# The extended COREIL corpus: first outcomes and methodological issues

Elisabeth Delais-Roussarie[1], Fabián Santiago[1,2] & Hi-Yon Yoo[1]

[1]UMR 7110-LLF, Université Paris Diderot, [2] UMR 7018-LPP, Université Sorbonne Nouvelle

elisabeth.roussarie@wanadoo.fr, fabian.santiago.ling@gmail.com, yoo@linguist.jussieu.fr

**Keywords:** L2 prosody, recording protocol, subject selection, intonation

## 1. INTRODUCTION

Since several years, numerous researches on second language acquisition have been based on corpus studies (cf. [1] and [2] among others), allowing a better evaluation of possible correlations between learner's L1, grammatical competence and proficiency level in L2. Among the studies focusing on the acquisition of L2 phonology, phonetics and prosody, a distinction can be made according to the way of collecting and analysing the data. Acoustic data, for instance, may be gathered by means of experimental procedures or may be extracted from larger corpora [3]. Experimental approaches have the advantage of allowing the control of various elements that may come into play in the production process and to focus on the targeted structures, but the obtained data may not always be a good sample of the learner's proficiency and may be limited. Corpus-based approaches, by contrast, allow gathering a large data set that may provide better insights on the L2 phonological /phonetic acquisition process (cf. [4], [5] and [6]). Even if the structure and content of the corpus used may vary greatly, corpus-based studies display several advantages in L2 research:

- Analysing a large amount of data allows investigating the interaction of several factors at the same time (for instance, the relation between proficiency level and distribution of tonal patterns, the relation between syntactic complexity and prosodic phrasing, etc.).
- It is possible to put in perspective several explanatory factors affecting the L2 acquisition process and competence (be they linguistic or not), e.g.: L1 transfer, speech style (reading vs. spontaneous speech), age when learners start acquiring the L2, etc.
- When the corpus includes data from a large array of languages and from various speaking styles, it is more representative and allows making generalisations on the acquisition of a foreign language. In addition, cross-comparisons between language pairs are possible and allow identifying the factors motivating L2 prosodic/ phonological 'errors'.

By taking all these issues, the COREIL corpus, a large learner speech database designed for studying the L2 prosody, was developed [6]. After presenting the principles at play in designing the data collection and annotation protocols, we will (i) describe the data that are currently compiled in the extended version of the COREIL corpus, (ii) briefly present some results obtained, and (iii) discuss some remaining methodological issues that must to be improved in order to successfully share these resources with the linguistic community.

## 2. THE EXTENDED COREIL CORPUS

### 2.1 Basic principles and facts

The protocols used to gather and annotate the data from the COREIL corpus were thought in such a way as to (i) avoid making strong presuppositions (such as the idea that L1 transfer is crucial); (ii) allow making contrastive analysis between learners' and natives' oral productions with comparable data sets; (iii) allow evaluating the learner oral competence and the L2 proficiency level while taking into account a large array of tasks/ skills (reading speech, monologual and interactive speech); and (iv) recording speakers with different L1.

### 2.2 Recording protocol and tasks

In order to obtain different speaking styles, the speakers were recorded while performing five distinct tasks that are classified into three groups. The first group includes two interactive oral production tasks (IOP). In one of them, the speakers were interviewed (they were asked to talk about their projects, their experience in French courses, etc.), while, in the second, they had to perform a role-play, in which they asked questions to complete an enrolment form. The second group consisted of two monologue oral production tasks (MOP), including first the description of a painting and second the narration/narration of a picture representing a group of people involved in an activity. The third group consisted in a reading task (RT), in which the speakers had to read short dialogues and several texts adapted from the

EUROM 1 corpus (cf., for more details, [5]). All participants were asked to read the texts and dialogues several times before the recording session.

The recordings took place in a quiet room and were done with an Edirol R09 digital recorder. The questions used in the current study were extracted from two types of tasks: IOP and RT.

### 2.3 Participants

Two groups of participants are distinguished in our corpus: the native speakers (or control groups) and the learners (experimental groups). In control groups speakers were recorded in their L1, whereas learners were recorded in the L2 target language. The COREIL corpus was designed to collect data in L2 French and English produced by speakers with English and Mandarin as L1. But in its current state, data from the following language background have been recorded, transcribed and are thus available: native spoken data from French, Mexican Spanish, German, Korean and Greek speakers, French L2 data from Mexican Spanish, German, Korean and Greek learners and French learners of Korean. As for the age of learners, no constraint was imposed, but the L2 was always acquired when the speakers had reach at least 10 years of age. Table 1 summarizes all these information. The proficiency level of learners was encoded according to the Common European Framework of Reference for Languages (CEFR).

Table 1. *Speakers' profile*

| Group | L1 | Participants | Level |
|---|---|---|---|
| L2 French | German | 8 | A2&B1 |
| | Greek | 4 | A1, A2 & B1 |
| | Korean | 3 | A1 & A2 |
| | Spanish | 15 | A2&B1 |
| L1 French | French | 10 | Native |
| L1 German | German | 8 | Native |
| L1 Greek | Greek | 2 | Native |
| L1 Korean | Korean | 2 | Native |
| L1 Spanish | Spanish | 15 | Native |

## 3. USE OF THE COREIL CORPUS

Up to know, the COREIL corpus has been used to develop or evaluate annotation tools and systems, but also to study specific intonational patterns.

Since most prosodic annotation systems have been developed at the phonological levels, it was important to evaluate how they could be used to encode learner prosody in oral data [7]. In addition, some of this data has been used to develop PROSOTRAN, a transcription tools that relies on the prosodic parameters at the syllabic level to provide a symbolic transcription [8].

Studies on the acquisition of intonation in an L2 have also been achieved with data extracted from the COREIL corpus. In studying and comparing the tonal configurations realized at the end of neutral yes/no questions and wh-questions in native French, Spanish and L2 French, [9] showed that some L2 intonational patterns, in particular high rises at the end of questions, cannot be attributed to an L1 transfer but rather to the effects of iconic universal tonal representations. In the same vein, [10] showed that German and Spanish learners of L2 French realize high rises at the end of non-final IP at early stages of the acquisition process, even if such form doesn't occur in their L1 nor in native French. This tonal pattern may be interpreted as a sign of linguistic insecurity.

## 4. CONCLUSION AND PERSPECTIVES FOR IMPROVEMENT

The COREIL Corpus had several features that allow getting interesting insights on the acquisition of intonation in an L2. As it includes data from different styles recorded by speakers with various proficiency levels, the results obtained are probably less disputable. The protocol used allows comparing productions from speakers with different L1, which offer promising results. Yet, the results of the studies achieved are still limited, several methodological issues requesting improvements:

- the protocol employed for eliciting questions does not allow gathering a large set of different question types (coordinated questions, echo-questions, etc.)
- the evaluation of the L2 proficiency levels is sometime problematic, as the CEFR does not really take into account suprasegmental features in the descriptors.

In order to improve these issues, we wish (i) to add to the protocol more interactive oral tasks for eliciting questions, and (ii) to develop tests for better identifying the proficiency level of the learners regarding prosody.

In order to reinforce some of the results obtained by comparing the productions of learners with different L1, it would be crucial to share the protocol for gathering a larger amount of data that can be compared.

## AKNOWLEDGEMENTS

This work was supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083). It also benefited from a doctoral grant from CONACyT (Mexico).

## REFERENCES

[1] 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.

[2] Hawkins J.A. & P. Buttery. (2009). Using learner language from corpora to profile levels of proficiency. *Studies in Language Testing*. Cambridge University Press.

[3] Delais-Roussarie E. & H. Yoo (2014) Corpus and research in phonetics and phonology (I): Methodological and formal considerations. In Durand J., Gut U. & G. Kristoffersen (eds), *Handbook of Corpus Phonology*. Oxford University Press

[4] Gut U. (2009). *Non-native Speech. A corpus-based analysis of the phonetic and phonological properties of L2 English and L2 German.* Peter Lang.

[5] Herment S., A. Tortel, B. Bigi, D. Hirst & A. Loukina (to appear) AixOx, a multi-layered learners' corpus: automatic annotation. In Díaz Pérez, J. & A. Díaz Negrillo (eds.), *Specialisation and variation in language corpora*. Peter Lang.

[6] 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* 471: 28-39.

[7] Delais-Roussarie, E & H. Yoo (2011). Transcrire la prosodie : un préalable à l'échange et à l'analyse de données. *Journal of French Language Studies 21/1* : 17-27.

[8] Bartkova K., Delais-Roussarie E. & F. Santiago-Vargas (2012). PROSOTRAN : a tool to annotate prosodically non-standard data. *Proceedings of Speech Prosody 2012*.

[9] Santiago, F. & E. Delais-Roussarie. (2015). The acquisition of question intonation by Mexican Spanish Learners of French. In Delais-Roussarie, E., M. Avanzi & S. Herment (eds.), *Prosody and languages in contact: L2 acquisition, attrition, languages in multilingual situations*. Springer Verlag.

[10] Santiago, F. & E. Delais-Roussarie. (2015). What motivates extra-rising patterns in L2 French: Acquisition Factors or L1 Transfer?, *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*.