THE INTERPHONOLOGY OF CONTEMPORARY FRENCH (IPFC): AN INTERNATIONAL CORPUS-BASED L2 PHONOLOGY RESEARCH PROGRAMME

Sylvain Detey, Isabelle Racine

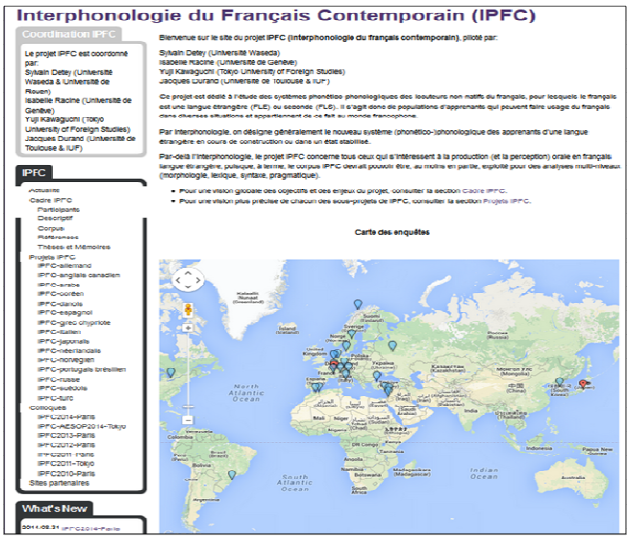
Waseda University (SILS), Geneva University (ELCF)
detey@waseda.jp, Isabelle.Racine@unige.ch

Keywords: corpus, French, interphonology, IPFC

1. ORIGINS

In 2014, the first *Handbook of Corpus Phonology* [10] was published by Oxford University Press, including a chapter describing the field of L2 corpus-based phonology written by Ulrike Gut [14], five years after the publication of her monograph about the LeaP corpus [13], which can be considered as a landmark corpus in the area. The inclusion of such a chapter in the handbook is an acknowledgement of the need to extend the methods and principles of corpus phonology to the field of non-native speakers, at a time when L2 pronunciation research is getting over the past theoretical or disciplinary boundaries (phonology vs phonetics; linguistics vs psycholinguistics; L2 acquisition studies vs L1 variationist sociolinguistic studies) and benefits from a new momentum. Unsurprisingly, L2 English has been the focus of most L2-oriented corpus-based phonology-phonetics projects (e.g. LeaP [13], AESOP [21]), but L2 Dutch has also been a proficient domain in connection with educational objectives in the field of Computer-Assisted Pronunciation Training programmes [16]. It is only recently that similar ventures have been launched for French [2, 15], with the *Interphonologie du Français Contemporain* (IPFC) project at its forefront. The IPFC project originally stemmed from the *Phonologie du Français Contemporain* (PFC) research programme [11]: a large-scale phonological survey across the French-speaking world launched in the late 1990s by three French phonologists, Jacques Durand, Bernard Laks and Chantal Lyche, who had become wary of the often limited set of data on which most French phonology had been relying. Standing against 'armchair' linguistics and in tune with the technical developments of oral databases and advances in sociophonetics, they designed a recording protocol inspired by William Labov's work which could be used with most speakers to ensure data comparability. The PFC corpus (www.projet-pfc.net) turned out to include multilingual speakers (in Africa, Canada, Louisiana), whose status as 'native' speakers of French could sometimes be debatable. Hence the launch of a dedicated 'non-native' avatar of PFC in 2008, the first of its kind for L2 French, with a Japanese team [5], immediately followed by a Spanish team, trying to combine the principles of PFC with the lessons learned from other projects, particularly in L2 English [13, 14]. The IPFC project now includes different teams representing 16 first languages (German, Alemanic, English, Arabic, Korean, Danish, Spanish, Greek, Italian, Japanese, Dutch, Norwegian, Portuguese, Russian, Swedish, Turkish) (Fig. 1), whose members (linguists and language educators) share an interest in L2 French pronunciation and are keen on exploring non-native varieties as part of the sociolinguistic continuum on which native, plurilingual and non-native speakers can be examined. It is important to note that the educational perspective is an integral part of the project, which partly explains some of its methodological orientations [9].

Figure 1: The website of the IPFC project with the different teams and survey points.



2. METHODOLOGY

2.1. Recording protocol

The protocol is made up of 6 tasks, two of which (b and d) are common to the PFC project for comparability purposes: a) repetition of an L1-specific wordlist[1], b) reading out of the PFC wordlist, c) reading out of the L1-specific wordlist[2], d) reading out of the PFC text, e) interview with a native speaker, f) interaction between two non-native speakers. A sociolinguistic questionnaire as well as a consent form are also included.

2.2. Data processing

Following the rationale adopted in PFC regarding data interoperability on the one hand and the pitfalls of phonetic transcriptions for large data-sets on the other hand, the audio recordings are orthographically transcribed with text-to-sound alignment in Textgrid files used with *Praat* [1]. Specific transcription conventions were designed to handle the characteristics of non-native speech [20].

2.3. Data analyses

Since one of the objectives of the corpus was to process as automatically as possible large sets of data, and following PFC's stance on variation, with a strong educational perspective in the case of IPFC, we decided to adopt and extend the coding system used in PFC for schwa and liaison, as an intermediate step between rough perceptual categorization (correct/incorrect) and fine-grained acoustic analysis (with its limits) [4]. The overall structure of each code is divided into four sections: 1) target structure, 2) left context, 3) right context, 4) perceptual assessment, primarily in terms of target-likeness (e.g. for nasal vowels: nasality, quality, postvocalic consonantal excrescence [8]). Alphanumeric codes were designed for consonants, oral vowels, nasal vowels, liaison and consonant clusters, and human coders, on the basis of their perceptual assessment of the production, insert the code in the orthographic transcription right after the structure under scrutiny, using separate tiers for each phenomenon (Fig. 2). The files are then analyzed with *Dolmen*, a phonological concordancer developed for the project by Julien Eychenne [12] (Fig. 3), which allows users to perform queries in the coded corpus and recover the requested items with several options. *Dolmen* provides descriptive statistics for code-based queries, and digs out the corresponding occurrences in concordance lines with the possibility of opening the sound files in *Praat*. A multiple-blind assessment option is included in the system.

Figure 2: An example of a sequence from the PFC text coded for nasal vowels by two different coders, and for liaison by one coder in a TextGrid file opened with Praat.

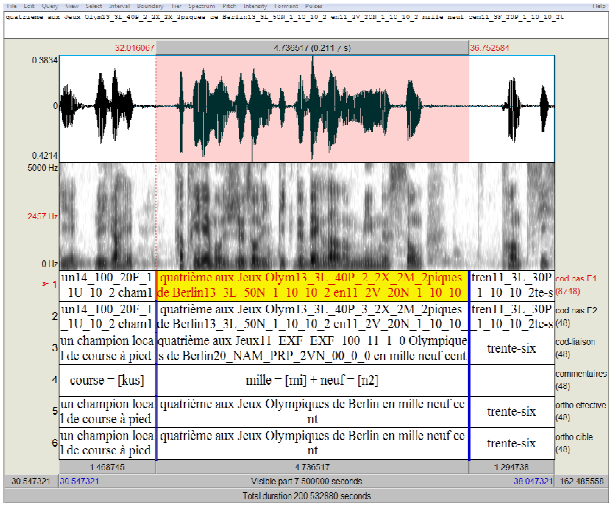
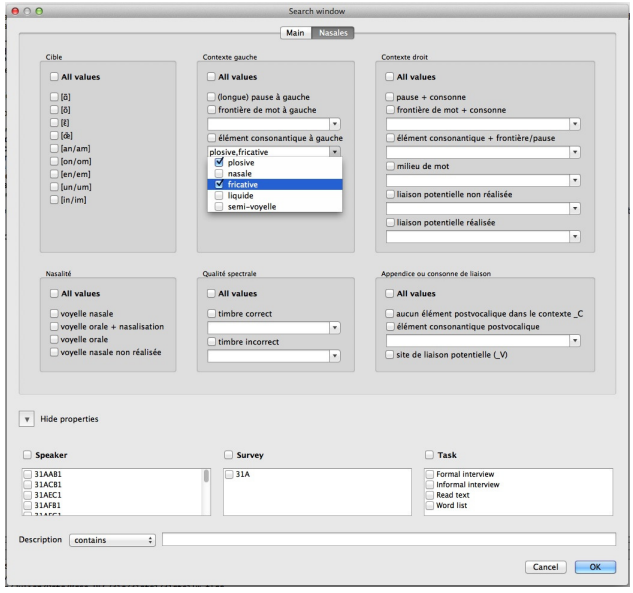


Figure 3: The Dolmen interface for the nasal vowels in the IPFC project.



Despite its obvious limitations in terms of phonetic description accuracy, this approach has proved successful so far (congruency between multiple subject psycholinguistic tests and coded results for isolated words), offering several advantages (especially for continuous speech): it can delineate specific data subsets for extensive analyses, and be used in the framework of perceptual studies, with a clear connection to perceptual norms. Last but not least, it really fits, in our view, with the overall objective and method of corpus phonology (but not laboratory phonetics at this stage).

3. ILLUSTRATIONS

Among the common objects of interest for all IPFC teams, nasal vowels and *liaison*, two typically difficult structures to be mastered by learners of French, have been extensively studied in the project [4, 7, 17, 18]. For several reasons (e.g. the relative difficulty of acoustical measures), the French nasal vowels have been a good benchmark to test and develop our approach, both with beginners and advanced learners. As for *liaison*, which has been a central object of study in the PFC project, it is also of particular interest in the case of L2 learners since it must be analyzed in a multidimensional manner: both from a segmental and a suprasegmental perspectives, but also at the interface between phonology, morphosyntax, lexis and orthography. Other elements have also been studied (high rounded vowels, voiced plosives, liquid consonants, lexical stress), and each team has its own focus (e.g. final consonant devoicing among Germanic languages speakers and vocalic epentheses among Japanese learners). Having a common coding system for all L1-specific surveys, with the *Dolmen* application to perform cross-corpus queries, is one of the main methodological assets of the IPFC project to carry out comparative analyses between different groups of learners (e.g. *liaison* production by Italian and Spanish learners of French, with 4788 coded liaison sites [19]).

4. PERSPECTIVES[3]

Building up an international database such as the one we are striving to achieve in IPFC takes time. Even though most of the methodological features of the project are now set, we are still in the process of developing: (i) a full-fledged searchable database, (ii) automatic functions in Dolmen to provide richer descriptions of the learners' productions, (iii) guidelines to evaluate our data with semi-manual acoustic analyses on the one hand and automatic machine assessment on the other hand, (iv) pedagogical applications for syllabus design and pronunciation training. For more information about the IPFC project, see: http://cblle.tufs.ac.jp/ipfc/.

5. REFERENCES

[1] Boersma, P., Weenink, D. 2015. *Praat: doing phonetics by computer*. Version 5.4.08.

[2] Delais-Roussarie, E., Yoo, H. 2011. Learner corpora and prosody: from the COREIL corpus to principles on data collection and corpus design. *Poznań Studies in Contemporary PSCIL*, 471: 28-39.

[3] Detey, S. 2005. *Interphonologie et représentations orthographiques. Du rôle de l'écrit dans l'enseignement / apprentissage du français oral chez des étudiants japonais*. PhD dissertation, University of Toulouse-Le Mirail.

[4] 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: John Benjamins, 229-250.

[5] 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, 11-13 Dec. 2008.

[6] Detey, S., Nespoulous, J.-L. 2008. Can orthography influence L2 syllabic segmentation? Japanese epenthetic vowels and French consonantal clusters. *Lingua. International Review of General Linguistics*, 118(1), 66-81.

[7] Detey, S., Racine, I., Eychenne, J., Kawaguchi, Y. 2014. Corpus-based L2 phonological data and semi-automatic perceptual analysis: the case of nasal vowels produced by beginner Japanese learners of French. *Proc. 15th Interspeech* Singapore, 539-544.

[8] Detey, S., Racine, I., Kawaguchi, Y. 2014. Des modèles prescriptifs à la variabilité des performances non-natives : les voyelles nasales des apprenants japonais et espagnols dans le projet IPFC. In: Durand, J., Kristoffersen, G., Laks, B. (eds), *La phonologie du français : normes, périphéries, modélisation. Hommage à Chantal Lyche*. Paris: Presses Universitaires de Paris Ouest, 197-226.

[9] Detey, S., Racine, I., Kawaguchi, Y., Zay, F. in press. Variation among non-native speakers: the InterPhonology of Contemporary French. In: Detey, S., Durand, J., Laks, B., Lyche, C. (eds), *Varieties of Spoken French: a Source Book*. Oxford: Oxford University Press.

[10] Durand, J., Gut, U., Kristoffersen, G. (eds) 2014. *The Oxford Handbook of Corpus Phonology*. Oxford: Oxford University Press

[11] Durand, J., Laks, B., Lyche, C. 2014. French phonology from a corpus perspective: the PFC programme. In: Durand, J., Gut, U., Kristoffersen, G. (eds), *The Oxford Handbook of Corpus Phonology*. Oxford: Oxford University Press, 486-497.

[12] Eychenne, J., Paternostro, R. in press. Analyzing transcribed speech with Dolmen. In: Detey, S., Durand, J., Laks, B., Lyche, C. (eds), *Varieties of Spoken French: a Source Book*. Oxford: Oxford University Press.

[13] Gut, U. 2009. *Non-native Speech: a Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. Bern: Peter Lang.

[14] Gut, U. 2015. Corpus phonology and second language acquisition. In: Durand, J., Gut, U., Kristoffersen, G. (eds), *The Oxford Handbook of Corpus Phonology*. Oxford: Oxford University Press, 286-301.

[15] Landron, S., Paillereau, N., Nawafleh, A., Exare, C., Ando, H., Gao, J. 2010. Le corpus PhoDiFLE : un corpus commun de français langue étrangère pour

une étude phonétique des productions de locuteurs de langues maternelles plurielles. *Cahiers de Praxématique*, 54/55, 73-86.

[16] Neri, A., Cucchiarini, C., Strik, H. 2006. Selecting segmental errors in L2 Dutch for optimal pronunciation training. *IRAL*, 44, 357-404.

[17] Racine, I., Detey, S. (eds) to appear. L'apprentissage de la liaison en français par des locuteurs non natifs : éclairage des corpus oraux. Special issue of *Bulletin VALS-ASLA*, 102.

[18] Racine, I., Detey, S. Buehler, 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. *Proc. 6th New Sounds* Poznan, 367-372.

[19] Racine, I., Paternostro, R., Falbo, C., Janot, P., Murano, M. 2014. La liaison chez les hispanophones et les italophones : du texte lu à la conversation. *Rencontres FLORAL 2014 : corpus oraux et enseignement de la prononciation en FLE & interphonologie et corpus oraux*. Paris, 8-9 Dec. 2014.

[20] Racine, I., Zay, F., Detey, S., Kawaguchi, Y. 2011. De la transcription de corpus à l'analyse interphonologique : enjeux méthodologiques en FLE. *Travaux Linguistiques du CerLiCO*, 24, 13-30.

[21] Visceglia, T., Tseng, C.-Y., Kondo, M., Meng, H., Sagisaka, Y. 2009. Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project). *Oriental-COCOSDA*, Urumuqi.

---

[1] The lists include items common to all surveys (e.g. words with nasal vowels, since these vowels are difficult to acquire for most learners of French) and words specifically chosen for specific groups of learners (e.g. items with consonantal clusters for Japanese learners).

[2] The L1-specific wordlist repetition-reading tasks included in the protocol aim at taking into account the impact of the orthographic factor in the elicitation process [3, 6].

[3] The research presented here has been partly supported by the Japanese Society for the Promotion of Science through two Grants-in-Aid for Scientific Research (B) n°2332012 and n°15H03227 to S. Detey, as well as by the University of Geneva and by the Fonds National Suisse de la Recherche Scientifique (subside n° 100012_1321441 to Isabelle Racine). We wish to thank Yuji Kawaguchi (co-director of the IPFC project), Julien Eychenne, Jacques Durand, Chantal Lyche, Bernard Laks, Mariko Kondo, Françoise Zay, Roberto Paternostro, as well as all the IPFC colleagues and participating students.