Oral lexical richness in French and Italian L2: a corpus-based perspective

Anna GUDMUNDSON (Stockholm University)
Camilla BARDEL (Stockholm University)
Christina LINDQVIST (Uppsala University)

This paper presents a series of corpus based studies on the lexical richness of Swedish learners of French and Italian as second languages (L2), as represented in oral production. The studies have been carried out within a research project concerned with the possibility of L2 speakers attaining native-like proficiency in the domain of vocabulary. Our overall aim is to develop a method for lexical profiling suitable for oral production data, which is able to distinguish between learners at different proficiency levels.

There is at present time a considerable interest in L2 vocabulary acquisition and the research area covers different aspects of the learner's lexicon, such as for instance cross-linguistic influences, lexical size, width and depth (cf. e.g. Bogaards and Laufer 2004; Daller et al. 2007; Milton et al. 2008). Previous studies on lexical richness in L2 mainly focus on written production, and our on-going work contributes to the field by providing empirical data based on oral production.

The method used is inspired by the lexical frequency profile (LFP) proposed and developed by Laufer and Nation (1995). The idea behind this method is that the use of a high proportion of low-frequency words in a learner's production indicates a rich or advanced vocabulary; a lexical profiling analysis is carried out by comparing the production of learners with that of native speakers of the target language, and concerns the proportions of words belonging to different frequency bands, traditionally based on written target language corpora. In Lindqvist (2010) the French version of LFP, vocabprofil, (Cobb and Horst 2004) was used to make lexical frequency profiles of spoken L2 data, but the results showed no significant differences between learners at different proficiency levels. In a qualitative analysis Lindqvist (2010) found that many of the words classified as rare by vocabprofil were quite frequent in oral production, and she pointed out that it might be more suitable to compare oral data with frequency bands based on oral data.

In our present project, we have created new frequency bands derived from spoken data, as opposed to the original version of LFP as well as its French version. The French frequency bands are based on the Corpaix corpus which has been compiled at the Université de Provence (Campione et al. 2005), and the Italian frequency bands are based on the lemmatized versions of the LIP corpus, originally created by De Mauro et al. (1993), developed at Karl-Franzens-Universität in Graz, and the Italian C-oral-rom corpus (http://lablita.dit.unifi.it/coralrom/lablita_corpus.html). The frequency bands are stored in an SQL data base (Structured Query Language), a suitable format when working with large amount of data.

In order to evaluate if a profiler based on the oral frequency bands is able to distinguish between learners at different proficiency levels, data from French and Italian oral L2 corpora compiled at Stockholm university was used. On the basis of morpho-syntactic criteria, such as VP and NP morphology, 10 intermediate learners and 10 advanced learners were chosen from the InterIta corpus (www.fraitaklass.su.se/interita), and 7 advanced low learners and 7 advanced high learners from the InterFra corpus (www.fraitaklass.su.se/interfra). The French learner data is the same as the one used in the study by Lindqvist (2010), and this enables us to evaluate the effect of using frequency bands based on oral corpora instead of written. We also used two control groups of native speakers to be able to study the possibility of L2 speakers attaining native-like proficiency in the domain of vocabulary.

The data of each learner group was compared to the frequency bands and the proportion of low frequency words was calculated. A one way ANOVA with post hoc tests showed significant differences between learner groups for both French and Italian ($p < 0.05$), i.e. the advanced learners in the Italian corpus display a significantly higher proportion of low frequency words when compared to the intermediate learners, and the French advanced high learners display a significantly higher proportion of low frequency words when compared to the advanced low learners. Both groups of Italian learners are significantly different from the native control group, showing a lower amount of low frequency words, but when it comes to the French data, there is no significant difference between the advanced high learners and the French control group. These results indicate that when making lexical frequency profiles of oral data it is better, or maybe even necessary, to compare it with frequency bands based on oral, and not written, data. Only when running the data from the Lindqvist (2010) study in the oral lexical frequency profiler, the LFP method was able to distinguish between learners at different proficiency levels and the method can be validated. The results also suggest that very advanced learners can reach the same level as native speakers in terms of lexical richness, as defined by low-frequency, in oral production, which is the case for the French advanced high learners.

Bogaards, P. & Laufer, B. (2004). Vocabulary in a second language: Selection, acquisition, and testing. Amsterdam: John Benjamins.

Campione, E., Véronis, J. & Deulofeu, J. (2005). 3. The French Corpus. In E. Cresti & M. Moneglia (eds.), C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages. Amsterdam: John Benjamins, 111-133.

Cobb, T. & Horst, M. (2004). Is there room for an academic wordlist in French? In P. Bogaards & B. Laufer (eds.), 15-38.

Daller, H., Milton, J. & Treffers-Daller, J. (2007). Modelling and assessing vocabulary knowledge. Cambridge: Cambridge University Press.

De Mauro, T., Mancini, F., Vedovelli, M., & Voghera, M. (1993). Lessico di frequenza dell'italiano parlato. Milano: Etaslibri.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. Applied Linguistics 16(3): 307-322.

Lindqvist, C. (2010). La richesse lexicale dans la production orale de l'apprenant avance de francais.

La Revue Canadienne des Langues Vivantes/The Canadian Modern Language Review 66(3): 393-420.

Milton, J., Daller, H., Malvern, D., Meara, P., Richards, B. & Treffers-Daller, J. (2008). Vocabulary. [Special issue]. Language Learning Journal 36(2).