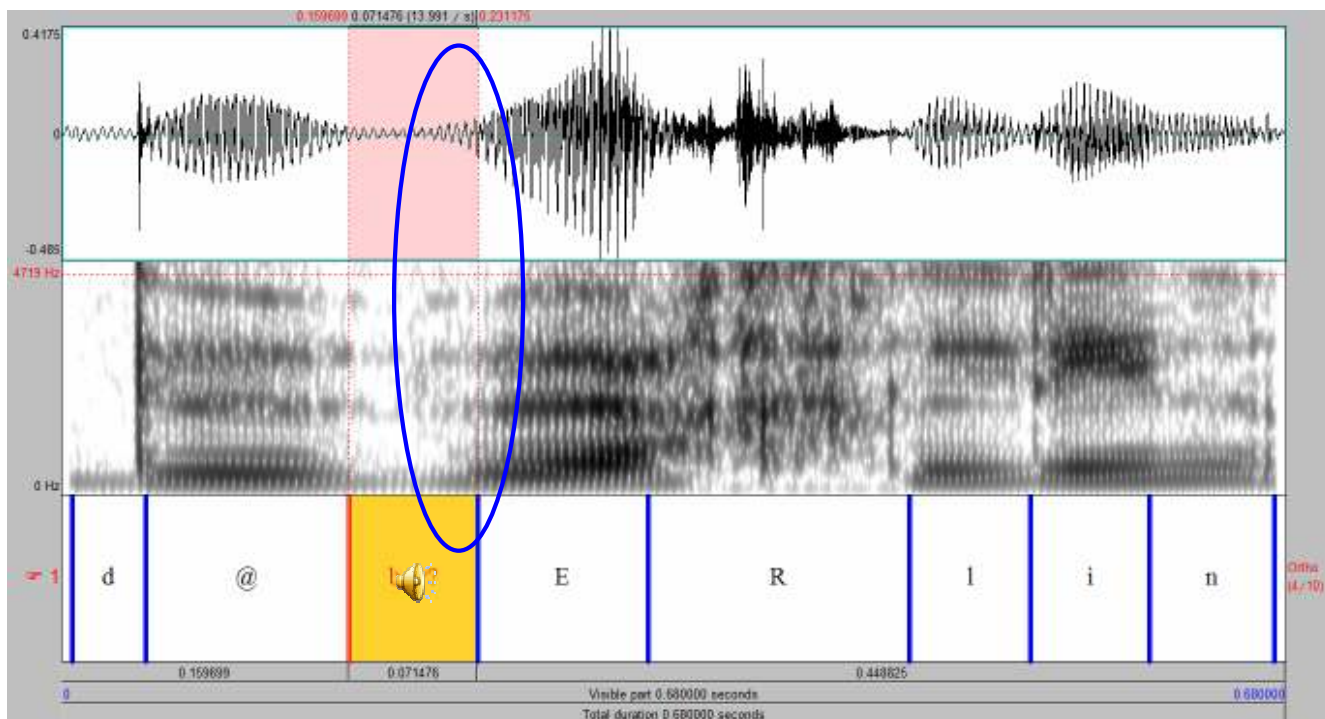
1)   ⇒ quels mots, quelle orthographe?

⇒ problème d'identification de la cible, mais une fois la cible identifiée, pas de problème de transcription orthographique («dorer» et «l'arabe», ni de notation avec l'API : [dɔʒe], [agab]).

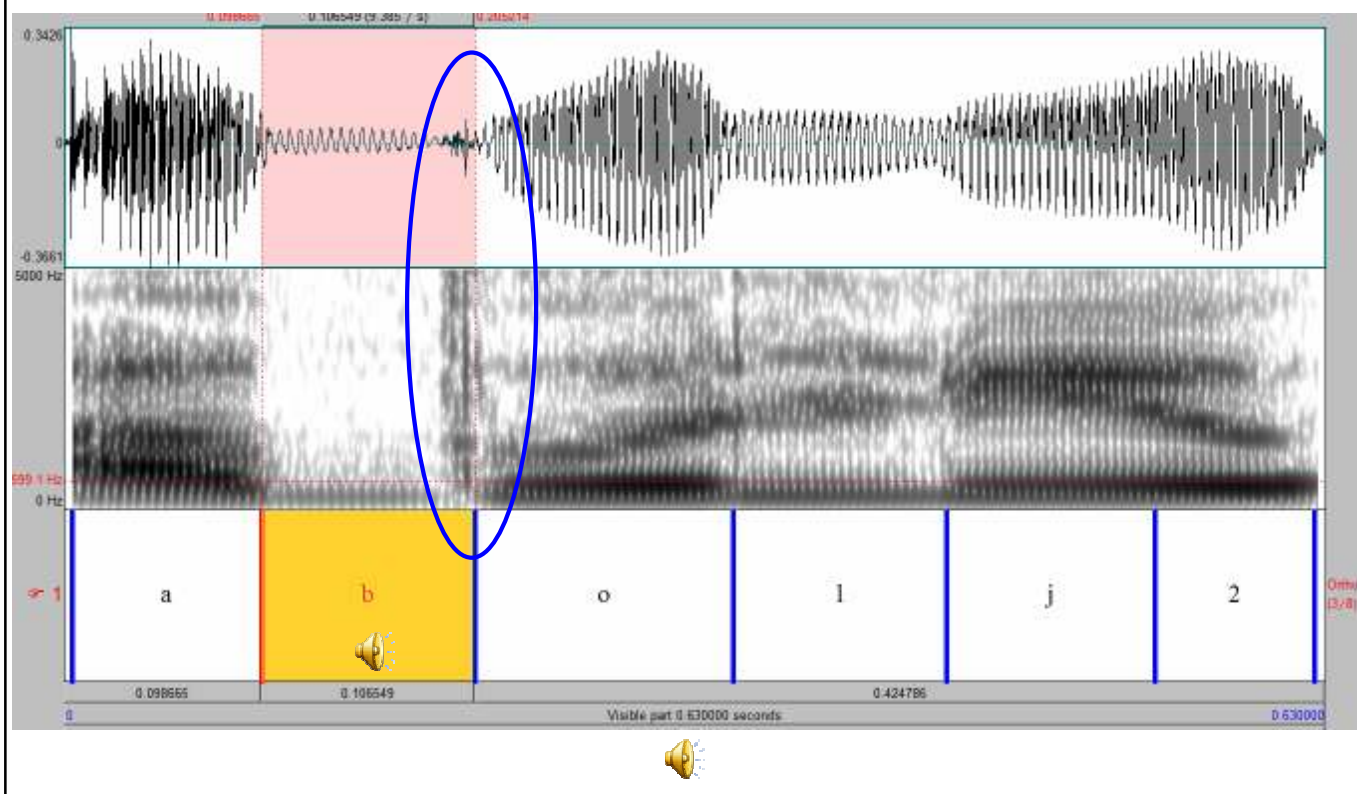
2)  « 4ème aux Jeux Olympiques de Berlin en 1936 »

⇒ quelle transcription phonétique?



Transcription phonétique fine



Transcription phonétique fine



Transcription phonétique fine

1)   ⇒ quels mots, quelle orthographe?

⇒ problème d'identification de la cible, mais une fois la cible identifiée, pas de problème de transcription orthographique («dorer» et «l'arabe», ni de notation avec l'API : [dɔʒe], [agab]).


2)  «4^{ème} aux Jeux Olympiques de **B**erlin en 1936»

⇒ problème d'identification d'une différence phonétique fine mais que l'on peut transcrire en ayant recours à l'API de la L1 ([β]) après avoir vérifié le signal sonore (présence ou non d'une barre de plosion).

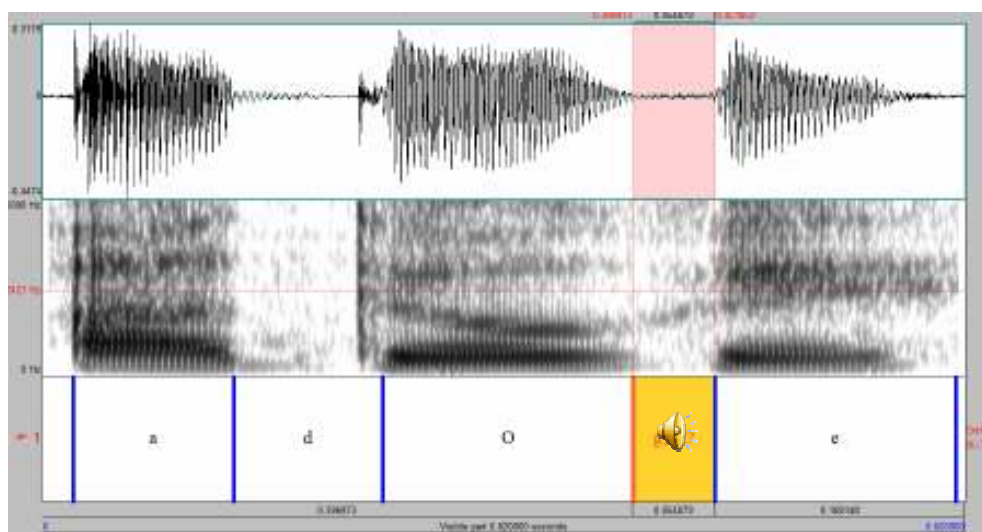
Autres exemples du même type avec [g, γ]



Transcription phonétique fine

3)  ⇒ Quel mot, quelle orthographe, quelle transcription phonétique?

⇒ différentes interprétations phonétiques possibles selon le transcripteur [g, γ, ʁ] et selon le signal [γ, ʁ], mais codage possible dans l'API.





Transcription phonétique fine

4)



⇒ Quelle voyelle nasale?



- ⇒ Variation selon si l'on dispose du contexte ou non.
- ⇒ Pas de recours au signal possible dans le cas des voyelles nasales.
- ⇒ Si la transcription phonétique se fait sur la base de la transcription orthographique (= avec contexte), risque de biais et de discordance entre transcrip-teurs!

Décision:

Ne pas « transcrire phonétiquement » mais se centrer sur un petit nombre de phénomènes, les examiner de manière très détaillée, avec recours au signal sonore lorsque c'est possible, et décrire les réalisations par le biais d'un système de codage basé sur un paramétrage de certains traits (nasalité de la voyelle, présence d'un appendice consonantique, etc.)



Transcription phonétique fine: conclusion

- L'usage de l'API implique un degré de catégorisation qui accorde beaucoup de poids à l'oreille native ou bilingue du transcripteur, avec tous les problèmes de fiabilité et de discordance que cela induit.
- Adopter un système de codage d'un certain nombre de paramètres permet d'éviter d'attribuer une catégorie phonémique à une réalisation dont l'appartenance à une catégorie phonémique donnée est problématique.
- Ce codage est coûteux donc implique de se focaliser sur un nombre limité de phénomènes dans l'analyse de l'interphonologie des apprenants. Pour IPFC-Espagne:
 - Les voyelles nasales
 - Les occlusives sonores
- Ce codage nous permettra d'aboutir à une image de l'interphonologie en comparant les codages à la cible phonologique établie sur la base de la transcription orthographique.
⇒ on évite ainsi les problèmes de catégorisation perceptive d'une réalisation qui n'appartient pas au système phonologique de la L2.



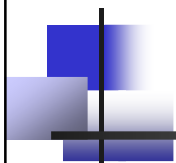
Transcription phonétique fine: conclusion

- La perception (= catégorisation phonémique) par des natifs se fait dans une étape ultérieure par le biais de tests perceptifs effectués par des sujets non spécialistes.
 - Ces tests perceptifs comprennent 2 étapes (*cf.* Strange et al., 2005):
 - l'identification du phonème ciblé.
 - Un indice du degré de représentativité de l'élément perçu comme membre d'une catégorie donnée.
- ⇒ Pour un exemple de ce type de travail sur les voyelles nasales, voir Racine et al. (2010).



Conclusion

- Cette phase d'analyse préliminaire et de prise de décisions nous semble une étape indispensable pour assurer la qualité des analyses qui seront ensuite effectuées sur la base du corpus.



Merci !



Références

- Cylwik, N., Wagner, A., Demenko, G. 2009. The EURONOUNCE corpus of non-native Polish for ASR-based Pronunciation Tutoring System. *Proceedings of SlaTE 2009 – 2009 ISCA Workshop on Speech and Language Technology in Education*. Birmingham, UK.
- Delais-Roussarie, E. (2009). *Conventions CHAT de Transcription des données*. Document interne, BDD Interlangue, janvier 2009.
- Delais-Roussarie, E. & Yoo, H. (2010). The COREIL corpus: a learner corpus designed for studying phrasal phonology and intonation. *Proceedings of New Sounds 2010*, 3-5 May 2010, Poznan.
- Detey, S. et Kawaguchi, Y. (2008). Interphonologie du Français Contemporain (IPFC) : récolte automatisée des données et apprenants japonais. *Journées PFC : Phonologie du français contemporain : variation, interfaces, cognition*, Paris, 11-13 décembre 2008.
- Detey, S., Racine, I., Kawaguchi, Y., Zay, F., Bühler, N., Schwab, S. (à paraître). Evaluation des voyelles nasales en français L2 en production: de la nécessité d'un corpus multitâches. *Actes de CMLF 2010*, Nouvelle Orléans, 12-15 juillet 2010.
- Durand, J., Laks, B. & Lyche, C. (2002). La phonologie du français contemporain: usages, variétés et structure. In : C. Pusch & W. Raible (eds.), *Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language*. Tübingen : Gunter Narr Verlag, 93-106.
- Durand, J., Laks, B. & Lyche, C. (2005). Un corpus numérisé pour la phonologie du français. In G. Williams (ed.), *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes, p. 205-217. Actes du colloque 'La linguistique de corpus', Lorient, 12-14 septembre 2002.
- Durand, J., Laks, B., Lyche, C. 2009. Le projet PFC: une source de données primaires structurées. In: Durand, J., Laks, B., Lyche, C. (eds), *Phonologie, variation et accents du français*. Paris: Hermès. 19-61.
- Gut, U. (2009). *Non-native Speech: A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. Wien: Peter Lang.



Références

- Meng, , Tseng, , Kondo, , Harrison, & Visceglia, (2009). Studying L2 suprasegmental features in Asian Englishes: a position paper. *Proceedings of Interspeech 2009*, Brighton, R-U.
- Neri, A., Cucchiaroni, C. & Strik, H. (2006). Selecting segmental errors in non-native Dutch for optimal pronunciation training.. *IRAL - International Review of Applied Linguistics in Language Teaching*, 44, 357-404.
- Pillot-Loiseau, C., Amelot, A. & Fredet, F. (2010). Contributions of experimental phonetics to the didactics of the pronunciation of the French as a Foreign language: stage 1: reflection around the establishment of a speaking materials. *Proceedings of New Sounds 2010*, 3-5 May 2010, Poznan.
- Racine, I., Detey, S., Zay, F., Y. Kawaguchi (à paraître). Des atouts d'un corpus multitâches pour l'étude de la phonologie en L2: l'exemple du projet « Interphonologie du français contemporain » (IPFC). In: Kamber, A., Skupiens, C. (eds). *Recherches récentes en FLE*. Berne: Peter Lang.
- Racine, I., Detey, S., Bühler, N., Schwab, S., Zay, F. & Kawaguchi, Y. (2010). The production of French nasal vowels by advanced Japanese and Spanish learners of French: a corpus-based evaluation study. *Proceedings of New Sounds 2010*, 3-5 May 2010, Poznan.
- Strange, W., Bohn, O.-S., Trent, S. A. & Nishi, K. (2005). Contextual variation in the acoustic and perceptual similarity of North German and American English vowels. *Journal of the Acoustical Society of America*, 118 : 1751-1762.
- Trouvain, J. & Gut, U. (eds) (2007). *Non-Native Prosody. Phonetic Description and Teaching Practice*. Berlin/New York: Mouton de Gruyter.
- Visceglia, Tseng, Kondo, Meng & Sagisaka (2009). Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project). *Proceedings of Oriental-COCOSDA*, Urumuqi, Chine.
- Zechner, K. (2009). What did they actually say? Agreement and Disagreement among Transcribers of Non-Native Spontaneous Speech Responses in an English Proficiency Test. *Proceedings of the ISCA SLaTE-2009 Workshop*, Wroxall, UK, September.