CBFC: a parallel L2 speech corpus for Korean and French learners

Hiyon Yoo*, Inyoung Kim+
*Laboratoire de Linguistique formelle CNRS 7110, Université Paris Diderot, Place Paul Ricoeur, Paris 75013, France
+ Naver Labs Europe, 6 chemin de Maupertuis, Meylan 38240, France
*yoo@linguist.univ-paris.diderot.fr , +inyoung.kim@naverlabs.com

Abstract

In this paper, we present the design of a bilingual corpus of French learners of Korean and Korean learners of French using the same experimental design. This language resource contains mainly speech data, gathered among learners with different proficiency levels and in different speaking contexts (read and spontaneous speech). We aim at providing a translated and annotated corpus to the scientific community which can be used for a large array of purposes in the field of theoretical but also applied linguistics.

Keywords: L2 corpora, Korean, French, phonetics, prosody, phonetic segmentation, alignment

1. Introduction

In the last years, especially with the development of new methodologies and new technologies in corpus and computational linguistics, language corpora have become more and more common and needed in linguistic research. Corpora vary a lot in size, uses or presentations but it appears clearly that doing research using linguistic corpora brings along new questions but also new analyses of linguistic phenomena. Recently, we saw a rise of large corpora for second language acquisition (cf. among others, Granger 2003 and 2012, Hawkins and Buttery 2009). Beside the fact that few corpora are freely available to the research community (see however Milde and Gut 2002, Tortel 2008, Herment et al. 2012a and 2012b, and Gut 2009), a glimpse at the « Learner corpora around the world » database (http://www.uclouvain.be/en-cecl-lcworld.html) reveals that:

- Most existing corpora concern English
- Most existing corpora are written and not spoken
- The pair of languages French-Korean is quasi-inexistent.

Still, the use of large corpora allows a better evaluation of possible correlations between the learner's L1 (first language), his grammatical competence and his proficiency level in L2 (second language) (following for example the Common European Framework of Reference for Languages). Corpus-based studies can be used to determine how some morpho-syntactic phenomena are acquired in English as a foreign language (see for example the project English Profile, Hawkins and Buttery 2009), how students' pronunciation of L2 can be influenced by his L1 (I-PFC project, Racine et al. 2011, 2012) and so on. Depending on how the corpus is built, such corpora can also be useful for understanding learners' opinion of the target language during education.

In this paper, we present a bilingual corpus of French learners of Korean and Korean learners of French which aims at giving an answer to this missing linguistic resource by putting in parallel two languages, French and Korean.

There are today numerous corpora of French as L1, be they written or spoken (the Frantext corpus (http://www.frantext.fr/), TCOF corpus (http://www.cnrtl.fr/corpus/tcof/), PFC (http://www.projet-pfc.net/) and so on). For instance, the TUFS corpus is a representative example of corpora gathered in Aix–en-Provence where the main concern is spontaneous French. Another ambitious project is the PFC corpus, which aims at gathering data of spoken French all over the world, using the same methodology. Thus, the PFC project is based on spoken tasks only (there are no written tasks), with reading tasks (words, short texts) and also spontaneous speech (mainly interviews). One of the recent developments of PFC was to include foreign learners of French (the I-PFC project, Racine et al. 2012) following the same corpus collection protocol but adding tasks in respects to the L1). The I-PFC includes also Korean learners of French. Moreover, the I-PFC concerns only French as a second language and the equivalent for Korean using the same material is not available. The CEFLE corpus (Corpus Ecrit de Français Langue Etrangère) developed at the Lund University (http://projekt.ht.lu.se/cefle/information) is an illustration of a written corpus of French as a Second Language, but concerns Swedish learners.

Most existing corpora for Korean are written (the Korean National Corpus http://www.sejong.or.kr/user/main.do) or concern the English-Korean pair (the Gachon Korean Corpus). The I-PFC is the only project that presents the pair of language French-Korean but, as mentioned before, concerns only Korean learners of French and more specifically on their phonological competence (Han, 2011). Moreover, there are not any published results yet, and it is impossible to say how much the project has advanced.

Section 2 presents in detail the experimental design of the CBFC corpus; section 3 gives information on how segmentation and annotation was dealt.

2. The CBFC[1] Project

The "Corpus Bilangue Français Coréen" (CBFC) is a L2 corpus putting in parallel French learners of Korean and Korean learners of French. The corpus has been gathered following a unique protocol for the two populations and is made to be as complete as possible, with both written and spoken data, but with special emphasis on the prosodic aspects of L2 speaker's productions. Our aim is to make available to the research community interested in L2 phonology and prosody from this resource for the French-Korean language pair. In this section, we detail the protocol we used for gathering the spoken data.

[1] CBFC stands for "Corpus Bilangue Français Coréen" (French Korean Bilingual Corpus). The term *bilangue* is to be distinguished from *bilingue* in that we are not dealing with a bilingual corpus but a parallel corpus between two languages.

2.1 Subjects and Recording Procedure

A total of 22 subjects participated to the recordings of the speech data: one native speaker for each language, and 10 learners for each language. We established a linguistic profile for each speaker, with information concerning their knowledge of other languages, their L2 proficiency and so on. Some of the detail of the participants is given in tables 1 and 2.

French subjects for L2 Korean	Age*	Level**	Bilingual	Other known second languages
F01	18	2y		English, German
F02	17	4y		English, Arabic, Spanish
F03	18	3y	Lingala	English, Spanish
F04	17	3y	Turkish	English, Spanish,
F05	27	2y	Antillian creole	English, Russian
F06	18	1.5y	Lingala	English, Italian, Spanish,
F07	17	2.5y	Italian	English, Italian
F08	19	5y		English, Russian, Italian
F09	n/a	6m		
F10	n/a	2y		

Korean subjects for L2 French	Age*	Level**	Stay in France***	Education
M01	23	1y	21d	International relations
F01	19	2y	15d	French language and literature
F02	15	5y	15d	French language and literature
F03	21	1m	15d	Economy
F04	16	5y	1m	French language and literature
M02	19	4y	n/a	French language and literature
F05	26	8y	8y	Costumier
M03	25	3.7y	3.7y	Cinema
F06	20	7y	7y	Architecture
M04	31	3y	3y	Pianist

Table 1: Details of French and Korean participants: *age when subject started learning L2, **years of study, ***period time of staying at the immerging country; years(y), months(m), days(d)

Recording sessions took place at the recording room of University Paris Diderot in France. We used a portable Roland R-26 and a XLR microphone. Each recording session lasted from 40 to 80 minutes depending on the subject. The subjects had the possibility to make a pause between two different tasks.

As subjects were recruited in Paris, the two language groups show different characteristics. As it is shown in Table 1, Korean participants had English as the only second language, and only one female participant considered herself as bilingual. There were several languages learned as second language, and five French participants considered themselves as bilingual. Korean participants had various education background compared to French who were recruited at the department of Korean language in Paris Diderot University.

2.2 The Corpus

The corpus is based on the COREIL protocol which was initially developed for the English/French (Delais-Roussarie and Yoo 2011) and extended to other language pairs such as Spanish/French (see Santiago, F. & E. Delais-Roussarie. (2012), Delais-Roussarie, Santiago, F. & Yoo H. 2015)). It is inspired by the AGILE corpus Voormann, H. & U. Gut. 2008, Gut 2009. This protocol which was initially designed for the study of L2 intonation and phrasing presents the advantage to be modular, and it is easy to add or remove tasks. It can also be adapted following the performance level of speakers. We thus enriched the initial protocol in order to obtain specific type of sentences.

We distinguish in this corpus three different groups of tasks:

- Reading
- Spontaneous Speech in monologue, dialogues, and interviews.
- ToBI questionnaire

The reading task is composed by three different texts: small monologues, small dialogues like those which appear in learners' textbooks (see Figure 1) and an excerpt of *Little Prince* (from Saint Exupéry). The texts are rich in type of sentences, and therefore, we expect subjects to produce various intonational contours in order to signal interrogatives or other illocutionary forces.

Server: *Bonjour. Une table pour deux personnes ?*
Client: *Oui, nous sommes deux. Vous avez un espace non-fumeur ?*
Server: *Bien sûr. Vous préférez cette table, ou celle-ci, près de la fenêtre ?*
Client: *Plutôt celle-ci.*
Server: *Très bien. Installez-vous. Voici le menu.*

Figure 1: Example of a dialogue reading task in L2 French

Spontaneous speech was obtained during four different tasks:

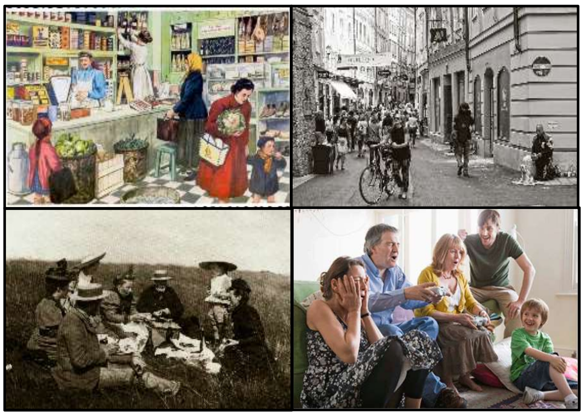


Figure 2: Images used for the image description task

- **Image description**: subjects are invited to describe and comment freely the image which is shown. Figure 1 gives an idea of the type of images they had to describe.
- **Question guided monologue**: the examiner asks questions and subjects are invited to answer freely.
- **Questionnaire completion**: subjects are provided with a questionnaire they had to fill, after interviewing the examiner.
- **"Who am I?" game**: subjects are invited to play a game where they had to ask questions in order to guess the face their co-gamer has chosen. Figure 3 gives an example of the type of data that were obtained. The motivation of the game setup is to obtain interrogative type of sentences with a coherent context given to each participant.

- mʌɾi-ga	noɾansɛg-iejo
hair-NOM	yellow-DEC *(Is the hair yellow?)*

- ne *(Yes.)*
- kopsɯlmʌɾi-ejo
  curly-DEC *(Does he or she has curly hairs?)*
- jak'an *(A little)*
- luk'a *(Lucas?)*
- madʒajo
  correct-DEC *(It's correct)*

Figure 3: Example of data from the *who am I?* game task

For the last task, we used the ToBI questionnaire used for gathering data of the IARI project and intonation in romance (Frota S. & Prieto P. 2015). In this project, the idea was to adopt the same framework and the same methodology in order to obtain comparable intonational contours for romance languages. This kind of questionnaire presents the advantage to force the subjects to perform certain utterances for a specific given context. However, it is a rather difficult task for beginners or intermediate learners.

Table 3 gives a mean value of the speech time for each task.

L2 French		Mean speech time (minutes)
Read speech	Texts	2.6
	Petit Prince	3.7
	Dialogues	4.5
Spontaneous speech	Monologue (Q-R)	16
	Image describing	5.3
	Free questionings	5.3
	Questioning game	9
Questionnaire ToBI		20
TOTAL		66

L2 Korean		Mean speech time (minutes)
Read speech	Texts	5
	Dialogues	8.6
Spontaneous speech	Monologue (Q-R)	14.2
	Image describing	5.8
	Free questionings	4.6
	Questioning game	7.1
Questionnaire ToBI		28
TOTAL		73.3

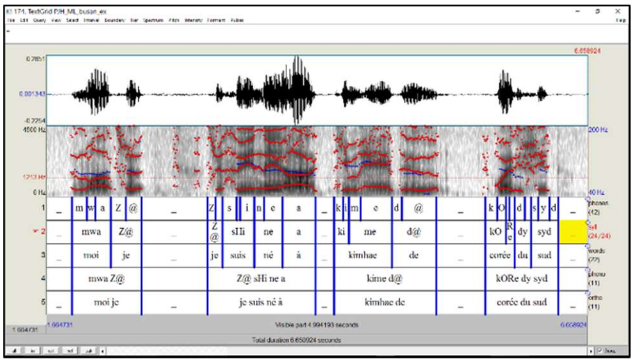
Table 3: Speech time by speech type for L2 French and L2 Korean

3. Segmentation and Annotation

As noted by Lüdeling and Hirschmann (2015), coding the learners' errors is a difficult task and it is important to minimize interpretation. For the moment, we proceeded at an orthographic transcription of the data. We then used semi-automatic tools in order to provide a segmentation and an annotation at the segmental level.

3.1 French Productions

French productions were aligned using the EasyAlign software (Goldman 2011). EasyAlign is an executable software that can be used in Windows environment only, which adds a plugin into the Praat software (2017). When provided with an orthographic transcription of the sound file, it gives as a result is a multi-tier annotation within the Textgrid in Praat, offering a segmentation into phones and words.



1 _ m w a Z @ _ Z s i n e a _ k i m e d @ _ k O R d s y d _
2 _ mwa Z@ _ Z@ sIIi ne a _ ki me d@ _ kO Re dy syd _
3 _ moi je _ je suis né à _ kimhae de _ corée du sud _
4 mwa Z@ Z@ sHi ne a kime d@ kORe dy syd
5 _ moi je _ je suis né à _ kimhae de _ corée du sud _
phones (42)
syl (24/24)
words (22)
phono (11)
ortho (11)
4500 Hz
1213 Hz
0 Hz
200 Hz
40 Hz
1.664731
6.658924
Visible part 4.994193 seconds
Total duration 6.658924 seconds

Figure 4: Image caption of a sentence in French produced by a Korean learner (M01, L2 French)

Figure 4 gives an image caption of the segmentation and annotation obtained after manual correction of the boundaries. EasyAlign provides five different tiers (phone, syllable, word, phonetization and orthographic level).

3.2 Korean Productions

Since the EasyAlign software is not yet available for Korean, we had to use another tool for segmenting and annotating Korean data. Korean data were aligned with the automatic alignment function supplied in the 'Interval' menu in Praat (version 6.0.25). The automatic alignment is proposed for many languages, and we tested the available option 'Korean-test'. We selected all the option boxes, i.e., 'word', 'phonemes', 'silences', and this function creates two subordinate tiers; '/word' tier and a '/phon' tier.

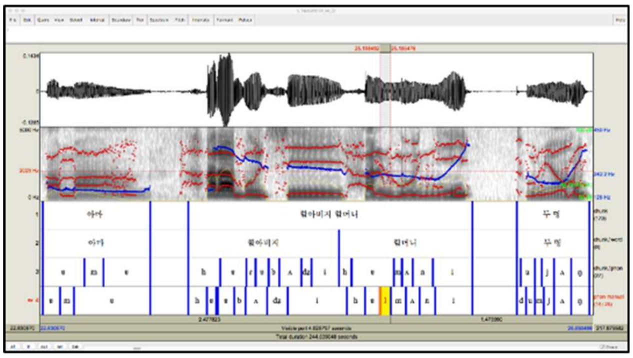
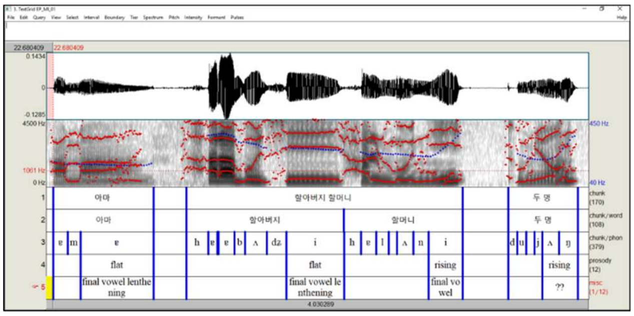


Figure 5: Image caption of a sentence in Korean produced by a French learner (F05, L2 Korean)

When Korean orthograph hangul is correctly transcribed and aligned in an interval tier (tier1 in Figure 5, named *chunk*), the second /*word* tier and the third /*phon* tier are created automatically within the boundary aligned to the first tier. The /*word* tier creates an *oejeol* unit, which can be a lexical word or a combination of a lexical word with a grammatical morpheme in Korean grammar. Consequently, when two lexical words are aligned as one unit in the first interval tier, the /*word* boundary of the two tiers between the words should be corrected also. Additionally, most of the time, the coda /l/ is ignored, and Korean intervocalic velar lenis stop is transcribed as /q/, and /g/ transcribes initial velar lenis stop by the automatic aligner. The fourth tier is duplicated from the third and manually corrected (tier 4).

In addition to the three tiers of segmentation and annotation, a *prosody* tier and a misc(miscellaneous) tier are created for the prosody annotation(Figure 6). The prosodic tier is completed manually, and the misc tier serves for remarks.



22.680409
0.1434
0
-0.1285
4500 Hz
1061 Hz
0 Hz
450 Hz
40 Hz
1 아마 할아버지 할머니 두 명 chunk (170)
2 아마 할아버지 할머니 두 명 chunk/word (108)
3 ɐ m ɐ h ɐ ɐ b ʌ dʑ i h ɐ l ʌ n i d u j ʌ ŋ chunk/phon (379)
4 flat flat rising rising prosody (12)
5 final vowel lenthening final vowel lenthening final vowel ?? misc (1/12)
4.030289

Figure 6: Example of an alignment after correction of the /*phon* tier and two additional tiers for prosody (F05, L2 Korean)

4. Conclusions and Perspectives

In this paper, we present the experimental design of a parallel L2 learner corpus for the pair of languages Korean/French. We aim at providing raw but controlled parallel material for research in comparative linguistics, (with special attention to syntax, morphology, phonology and prosody), but also in applied linguistics, and bring to teachers, material for building assessments and language evaluation tools. We believe that such a parallel corpus can enable to evaluate the weight and role of the learner L1 as well as the differences and/or similarities between L1 and L2 acquisition. We will contribute at bringing to the scientific community working on Korean and French but also second language acquisition a valuable linguistic resource, and more collaborations to be established.

At the moment, we are finalizing annotation corrections for Korean data, and we are working on the deposition of the linguistic resource.

5. Acknowledgements

This project was initiated by co-funding from University Paris Diderot and Seoul National University, and followed by Labex Empirical Foundations for Linguistics funding.

6. Bibliographical References

Boersma, Paul & Weenink, David (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.33, retrieved 26 September 2017 from http://www.praat.org/

Common European Framework of Reference for Languages: Learning, Teaching, Assessment, Language Policy Unit, Strasbourg (www.coe.int/lang-CEFR)

Delais-Roussarie E. and H. Yoo. (2011) Learner corpora and prosody: from the COREIL corpus to principles on data collection and corpus design. Poznań Studies in Contemporary PSCIL 47: 28-39.

Delais-Roussarie, E. Santiago, F. & Yoo, H. (2015) the extended COREIL corpus: first outcomes and methodological issues, Satellite workshop of ICPhS 2015, August 2015 Glasgow

Detey, S., Durand, J., Laks, B. and Lyche, C. (2016) (éds). Varieties of Spoken French: a source book. Oxford: Oxford University Press.

Frota, S. & Prieto, P. (2015), Intonation in romance, OUP.

Gut, U. (2009). Non-native speech. A corpus-based analysis of phonological and phonetic properties of L2 English and German. Frankfurt: Peter Lang.

Granger, S. (2012). Learner Corpora. In: C.A. Chapelle, The Encyclopedia of Applied Linguistics, Wiley-Blackwell: Oxford, 2012. 978-1-4051-9473-0.

Granger, S. (2003). The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. TESOL 37 (3). 538-546.

Goldman J.-Ph. (2011) *EasyAlign: an automatic phonetic alignment tool under Praat* Proceedings of InterSpeech, September 2011, Firenze, Italy (http://latlcui.unige.ch/phonetique/easyalign.php)

Gut, U. (2009): Non-native speech. A corpus-based analysis of phonological and phonetic properties of L2 English and German. Frankfurt: Peter Lang.Han, M.-H. (2011). Fautes de prononciation des Coréens apprenant le français et correction phonétique. Synergies Corée 2 : 73-82.

Hawkins, J. & Buttery, P. (2009). "Using learner language from corpora to profile levels of proficiency: insights from the English Profile Programme". In Taylor, L. & Weir, C. J. (dir.), Language testing matters: investigating the wider social and educational impact of assessment. Cambridge: Cambridge University Press. pp. 158-175.

Herment, S., Tortel,A., Bigi, B., Loukina, A., Hirst, D. J., Kochanski, G. (2012a). AixOx, a multi-layered learners corpus: automatic annotation, Proceedings of the 4th International Conference on Corpus Linguistics (4 : 2012 mars 21-24 : Jaèn, Spain).

Herment, S., Loukina, A., Tortel, A. (2012b). AixOx. Available on SLDR (Speech Language Data Repository), http://sldr.org/sldr000784/fr

Lüdeling, A., Hirschmann, H. (2015). Error annotation. In: Granger, S., Gilquin, G., Meunier, F.,(eds), The Cambridge Handbook of Learner Corpus Research Cambridge: Cambridge University Press.

Milde & Gut 2002 "A prosodic corpus of non-native speech", In Proceedings of Speech Prosody 2002, Aix-en-Provence, pp.503-506.

Racine, I., Zay, F., Detey, S. and Kawaguchi, Y. (2011), « De la transcription de corpus à l'analyse interphonologique: enjeux méthodologiques en FLE ». In Travaux Linguistiques du CerLiCO 24, Rennes: PUR.

Racine, I., Detey, S., Zay, F. and Y. Kawaguchi (2012). 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.

Santiago, F. & E. Delais-Roussarie. (2012). Acquiring phrasing and intonation in French as a second language: The case of Yes-No questions produced by Mexican Spanish Learners. In Ma, Q., Ding, H. & Hirst, D. [Eds.]. Proceedings of Speech Prosody 2012, Shanghai, China, 338-341

Tortel, A. 2008. ANGLISH : base de données comparative L1 & L2 de l'anglais lu, répété, parlé. TIPA 27: 111-122.

Voormann, H. & U. Gut. 2008. Agile Corpus Creation. Corpus Linguistics and Linguistic Theory 4 (2): 235-251.

7. Resources

CEFLE corpus: http://projekt.ht.lu.se/cefle/information

Frantext project: http://www.cnrtl.fr/corpus/frantext/

IARI project: Prieto, P., Borràs-Comes, J., & Roseano, P. (Coords.) (2010-2014). Interactive Atlas of Romance Intonation. Web page: <http://prosodia.upf.edu/iari/>

Korean National Corpus : http://www.sejong.or.kr/user/main.do

Learner corpora around the world database: http://www.uclouvain.be/en-cecl-lcworld.html

PFC Project : http://www.projet-pfc.net/

TUFS project: http://www.coelang.tufs.ac.jp/multilingual_corpus/fr/index.html?contents_xml=corpus&menulang=en

TCOF project: http://www.cnrtl.fr/corpus/tcof/