Enjeux méthodologiques dans les corpus oraux en L2: transcriptions, annotations et codages

Isabelle Racine[1], Sylvain Detey[2], Françoise Zay[1] & Yuji Kawaguchi[3]

[1]ELCF, Université de Genève,
[2]SILS, Waseda University & LiDiFra, Université de Rouen,
[3]Tokyo University of Foreign Studies

IPFC 2010 « Interphonologie, corpus et français langue étrangère »

Paris – 8 décembre 2010

UNIVERSITÉ DE GENÈVE WASEDA University UNIVERSITE DE ROUEN 東京外国語大学 Tokyo University of Foreign Studies

Plan

I. Introduction:

Transcrire des données orales non natives : enjeux

II. Transcriptions et annotations dans le projet IPFC

a) Conventions générales

b) Niveau orthographique

c) Niveau phonético-phonologique

III. Codages segmentaux

IV. Conclusion

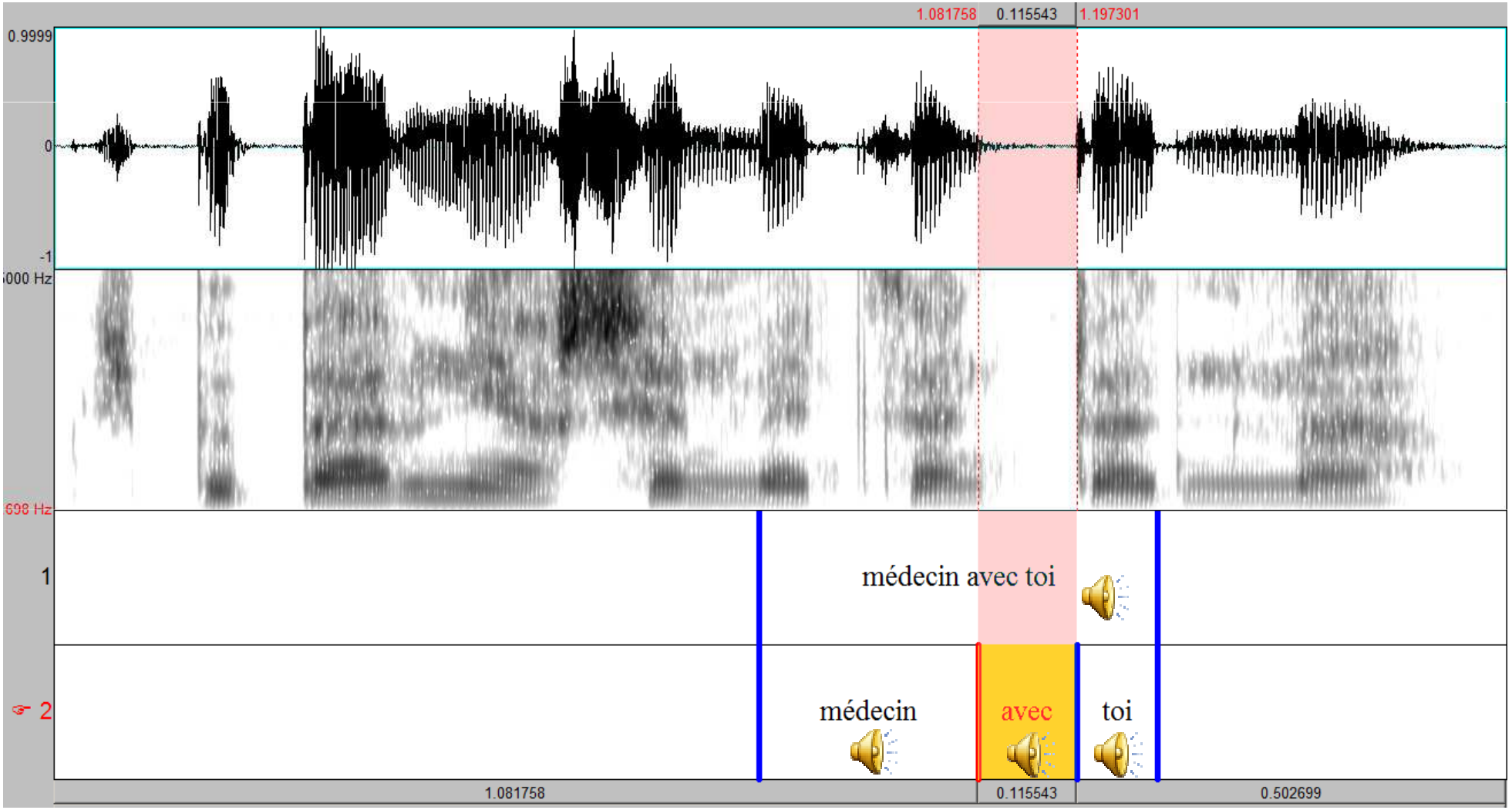
1. Transcrire des données orales non natives : enjeux

- Transcrire, pourquoi ?
  - accessibilité des données
    - La transcription = point d'entrée dans le corpus
    - conséquence = exigence de lisibilité
  - questionnement méthodologique
    - transcrire = participer à la construction des faits analysés, mesurer la distance entre ce qui est effectivement produit et les formes langagières interprétées.
- Transcrire, comment?
  - « *fournir une représentation symbolique du signal* »

    → abstraction du réel
  - « *transcrire ce qui est dit* », « *éviter au maximum de faire des interprétations* » (Delais-Roussarie 2009)

    → fidélité au réel
  - Une grande difficulté : la reconstruction perceptive effectuée par le transcripteur.

La reconstruction perceptive



Reconstruction *vs* fidélité?

- 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
  - ⇨ Impossible parfois d'arriver à une interprétation univoque
- 2 impératifs:
  - Nécessité de lisibilité des données ⇨ besoin d'un niveau orthographique simple (qui sélectionne des formes morphologiques abstraites)
  - 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?

II. Transcriptions et annotations dans le projet IPFC

II.1. Conventions générales dans IPFC

- Transcriptions effectuées sous Praat (Boersma & Weenink, 2009).
- Conventions largement inspirées de celles de PFC.
- Rendu minimum: transcription orthographique du texte et des deux conversations alignées en respectant les conventions générales suivantes:
  - Vérification de toutes les transcriptions par un 2ème transcripteur et discussions en cas de désaccord
  - Transcription orthographique sans ponctuation (cf. p. ex. *Buckeye Speech Corpus*)
  - Frontière dans le grid quand il y a une pause longue (idem PFC)
  - Transcrire des formes complètes

    Ex. « il y a » et non « y'a »
  - Si homophonie complète (≠ forme déviante produite par un apprenant), utiliser les parenthèses

    Ex. « on (n')a pas »
  - Pas de caractères phonétiques dans la tire orthographique car tire spécifique réservé aux phénomènes phonético-phonologiques

II.1. Conventions générales dans IPFC

- Conventions plus « standard »:
  - Pour les chevauchements ⇨ chevrons

    Ex. XG: xxxxxx <E: xxx> XG: xxx
  - Transcrire les chiffres en toutes lettres
  - Sigles ⇨ ONU mais U.E.F.A.
  - Syllabes incompréhensibles ⇨ X (un X = une syllabe)
  - Bruit, rire ou autres commentaires ⇨ [toux]
  - Troncation ⇨ sy- syllabe
  - Acquiescement ⇨ hum hum
  - Onomatopées ⇨ voir liste (à compléter au fur et à mesure)
  - Hésitation: euh (standard), èh, mh (liste à compléter)

Niveau orthographique

1) Formes inintelligibles : « X »

2) Morphologie correcte – réalisation phonétique déviante

- «**le** hasard» : tâche de lecture, le locuteur encode un article singulier. Cible phonologique = /lə/, réalisation phonétique = [də]
- « le **premier** Ministre » : également en lecture, mauvaise correspondance graphie-phonie. Cible = « premier », réalisation phonétique = [pʀœmjɛʀ]

Niveau orthographique

3) Morphologie erronée mais forme existante – bonne réalisation phonétique

- « [ilaʁevenu] » : - « est » avec réalisation déviante ? Non, « il a revenu » → généralisation de l'auxiliaire « avoir »

4) Morphologie erronée et forme inexistante – adaptation orthographique

- « les gens (…) ils habitent (…)
  ils ne **connaient** pas / il ne **connaît** pas

Niveau orthographique

5) Ambiguïtés – multitranscriptions

- «c'était **le/les** choix» : le locuteur généralise [e] à la place de [ə], 2 cibles phonologiques possibles : /lə/ - /le/

  Le contexte ne désambiguïse pas :

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

- «une partie **de/le/du** Moyen Age» : il y a un [d] à l'initiale du déterminant. « du » est la forme attendue, « de » est une forme probable, « le » est possible vu le contexte large…

  → Quel niveau impliquer? La morphologie ou la phonétique?

Niveau orthographique

6) Emprunts et alternances codiques

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

[lɛnɔ̃bʁɛdɛnɔnbudist]  

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 entre **alternance codique** (retour à la L1) et **emprunt** (forme lexicale de la L1 mais adaptation phonétique à la L2)

Niveau 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

II.3. Niveau phonético-phonologique

- Transcription phonétique ⇨ question peu traitée jusqu'à récemment avec l'apparition de corpus intégrant ce type de transcriptions (cf. Neri et al., 2006; Cylwik et al., 2009; Gut, 2009; Visceglia et al., 2009; Delais-Roussarie & Yoo, 2010; Pillot-Loiseau et al. 2010).
- Unanimité: travail coûteux en temps, fastidieux et qui requiert des compétences spécifiques (cf. Gut, 2009; Makino, 2007; Wester et al., 2001).
- Problème principal: **subjectivité** due à la reconstruction perceptive (« filtre phonologique »).
- Quelques exemples de difficultés:

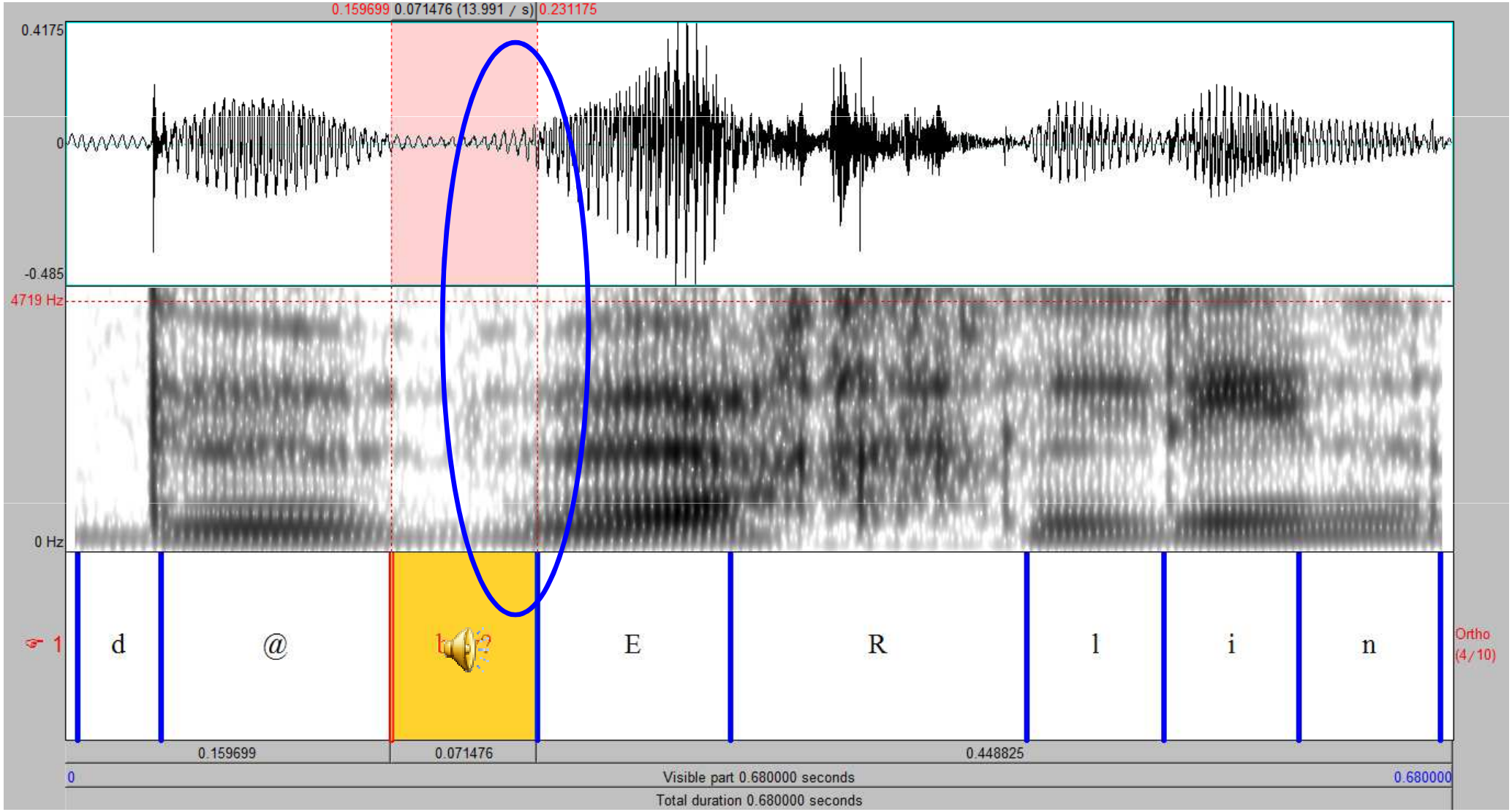
1) Identification des formes « déviantes »

« 4ème aux Jeux Olympiques de Berlin en 1936 »

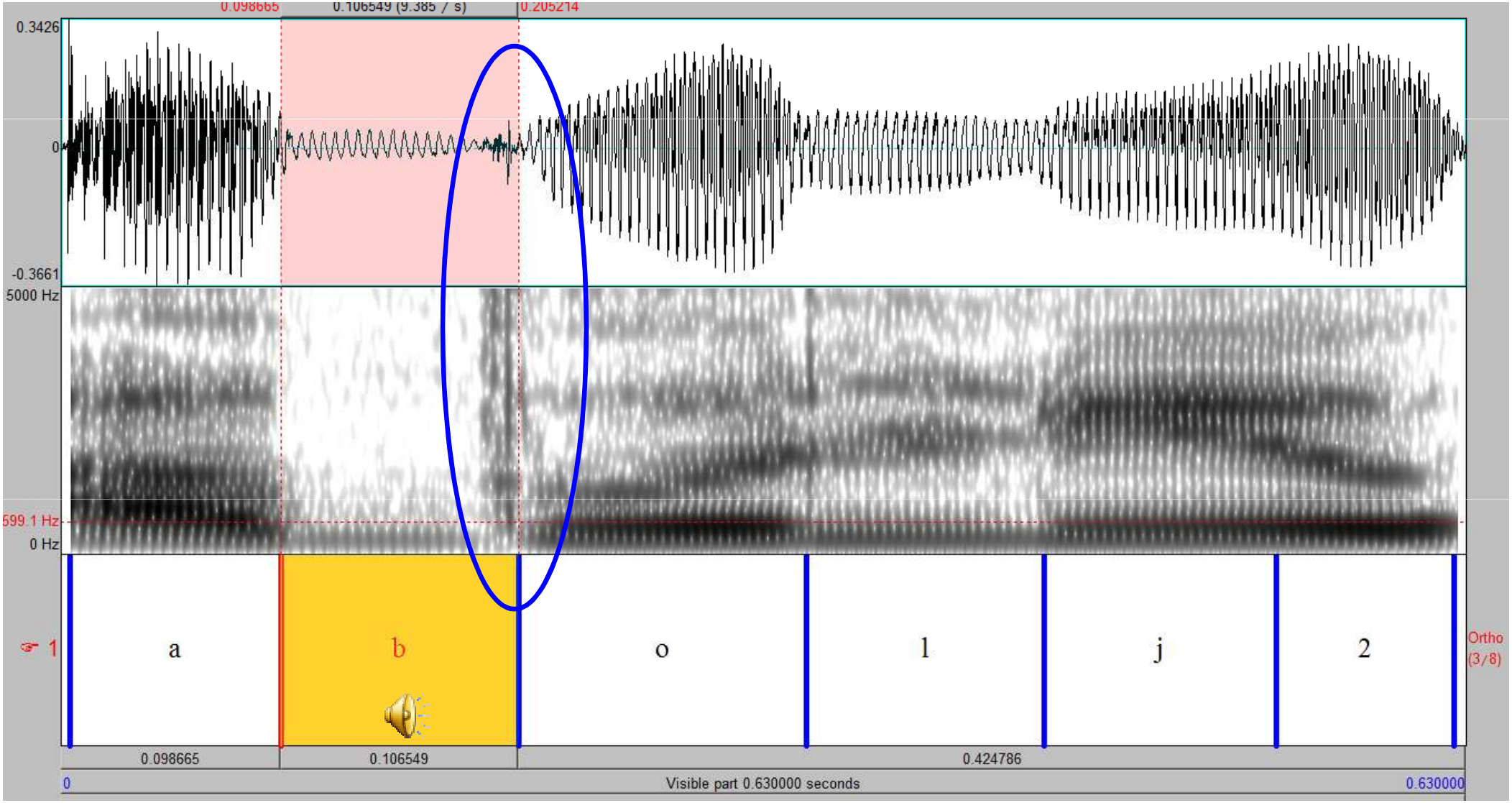
⇨ qu'est-ce qui doit être considéré comme « déviant »?

⇨ quelle transcription phonétique?

II.3. Niveau phonético-phonologique



II.3. Niveau phonético-phonologique



II.3. Niveau phonético-phonologique

1) Identification des formes « déviantes »

«4ème aux Jeux Olympiques de **B**erlin en 1936»

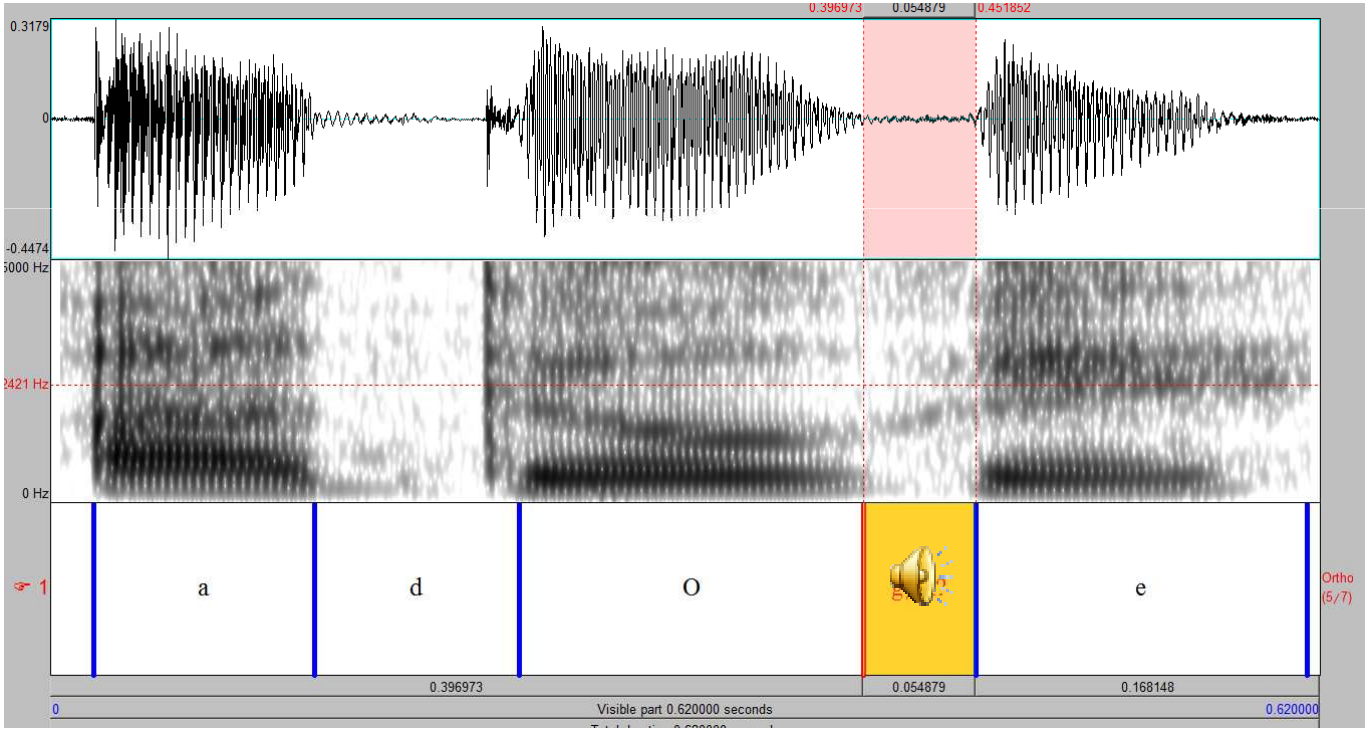
⇨ problème d'identification d'une différence phonétique fine liée à la reconstruction perceptive effectuée par le transcripteur mais que, une fois identifiée, on peut transcrire à priori en ayant recours à l'API de la L1 ([β]) après avoir vérifié le signal sonore (présence ou non d'une barre d'explosion).

II.3. Niveau phonético-phonologique

2) Interprétations multiples

Quelle transcription phonétique?

⇨ différentes interprétations phonétiques possibles selon le transcripteur [g, ɣ, ʁ] et selon le signal [ɣ, ʁ], mais codage possible dans l'API ⇨ multitranscription



II.3. Niveau phonético-phonologique

3) Catégorisation phonétique problématique

- ⇨ Quelle voyelle nasale?
- ⇨ Pas de recours au signal possible dans le cas des voyelles nasales.
- ⇨ Aide par le contexte (cible phonologique identifiable).

⇨ Comment transcrire cette réalisation?

⇨ Limites de l'API?

Autres exemples: « ma**g**asin »

II.3. Niveau phonético-phonologique: conclusion

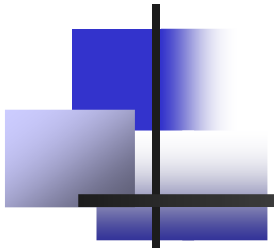
- La subjectivité induite par la double catégorisation effectuée par:
  - « l'oreille » du/des transcripteur(s)
  - la représentation catégorielle qui découle de l'usage d'un alphabet phonétique.

  pose problème.

- Il semble donc nécessaire:
  - de s'appuyer sur la visualisation du signal sonore
  - de recourir à plusieurs transcripteurs
  - que ceux-ci connaissent les deux systèmes en jeu
  - de tester la perception (= catégorisation phonémique) dans une étape ultérieure par le biais de tests perceptifs effectués par des sujets natifs non spécialistes (cf. Strange et al. 2009; Racine et al., 2010).

↳ **Tâche de transcription phonétique extrêmement lourde!**

⇨ adopter plutôt un système de **codage** de certains phénomènes ciblés.



III. Codage: réflexions préalables

III.1. Coder pour quoi?

Une réponse possible: (pour un phénomène donné)
un panorama automatisé des réalisations phonétiques intégrant des facteurs contextuels (environnement *au moins* phonologique…)

- « Panorama »:
  - réalisations conformes (à la cible en L1)

  et
  - réalisations déviantes
- Interprétation de ce panorama phonétique grâce aux facteurs contextuels pris en compte :

  → vue d'ensemble du système phonologique du sujet

Malgré: système INTERphonologique (Interlangue (IL)-provisoire et instable)

Evolutif → Nécessité d'études longitudinales

Variable →Nécessité de (macro-)corpus - système émergent

III.2. Coder comment?

- Avec un code adapté
  (quelle fonction assignée au codage?)

  (codage ≠ transcription bis, en particulier en L2)
- Pour coder les réalisations phonétiques: principe de PFC
  ne pas pré-interpréter, rester aussi descriptif que possible
  (p. ex: la liaison, pas de classification a priori (ob/fac/int) de manière à faire émerger des catégories robustes (cat/var/n.a.))
- Mais le codage repose en partie sur la connaissance des phénomènes et des facteurs impliqués
  (cf. codages PFC schwa et liaison).
- Dans le cas de l'IL, il faut connaitre:
  - La L1 des sujets
    (p. ex. ne pas confondre [ə] et [ɯ] chez les japonophones)
  - Certaines caractéristiques du processus d'apprentissage
    (p. ex. hypercorrection dans la production des liquides chez les japonophones ou antériorisation des voyelles arrondies postérieures chez les hispanophones)

Connaissances sur la L1 et l'IL requises pour:

- L'élaboration du code
- L'activité de codage

→ Lors du codage: problème de compétence métalinguistique et de reconstruction perceptive du codeur (comme pour le transcripteur):

→ sensibilité (phonético-phonologique) du codeur

- p. ex: longueur vocalique

→ norme du codeur

- p. ex.: brun/brin

Et ne pas négliger la tâche (impact orthographique)!

- p. ex.: multiplication des consonnes en lecture

III.3. Travaux antérieurs?

- Deux types:
  - *Evaluation* (« rating »/ « assessment »)
    - Echelle scalaire: 0-5 / un peu-beaucoup
  - *Catégorisation*
    - Substitution/ Effacement / Insertion

Exemple: Bent Bradlow & Smith (2007), Yoon et al. (2009)

- Problèmes:
  - Evaluation: quel facteur précisément (degré d'accent (accentedness), intelligibilité (intelligibility), bonne formation (goodness), précision (accuracy), absence d'accent (nativeness), prototypicalité, etc.).
  - Catégorisation: pas assez informatif & problèmes identiques à l'évaluation (seuils des substitutions, etc.)

III.4. Dans PFC ?

- Adaptation du codage:
  - cas d'effacement ou de réduction consonantique ayant une incidence sur le schwa en français ivoirien (Boutin 2006)

- Etudes en L2: traitement du schwa
  - chez les apprenants marocains (Grüter 2008)
  - chez les apprenants néerlandais (Nouveau & Berns 2008)

- Codage segmental autre que schwa ou liaison:
  - Opposition entre voyelle haute et glide en français canadien (Poiré et Gurski 2004)
  - Alternance en position post-vocalique entre /r/ réalisé et /r/ non-réalisé en français de la Réunion (Bordal 2006)
  - Opposition entre /r/ antérieur et postérieur en français canadien (Poiré, Moroz et Kelly 2008)

En résumé:

- Codage binaire:
  - bon (inclusion) /mauvais (exclusion)
  - absent/présent
  - antérieur/postérieur
- Et gradience…
- Dans tous les cas: impossible de tout coder *a priori*.
- Autre contrainte: développement du codage ET de l'outil d'analyse
  - Plateforme PFC ? PHON? Outil ad hoc?
  - Si plateforme PFC: mêmes facteurs que pour le schwa?
  - Quels facteurs intégrer dans le codage?
  - Impossibilité d'intégrer tous les facteurs…

→ plusieurs codages?

- Selon l'objectif de l'étude?
- Selon la L1?
- Codage générique et codage spécifique?

Exemple: le cas des nasales

Plusieurs options de codage possibles (+/– fin):

- Substitution
- Substitution en voyelle nasale/nasalisée
- Substitution en voyelle nasale X du français
- Substitution en voyelle nasale/nasalisée Y (non français)
- Substitution en voyelle orale X du français (= effacement du trait de nasalité?)
- Substitution en voyelle orale Y (non français)
- Substitution en voyelle orale (laquelle?) + consonne nasale (laquelle?)
  - (= substitution + insertion ? = effacement + insertion?)
- Substitution en diphtongue (avec ou sans nasalisation) (= insertion?)
- Etc.

Domaine français/non français: problèmes de délimitation des variétés (ex: voyelles relâchées en français canadien pour des apprenants anglophones canadiens…).

Le cas de IN / UN est plus complexe :

- Que faire quand UN est réalisé IN ?
- Que faire quand IN est réalisé UN ?
- Que faire en cas de réalisations intermédiaires ?
- Que faire des incidences lexicales (« un », « brun », etc…)

Exemple de proposition provisoire

- 1er chiffre: contexte phonologique segmental de la voyelle nasale (1 = V ; 2 = VC ; 3 = CVC, 4 = CV ; 5 = CCVC, etc.)
- 2ème chiffre: qualité de la nasalité (1 = voyelle nasale ; 2 = voyelle orale avec nasalisation partielle ; 3 = voyelle orale)
- 3ème chiffre: qualité du timbre (1 = timbre correct ; 2 = timbre incertain ; 3 = timbre erroné)
- 4ème chiffre: présence/absence d'une consonne nasale postvocalique (1 = pas de consonne ; 2 = consonne ou appendice)

Exemple: « j'ai trente-trois ans » 

- « trente »: CCVC,

  Timbre et nasalité : pas de problème et pas de consonne postvocalique

  Codage: « 5111 ».

- « ans »: V 

  Voyelle nasale et non orale, mais timbre non catégorisable en français avec appendice consonantique postvocalique non identifiable.

  Codage: « 1122 ».

Perspectives

Codage encore perfectible…

Et encore des questions à envisager…

- Par exemple:
  - Domaine de codage: phonème ? trait ? → complexité du trait (phonétique/phonologique et architecturé) en particulier en IL… à réserver pour des études particulières?
  - Codage du degré de transparence L1/L2 ?

    → p. ex: *Berlin* vs *Enghien*

    (pb. de fréquence et de disponibilité lexicale a minima)
  - En lecture:
    - Codage des graphèmes?
    - nombre et nature des digrammes (*pin* vs *pain; en* vs an)?

    (en lecture (et répétition): listes et textes peuvent être pré-codés)

Donc: codage : *chantier en cours!*

Conclusion

- Le traitement d'un corpus L2 à des fins d'analyses phonético-phonologiques exige un gros travail de réflexion à différents niveaux:
  - Transcription orthographique
  - Phonético-phonologique ⇨ développement d'un système de codage approprié
  - Outils à utiliser

mais

- 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

- The Buckeye Speech Corpus, http://vic.psy.ohio-state.edu/.

- Bent, T., Bradlow, A. R. and Smith, B. L. (2007). Segmental errors in different word positions and their effects on intelligibility of non-native speech: All's well that begins well. In Munro, M. and Bohn, O.-S. (eds.), *Language Experience in Second Language Speech Learning: In honor of James Emil Flege*. Amsterdam: John Benjamins, 331-347.
- Bordal, G. (2006). *Trace de la créolisation dans un français régional : le cas du /r/ à l'île de la Réunion*. Mémoire de Master. Université d'Oslo (Norvège).
- Boutin, B. (2006). PFC-Abidjan : choix méthodologiques liés à l'extension d'un corpus. In Rastier, F. & Ballabriga, M. (dir.) Actes du colloque international d'Albi, juillet 2006 : *Documents numériques et interprétation : corpus en lettres et sciences sociales*, 288-292. Publiés par C. Duteil & B. Foulquié, *Revue Texto*, Presses de l'Université de Toulouse Le Mirail, http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Boutin.pdf.
- 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.
- Gut, U. (2009). Non-native Speech: *A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. Wien: Peter Lang.

Références

- Grüter, M. (2008). Schwa et structure syllabique: variation dans le français des apprenants marocains. Colloque *Phonologie du français contemporain : variation, interfaces, cognition*. MSH, Paris.
- Makino, Takehiko, 2007, « A corpus of Japanese speakers' pronunciation of American English : preliminary research », *PTLC 2007*, UCL, London.
- Meng, , Tseng, , Kondo, , Harrison, & Viscelgia, (2009). Studying L2 suprasegmental features in Asian Englishes: a position paper. *Proceedings of Interspeech 2009*, Brighton, R-U.
- Neri, A., Cucchiarini, 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.
- Nouveau, D. (2008). *Schwa et apprenants néerlandophones : Impact de l'effacement sur la reconnaissance du mot.* Colloque *Phonologie du français contemporain : variation, interfaces, cognition*. MSH, Paris.
- 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.
- Poiré, F. & Gurski, C. (2004). Distribution de schwa et des voyelles hautes en français de Windsor. Colloque Phonologie du français et théorie phonologique. Université de Calgary (Canada).
- Poiré, F., Moroz, N. & Kelly, S. (2008). PFC et variation diachronique : Montréal 1971-2008. Colloque *Structure des français en contact*. Université Tulane, Nouvelle-Orléans (Etats-Unis).
- Racine, Isabelle, Detey, Sylvain, Bühler, Nathalie, Schwab, Sandra, Zay, Françoise & Kawaguchi, Yuji, 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*, Poznan.

Références

- 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.
- Yoon S.-Y., Pierce L., Huensch A., Juul E., Perkins S., Sproat R. & Hasegawa-Johnson M. (2009). Construction of a rated speech corpus of L2 learners? speech. *CALICO Journal* 26(3): 662-673.
- Wester, Mirjam, Kessens, Judith, Cucchiarini, Catia & Strik, Helmer, 2001, Obtaining Phonetic Transcriptions : A Comparison between Expert Listeners and a Continuous Speech Recognizer, *Language and Speech*, 44 (3), 377-403.
- Zechner, K. (2009). What did they actually say? Agreement and Disagreement among Transcribers of Non-Native Spontaneous SpeechResponses in an English Proficiency Test. *Proceedings of the ISCA SLaTE-2009 Workshop*, Wroxall, UK, September.