LINDSEI-BR: an oral English interlanguage corpus

Heliana MELLO (UFMG)
Luciana ÁVILA (UFMG)
Barbara Malveira ORFANÒ (UFSJ)
Tufi NEDER NETO (UFLA)

Corpus Linguistics has been more than instrumental in the study of interlanguage. It has made it possible for researchers not only to have access to large quantities of varied interlanguage samples but also process these data both for individual language features as well as for a host of other elements, such as interlanguage features at a given acquisition stage, comparative error analysis, among others. Presently there are many interlanguage corpora available to researchers and teachers, both written and oral, and this has afforded a spurt in interesting findings as far as the manyfold processes involved in language acquisition are concerned.

In this paper, we will present a new English interlanguage corpus under compilation in Brazil. It is associated with a larger project - the COBAI; the Brazilian Oral Corpus of Learner English is a repository of spoken interlanguage data and aims to gather varied subcorpora of Brazilian learner English with the main purpose of providing data for the study of interlanguage features within the frame of second language acquisition research. The project was launched in 2011 and so far it is concerned with the compilation of the LINDSEI-Brazil component which will be presented in this paper.

The Louvain International Database of Spoken English Interlanguage (LINDSEI) project is an international initiative coordinated at the Centre for English Corpus Linguistics, at the Université Catholique de Louvain (cf. Gilquin, De Cock, Granger, 2010). The LINDSEI project encompasses seventeen different interlanguage subcorpora, compiled with the same parameters and transcribed following the same guidelines. The LINDSEI project is the oral counterpart for the ICLE – International Corpus of Learner English, compiled by the same team of researchers under the direction of Sylviane Granger (cf. Granger, 2003 and Granger et al. 2009).

The LINDSEI-BR is being compiled following the international project guidelines. At present we have achieved our recording goal of fifty recordings and their transcription is underway. The recording informants were university, high intermediate to advanced level students of English as a second language. The recordings covered three different tasks: a narrative about a chosen set topic by the informant, free discussion with the interviewer and the description of a pictured scene. Each recording is on average twenty minutes long and features quasi-spontaneous speech patterns. For each recording there is an accompanying learner profile that covers the learner's language history and other elements that might have contributed to her/his process of language acquisition, besides having information about the interviewer and the actual interview itself. The transcription guidelines include a code for each recording, speakers' turns, and the marking of several speech features, such as: overlapping, pauses, backchannelling, contractions, truncation, among others. An excerpt of BR001 follows below as illustration:

<A> yeah everybody is polite (em) people. I I don't know they are very different from us but. they are very polite and not so cold as people say that oh they are very cold no they're. wonderful. people.. (eh) what else let me see </A>
<B> but you… weren't you in Lisbon </B>

After the transcription process is concluded and revisions and transcription validation are undertaken, we will proceed to the PoS tagging of the corpus and will also generate specific

statistics that describe the corpus participants and the corpus itself, as far as number of words, turns, and specific speech phenomena are concerned. A second stage in this corpus treatment process will be to create a mark up system that specifies interlanguage features in the LINDSEI-BR, in order to make it possible to carry searches based on linguistic features alone and cross them over with other parameters of corpus.

Although the LINDSEI-BR project is still on the making, we believe that when it is concluded it will be a valuable resource for both researchers and teachers of English as a second language in Brazil.

References

GILQUIN, Gaëtanelle; De COCK, Sylvie; GRANGER, Sylviane. The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM. Presses universitaires de Louvain, Louvain-la-Neuve. 2010.

GRANGER, Sylviane. The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. TESOL Quarterly, 37(3), 538-546. 2003.

GRANGER, Sylviane; DAGNEAUX, Estelle; MEUNIER, Fanny; PAQUOT, Magali. The International Corpus of Learner English (Version 2). Handbook and CD-ROM. Presses universitaires de Louvain, Louvain-la-Neuve. 2009.