Title: Construction of a French-Korean bilingual corpus for research and application purposes

Principal investigators:

Université Paris Diderot	Seoul National University
Hiyon Yoo Department of Linguistics) UMR 7710 Laboratoire de Linguistique Formelle Case 7003, Université Paris Diderot 75205 Paris Cedex 13 Tel : +33 1 57275770 Fax : +33 1 57275781 Mail : yoo@linguist.univ-paris-diderot.fr	Dongyeol PARK Seoul National University College of Education Depart. of French Language Education 599 Gwanak-ro, Gwanak-gu, Seoul, 151-748, Republic of Korea Tel : +82-2-880-7692 Fax : +82-2-875-4884 Mail: bondieu@snu.ac.kr

Involved faculties:
Université Paris-Diderot
UFR Linguistique
UFR LCAO
UFR EILA

Seoul National University
College of Education, department of French Language Education
College of Education, department of Korean Language Education
College of Humanities, department of Linguistics

Summary

The present project proposes to build a bilingual corpus of French learners of Korean and Korean learners of French using the same experimental design. This language resource will contain written and speech data, gathered among learners with different proficiency levels and in different contexts. 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 linguistics, computational linguistics, second language acquisition but also in the field of applied linguistics.

Presentation of the project

In the last years, especially with the development of methodologies in corpus linguistics 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 the linguistic phenomena. Recently, we saw a rise of large corpora for second language acquisition (cf. among others, Granger 2003 and 2012, Hawkins & Buttery 2009). Beside the fact that few corpora are freely available to the research community (see however Milde & Gut 2002, Tortel 2008, Herment et al. 2012, 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 & Buttery 2009), how students' pronunciation of L2 can be influenced by his L1 (I-PFC project, Racine et al., to appear) 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.

One of the main goals of the present project is to propose an answer to this missing linguistic resource by putting in parallel two languages, French and Korean, and by building a bilingual corpus of French learners of Korean and Korean learners of French. This corpus will be gathered following a unique protocol for the two populations and constitute a linguistic resource as complete as possible, with both written and spoken data. Such a resource will represent a large asset in the field of theoretical but also applied linguistics, especially in developing teaching material, establishing tools and research for the education of future teachers of foreign language (pronunciation, communicative and grammatical competence, and so on).
In the following sections, we will present an overview of existing corpora in order to show how such a project can fit in the research community, the goals of the project and a more detailed organization of the project.

Overview of existing second language acquisition corpora for French and Korean

There are today numerous corpora of French as L1, be they written or spoken (the Frantext corpus, TCOF corpus, PCF 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 word, 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. to appear) 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 de say how much the project has advanced.

Aims and ambitions of the presented project

As presented in the previous section, the pair Korean-French is quasi-inexistent in the field of second language acquisition corpora. The only project that puts along the two languages is I-PFC but the latter is focused on French as a second language (and not Korean as a second language) and more specifically on some phonological competence of the learners only. Compared to I-PFC, the present project aims at building a bilingual corpus for the pair of Korean-French where Korean learners of French and French learners of Korean have to perform the same tasks.

Most existing corpora are based on written tasks (CEFLE), reading tasks (the IFCASL Project, Trouvain et al. 2013, I-PFC, Racine et al. to appear, Longdale, Ballier & Martin 2010) or free spoken language (I-PFC, Longdale, rated speech corpus, Yoon et al. 2009), but they are not always suitable when we want to look at a specific linguistic phenomenon. Furthermore, most existing corpora are mono-directional and do not bring along data in both sides, that is when L1 and L2 languages are inversed (the IFCASL Project is one the few bilingual projects that exist, putting in parallel French and German). We believe that such a confrontation (in our case, Korean learners of French and French learners of Korean) can shed light to a better comprehension of the specific interlanguage but also of the two languages.

The aim of this project is two-fold:

- Build a French-Korean bilingual L2 corpus with written and speech data gathered among learners with different proficiency levels and in different contexts (taking account for instance the country where the L2 is acquired). 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 linguistics, computational linguistics and second language acquisition, both for French and Korean as a second language. Some examples are given below:
    - Building computer assisted language learning tools adapted to the learners population,
    - Building teaching materials (courses and exercises) taking account the particularities of each population,
    - Providing learners with individualized feedback,

    - 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, language evaluation tools, and so on.

- Following the domains of interests of the participants of the project, specific research on the acquisition of some linguistic phenomena, especially by observing and ranking recurring errors made by learners, both in written and spoken material. A special attention will be given at:
    - Comparative linguistics between Korean and French
    - The phonological and phonetic systems of French and Korean learners
    - The realization of prosodic patterns
    - Research on the specific interlanguage that is at stake when comparing the two languages.

The present project involves researchers of different departments (UFRL, EILA, LCAO in Paris Diderot, department of French Education and department of Linguistics in SNU) and represents a good opportunity to establish further projects in the field of linguistic and language pedagogy. Moreover, this project can lead to future collaborations between French and Korean research laboratories in linguistics, especially in the domain of second language teaching (similar undergraduate and graduate programs can be found in both universities).

Organisation of data collection

The project can enjoy benefits from the fact that both languages (French and Korean) are taught as a second language in both universities, providing potential participants with defined proficiency levels. For instance, data gathering can be done at the end of each semester, which allows the determination of the acquired proficiency (following the different syllabus at each university). Moreover, the fact that both languages are taught as a second language in both universities allows taking into account complementary parameters such as:

- L2 learned in immersion or not
- Proficiency level

Using online exercises or exams that can be built for the two languages and used in both universities can also facilitate data collection.

For the moment, we aim at gathering 40 L2 learners per language (10 by proficiency level determined by their year of study) and 10 learners in their L1 (A total of 80 speakers). It is possible to have a longitudinal collection, where some learners will participate to the data during their whole education.

The first step of the project (we hope during the workshop in October 2015) will be to determine together the tasks that will constitute the corpus. The fact that we use a task-based corpus makes the protocol adaptable and modular, following proficiency level, learning situations. For written tasks, we will think about the different tasks than can be performed by the students.

For spoken tasks, we could lay on protocols such as COREIL (Delais-Roussarie & Yoo 2011, inspired by the AGILE corpus, Voormann, H. & U. Gut. 2008, Gut 2009), including

reading of short texts, discourse completion tasks (Blum-Kulka et al. 1989) study questionnaire where subjects have to give the proper answer to a given communicative situation (see for instance the IARI project, Prieto et al. 2014), a maptask, a discussion between at least three persons.

The different points developed during the workshop in October will hopefully lead to data collection as the end of each semester. The second workshop that will be held in Paris in June 2016 will be the opportunity for the involved researchers to make a point on the advancement of data gathering, to see how

Such a bilingual corpus can enables 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 an French but also second language acquisition a valuable linguistic resource, and more collaborations to be established.

External funding

The project can be part of the existing LABEX Empirical Foundations for Linguistics (http://www.labex-efl.org/?q=fr/accueil). Axe 6 (language ressources) can provide complementary funding as soon as the project begins, especially for logistic assistance but also for financing students or researchers during fieldwork (data collection). Following the issues that will be studied, funding can be asked among axes 1 (phonetics and phonology) and 3 (linguistic typology).

We also plan to ask for complementary funding through the STAR program (French-Korean partnership Hubert Curien) at the National Research Foundation of Korea and at the Ministry of Education in France (MAEDI et MENESR) (http://www.campusfrance.org/fr/star).

Finally, following the evolution of the project, we plan to deposit a more consequent project at the ANR (Agence National de la Recherche (ANR) and at the National Research Foundation of Korea (NRF). The project aims at gathering researchers working in this field beyond the collaboration between the two universities.

Selected references

Ballier N. & Martin Ph, “The Charles V phonetically annotated corpus: a research project at the University of Paris-Diderot”, la Societas Lingvistica Europaea, 43rd Annual Meeting, Vilnius, 2-5 September 2010

Blum-Kulka, Shoshanna; House, Juliane, and Kasper, Gabriele. (1989). ‘Investigating cross-cultural pragmatics: An introductory overview’, in S. Blum-Kulka, J. House & G. Kasper (eds), Cross-cultural pragmatics: Requests and apologies (pp. 1-34). Norwood, NJ: Ablex, p. 13-14.

COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES: LEARNING, TEACHING, ASSESSMENT, Language Policy Unit, Strasbourg (www.coe.int/lang-CEFR)

Delais-Roussarie E. & 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.

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

Han, Mun Hi (2011). Fautes de prononciation des Coréens apprenant le français et correction phonétique. Synergies Corée 2 : 73-82.

Granger, Sylviane. 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.
Gut, U. (2009): Non-native speech. A corpus-based analysis of phonological and phonetic properties of L2 English and German. Frankfurt: Peter Lang.
Hawkins, J.A. & P. Buttery. 2009. Using learner language from corpora to profile levels of proficiency. Studies in Language Testing. Cambridge: Cambridge University Press.
Herment ; S. ; Tortel,A. ; Bigi,B. ; Loukina ,A. ; Hirst,D.J. ; Kochanski, G. (2012). 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. (2012). AixOx. Available on SLDR (Speech Language Data Repository), http://sldr.org/sldr000784/fr
Racine, I., Detey, S., Zay, F. & Y. Kawaguchi (to appear). 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. (to appear), « De la transcription de corpus à l'analyse interphonologique: enjeux méthodologiques en FLE ». In Travaux Linguistiques du CerLiCO 24, Rennes: PUR.
Sugiyama, K. (2012). Lexical Profile of French Learner Speech: The Case of Japanese University Students. In Tono, Y., Kawaguchi, Y. & Minegishi, M. (éds.), Developmental and Cross linguistic Perspectives in Learner Corpus Research Amsterdam/Philadelphia, John Benjamins.
Tortel, A. 2008. ANGLISH : base de données comparative L1 & L2 de l'anglais lu, répété, parlé. TIPA 27 : 111-122.
Trouvain, J., Laprie, Y., Möbius, B., Andreeva, B., Bonneau, A., Colotte, V., Fauth, C., Fohr, D., Jouvet, D., Mella, O., Jügler, J. & Zimmerer, F. 2013. Designing a bilingual speech corpus for French and German language learners. Proc. Corpus et Outils en Linguistique, Langues et Parole: Statuts, Usages et Mésuages, Strasbourg, pp. 32-34.
Voormann, H. & U. Gut. 2008. Agile Corpus Creation. Corpus Linguistics and Linguistic Theory 4 (2) : 235-251.
Yoon S..-Y., Pierce L., Huensch A., Juul E., Perkins S., Sproatt R. & Hasegawa-Johnson M. (2009). Construction of a rated speech corpus of L2llearners' speech. CALICO Journal 26(3):662-673.

Online Ressources:

IARI project: Prieto, P., Borràs-Comes, J., & Roseano, P. (Coords.) (2010-2014). Interactive Atlas of Romance Intonation. Web page: <http://prosodia.upf.edu/iari/>
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/
Frantext project: http://www.cnrtl.fr/corpus/frantext/
Longdale project:
http://www.univ-paris-diderot.fr/EtudesAnglophones/pg.php?bc=CHVR&page=ProjetsencoursCLILLAC-ARP1&g=sm

Anticipated budget and format for the workshops:

The project gathers researches from different departments and faculties from each university. The workshop will be the opportunity for the main investigators from the two universities to meet and to establish the roadmap of the project. The workshop can be held during three days in the format of roundtables, around the themes we want to develop in the project.
Some of the themes could be:

- Presentation of the different departments,
- Presentation of existing corpora for French and Korean and presentation of the present project
- Discussion and Establishment of the protocol for written and tasks
- Discussion on annotation and used tools for gathering data

There will be also discussion on the funding for setting up the project. In this section, we detailed the evaluated budget for the workshops only, plus a small budget in order to set up the project between the two workshops. An additional budget for the project itself can be provided if needed.

Expected costs by categories

Workshop to be held in Paris

Travel expenses	Travel and living expenses for the Korean researches	1000€*3	3 000
Living expenses	1 week living expenses for the researches	100*7*3	2 100
Logistic assistance for the workshop hosting	Coffee breaks, lunchs, logistics		500
Sub-total			5 600

Workshop to be held in Seoul

Travel expenses	Travel and living expenses for 4-5 researches	1000€*4	4 000
Living expenses	1 week living expenses for 4-5 researches	100*7*4	2 800
Logistic assistance for the workshop hosting	Coffee breaks, lunchs, logistics		500
Sub-total			7 300

Setting the beginning of the project

Human resources	Payment of students during data collection		4 000 (2 000 per university)
Material	Recording material		4 000 (2 000 per university)

Total			**22 900**