Interphonologie, corpus et français langue étrangère: le projet IPFC

Sylvain DETEY[1] & **Isabelle RACINE**[2]

[1]SILS, Waseda University (Japon) & E.A. 4305 LiDiFra, Université de Rouen (France)

[2]ELCF, Université de Genève (Suisse)

IPFC2010 – 8 décembre 2010 – Paris MSH

Plan

- 1) Introduction
- 2) Le projet IPFC
- 3) Le cadre IPFC
- 4) Travaux en cours
- 5) Perspectives
- 6) IPFC2010

1) Introduction

Objectif de cette journée d'étude:

Interphonologie, corpus oraux, français langue étrangère

Interphonologie:

- *regain d'intérêt dans les 90's grâce à OT, mais assez peu dans le domaine francophone*
- *davantage de travaux en phonétique qu'en phonologie*

Corpus oraux:

- *essentiellement en L1*
- *axés sur la morphologie, le lexique et la syntaxe (cf. Flloc)*

Français langue étrangère:

- *longue histoire de l'enseignement/apprentissage de la prononciation, mais une diversité "non-unifiante" (phonétique, phonologie, didactique (très hétéroclite) intégration insuffisante de la psycholinguistique (cf. SLA))*
- *Absence de grande manifestation (cf. New Sounds)*

Corpus oraux en L2: répertoire des corpus d'apprenants (écrits et oraux) du *Center for English Corpus Linguistics* de l'Université Catholique de Louvain (S. Granger & D. Goosens)

UCL Université catholique de Louvain — Humanities and Social Sciences Sector

CENTRE FOR ENGLISH CORPUS LINGUISTICS

s and Social Sciences Sector > Centre for English Corpus Linguistics > Learner corpora around the world

Learner corpora around the world

This list is very much work in progress. We would like it to be as comprehensive as possible. If you have a learner corpus or know of one that is not listed on this webpage, send a me
Sylviane Granger and we'll add it to the list. We hope you will find the list useful for your research!

Corpus	Target language	First language	Medium	Text type/ task type	Proficiency level	Size in words	Project director	Av
The Pilot Arabic Learner Corpus PDF	Arabic	English	written		Intermediate and advanced	c. 9,000	Ghazi Abuhakema Reem Faraj Anna Feldman Eileen Fitzpatrick Montclair State University, USA	
Leerdercorpus Nederlands als Vreemde Taal	Dutch	French	written				Liesbeth Degand Université catholique de Louvain, Belgium	
Asao Kojiro's Learner Corpus Data	English	Japanese	written	Essays and stories written or reproduced by Japanese college students.			Asao Kojiro	Te
The Barcelona English Language Corpus (BELC)	English	Spanish Catalan	spoken and written	4 tasks: Written composition Oral narrative Oral interview Role-play Longitudinal data (children and young adults learning English)			Carmen Muños University of Barcelona, Spain	Sp av

- **103 corpus!**

(dont 64 corpus exclusivement consacrés à l'apprentissage de l'anglais)

Corpus de français langue étrangère :

The *Chy*-FLE (Cypriot Learner Corpus of French)	French	Modern Greek (and Cypriot Greek)	written	Argumentative and descriptive essays	From intermediate to advanced	c. 250,000 (under development)	Freiderikos Valetopoulos Université de Poitiers, France In collaboration with the University of Cyprus
The COREIL corpus	French English		spoken				Elisabeth Delais-Roussarie Hiyon Yoo Université Paris-Diderot, France
The "Dire Autrement" corpus	French (Second Language)	Mainly L1 speakers of English	written	Narrative, injunctive, persuasivle and informative texts		48,114	Marie-Josée Hamel Jasmina Milicevic Dalhousie University, Canada
French **I**nterlanguage **Da**tabase (FRIDA)	French	various	written				Sylviane Granger Centre for English Corpus Linguistics Université catholique de Louvain, Belgium
French Learner Language Oral Corpora (FLLOC)	French	various	spoken	See description of the 7 corpora	various		Florence Myles Newcastle University Rosamund Mitchell University of Southampton, UK
The InterFra corpus	French	Swedish	spoken	Interviews, retellings of video clips and picture stories	various		Inge Bartning Stockholm University, Sweden. interfra@fraita.su.se
The "Interphonologie du Français Contemporain" (IPFC) corpus	French	Japanese Spanish ?	spoken	Reading aloud, repeating words, guided interviews, interactions between two learners.	various	under development	Sylvain Detey Waseda University, Japan Université de Rouen, France Isabelle Racine Université de Genève, Switzerland Yuji Kawaguchi Tokyo University of Foreign Studies, Japan
The LCF corpus (Learner Corpus French)	French	Dutch	written	Argumentative essays Informative texts Journalistic texts Formal letters Summaries Written compositions by Flemish students of French	From intermediate to advanced	490,000	K.U.Leuven Campus Kortrijk, UGent and Lessius Hans Paulussen
The Lund CEFLE Corpus (Corpus Écrit de Français Langue Étrangère)	French	Swedish	written	Descriptive and narrative essays; picture-based stories.	various	100,000	Malin Ågren Lund University, Sweden
The UWi (University of the West Indies) learner corpus	French	English and Jamaican Creole	spoken	Conversations during oral exams and in informal contexts	various		Hugues Peters University of New South Wales, Sydney, Australia

- Français:
  - 10 corpus
  - 5 corpus oraux :
    - FLLOC (R-U),
    - Interfra (Suède),
    - COREIL (France),
    - UWI (Jamaique) et…
    - IPFC (International).

→ ***IPFC:***

InterPhonologie du Francais Contemporain:

usages, variétés, structureS

(vs* PFC : ... structurE ?*)

2) Le projet IPFC

- Travaux sur corpus en phonétique / phonologie en L2?

  → Travaux d'actualité 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)

  ⇨ Focus sur les aspects segmentaux et suprasegmentaux (Trouvain & Gut, 2007 ; Meng, Tseng, Kondo, Harrison & Viscelgia, 2009)

  ⇨ Absence de données similaires pour le français...

- Or dans PFC:
  - Intérêt pour l'enseignement/apprentissage du français

    (programme PFC-EF depuis 2006)
  - Présence de sujets bilingues (Canada, Belgique, Suisse...) et plurilingues (Algérie, Sénégal, Côte d'Ivoire...)

→ Sujets non-natifs?

Le projet « InterPhonologie du Français Contemporain »

(lancé en 2008, Detey & Kawaguchi)

IPFC: historique

Historique

Décembre 2008 : lancement du projet IPFC (IPFC-japonais – S. Detey et Y. Kawaguchi)

Juin 2009 : ajustement du protocole (IPFC-espagnol – I. Racine (avec F. Zay et S. Schwab)).

Juillet 2009 – décembre 2009 : récolte et première exploitation des corpus japonais et espagnol.

Printemps 2010 : préparation de Moodle multilingue et du site IPFC (Y. Kawaguchi).

Septembre 2010 : lancement d'IPFC-norvégien, allemand, néerlandais, anglais canadien, grec.

Décembre 2010 : organisation de la première Journée d'étude IPFC à Paris IPFC2010 : Interphonologie, corpus et français langue étrangère.

- Travaux récents dans IPFC:

Evaluation des voyelles nasales chez les japonophones et hispanophones

→ *New Sounds 2010* (Pologne)

→ *CMLF 2010* (Etats-Unis)

Les deux premiers corpus:

- japonais (Y. Kawaguchi, TUFS)
  - 100 pour répétition & lecture
  - 12 pour le protocole complet

- espagnol (I. Racine, U. Genève)
  - 14 à Genève (complet)
  - 5 à Madrid (complet)

3) Le cadre IPFC

- Protocole adapté de celui de PFC et quasiment identique pour les différentes L1s ⇨ assure la comparabilité des données:
  - Entre natifs (projet PFC) et apprenants
  - Entre apprenants de différentes L1

- Le protocole comprend 6 tâches (environ 1h de données orales par apprenant):
  - Répétition d'une liste de mots spécifique à la L1 des apprenants comprenant
    - 34 éléments communs pour tous les apprenants (voyelles nasales, [y/u], etc.)
    - Entre 25-35 mots avec difficultés spécifiques à la L1 (ex. japonais et espagnol: b/v)
  - Lecture de la liste de mots PFC
  - Lecture de la liste de mots spécifique
  - Lecture du texte PFC
  - Entretien guidé avec un natif (2 niveaux pour les questions ouvertes: A1-B1 et B2-C2)
  - Interaction semi-contrainte entre deux apprenants

+ questionnaire biographique (22 questions)

- Récolte des données:
  - Manuelle
  - Semi-automatique ⇨ plateforme Moodle pour les 4 premières tâches (cf. présentation Matsuzawa)

  ⇨ format: wav, mono, échantillonage 22’050KHz, 16 bits

- Transcription des données sous Praat (Boersma & Weenink, 2009):
  - Transcription orthographique (et, si passation manuelle, alignement) des listes de mots
  - Transcription orthographique du texte avec transcription cible et effective
  - Transcription orthographique des conversations (cf présentation Racine, Detey, Zay & Kawaguchi)

  (+ question du codage ⇨ cf. présentation Racine, Detey, Zay & Kawaguchi)

4) Travaux en cours

Technique

- plateforme Moodle
- site IPFC

Méthodologique

- transcription
- codage

Expérimental

- b-v (IPFC-japonais et IPFC-espagnol)
- Voyelles nasales, occlusives sonores et accentuation dans IPFC-espagnol (projet FNS de I. Racine)

Corpus

- En préparation: norvégien, anglais canadien, grec chypriote, néerlandais, allemand
- En discussion: mandarin standard (Taiwan), suédois, danois, italien

Rencontre

- Workshop à TUFS (Tokyo) en mars 2011
- Workshop à Paris en décembre 2011 (IPFC2011)

Publication

- chapitre sur la variation chez les non-natifs (vol. OUP Detey, Durand, Laks, Lyche - en préparation)

5) Perspectives

- Nombreuses perspectives:
  - Exploitation des corpus existants:
    - Comparaisons inter-tâches
    - Comparaisons inter-L1
    - Comparaisons natifs/non-natifs
    - Analyses d'autres éléments (/b-v/, nasales, liquides, accentuation…)
    - Exploitation pour des analyses lexicales, morphologiques, syntaxiques, etc.
  - Développement du protocole: production, perception et longitudinal
  - Développement du corpus: autres L1
  - Collaboration entre projets (AESOP…)
  - Collaboration dans le domaine du TAL
  - …
  - Et exploitations pédagogiques!

  (pour la formation des enseignants mais aussi pour les étudiants)

  → IPFC: un projet d'actualité

  (cf. *Wildcat Corpus of Native- and Foreign-accented English* (Van Engen *et al.* 2010), *Rated L2 speech corpus* (Yoon *et al.* 2009))

6) IPFC2010

- *Juana GIL-FERNANDEZ*
  Le rapport entre théorie et pratique dans l'enseignement de la prononciation des langues étrangères

- *Sandra SCHWAB & Joaquim LLISTERRI*
  Vers une méthodologie pour l'étude de l'accentuation en langue étrangère

- *Takeki KAMIYAMA*
  Production des /u y ø/ français chez des apprenants japonophones : des phones phonétiquement et/ou phonémiquement nouveaux

- *Mariko KONDO & Yoshinori SAGISAKA*
  Constructing Asian English Corpus for Universal Purposes

- *Isabelle RACINE, Sylvain DETEY, Françoise ZAY & Yuji KAWAGUCHI*
  Enjeux méthodologiques dans les corpus oraux en L2 : transcriptions, annotations et codages

- *Jacques DURAND & Chantal LYCHE De PFC à IPFC : perspectives*

- *Chantal LYCHE & Unni LUNDSTEDT IPFC-norvégien : protocole et contexte*
- *Jeff TENNANT, Jade SHAPIRO & Nerissa TAYLOR IPFC-anglais canadien : protocole et contexte*
- *Freiderikos VALETOPOULOS & Marina CHRISTOFI IPFC-grec : protocole et contexte*
- *Dominique NOUVEAU & Janine BERNS IPFC-néerlandais : protocole et contexte*
- *Juri CHERVINSKI & Elissa PUSTKA IPFC-allemand : une pré-enquête auprès de quelques étudiants munichois*
- *Trudel MEISENBURG & Majana GRÜTER IPFC-allemand : aspects prosodiques*
- *Mito MATSUZAWA & Yuji KAWAGUCHI Les supports informatiques de IPFC : plateforme Moodle et site IPFC*

Merci et bonne journée IPFC2010 !

Quelques références IPFC

- Detey, S. (2009). Phonetic input, phonological categories and orthographic representations: a psycholinguistic perspective on why oral language education needs oral corpora. The case of French-Japanese interphonology development. In Kawaguchi, Y., Minegishi, M. & Durand, J. (eds), *Corpus Analysis and Variation in Linguistics*. Amsterdam/Philadelphia: John Benjamins, 179- 200.
- Detey, S., Durand, J. & Nespoulous, J.-L. (2005). Interphonologie et représentations orthographiques. Le cas des catégories /b/ et /v/ chez des apprenants japonais de FLE. *Revue PArole* 34/35/36 (supplément), 139-186.
- Detey, S. & Kawaguchi, Y. (2008). Interphonologie du français contemporain (IPFC) : récolte automatisée des données et apprenants japonais. *Phonologie du français contemporain : variation, interfaces, cognition*. Paris.
- Detey, S. & Nespoulous, J.-L. (2008). Can orthography influence L2 syllabic segmentation? Japanese epenthetic vowels and French consonantal clusters. *Lingua* 118 (1), 66-81.
- Detey, S., Racine, I., Kawaguchi, Y., Zay, F., Bühler, N., Schwab, S. (2010). Evaluation des voyelles nasales en français L2 en production : de la nécessité d'un corpus multitâches. In Neveu, F., Muni-Toké V., Durand, J., Klingler, T., Mondada, L. et Prévost S. (éds.), Actes de CMLF'10, « Phonétique, phonologie et interfaces », Paris : ILF, 1289-1301.
- Racine, I., Detey, S., Zay, F. & Kawaguchi, Y. (2009). Interphonologie du français contemporain : réflexions méthodologiques et premières données d'apprenants hispanophones et japonophones. *AFLS 2009*, *Langue française en contextes,* Université de Neuchâtel (Suisse), Sept. 2009.
- 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. In Dziubalska-Kolaczyk, K, Wrembel, M. & Kul, M. (éds.), Proceedings of New Sounds 2010, Poznan: Adam Mickiewicz University, 367-372.
- 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 A. Kamber et C. Skupiens (éds). *Recherches récentes en FLE*. Berne : Peter Lang.
- Racine, I., Zay, F., Detey, S. & Kawaguchi, Y. (à paraître), « De la transcription de corpus à l'analyse interphonologique: enjeux méthodologiques en FLE ». In *Travaux Linguistiques du CerLiCO* 24, Rennes: PUR.

Références présentation

- **Répertoire des corpus oraux en L2: Granger, S. & Goosens, D. http://www.uclouvain.be/en-cecl-lcWorld.html**

- 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. Pp. 331-347.
- Boersma, P. & Weenink, D. (2009). *Praat : doing phonetics by computer* (version 5.0), http://www.praat.org.
- 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.
- Detey, S., Durand, J., Laks, B. & Lyche, C. (éds). (en préparation). *Varieties of Spoken French : a source book*. Oxford : Oxford University Press.
- Gut, U. (2009). Non-native Speech: *A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. Wien: Peter Lang.
- 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.
- 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 SpeechResponses in an English Proficiency Test. *Proceedings of the ISCA SLaTE-2009 Workshop*, Wroxall, UK, September.
- Van Engen K., Baese-Berk M., Baker R.E., Choi A., Kim M., Bradlow A.R. (2010). The *Wildcat Corpus* of Native- and Foreign-Accented English: communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*. published online Sept. 1st 2010.
- 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.