Oral Expression in Spanish as a Foreign Language: Interlanguage and Corpus-Based Error Analysis

La expresión oral en español como lengua extranjera: interlengua y análisis de errores basado en corpus

Leonardo Campillos Llanos
Laboratorio de Lingüística Informática. Universidad Autónoma de Madrid.
Facultad de Filosofía y Letras. Dpto. de Lingüística y Lenguas modernas
c\ Francisco Tomás y Valiente, 1. Campus de Cantoblanco. Madrid 28049
leonardo.campillos@uam.es / leonardo.campillos@gmail.com

Abstract: PhD Thesis written by Leonardo Campillos Llanos under the supervision of Dr. Antonio Moreno Sandoval and Dr. Paula Gozalo Gómez (Universidad Autónoma de Madrid). The thesis was defended on December 17th, 2012, at the Facultad de Filosofía y Letras (Universidad Autónoma de Madrid), and the committee consisted of: Dr. Francisco Marcos Marín (University of Texas at San Antonio), Dr. Joaquín Garrido (Universidad Complutense de Madrid), Dr. Sonsoles Fernández López (Escuela Oficial de Idiomas), Dr. Isabel García Parejo (Universidad Complutense de Madrid), and Dr. Ana Serradilla (Universidad Autónoma de Madrid). The PhD thesis was awarded Summa cum laude (International Doctorate).
Keywords: corpus linguistics, Spanish as a Foreign Language, oral expression, error analysis, interlanguage

Resumen: Tesis realizada por Leonardo Campillos Llanos y dirigida por los doctores Antonio Moreno Sandoval y Paula Gozalo Gómez (Universidad Autónoma de Madrid). Fue defendida el 17 de diciembre del 2012 en la Facultad de Filosofía y Letras (Universidad Autónoma de Madrid) ante un tribunal formado por los doctores Francisco Marcos Marín (University of Texas at San Antonio), Joaquín Garrido (Universidad Complutense de Madrid), Sonsoles Fernández López (Escuela Oficial de Idiomas), Isabel García Parejo (Universidad Complutense de Madrid), y Ana Serradilla (Universidad Autónoma de Madrid). La tesis obtuvo la calificación de Sobresaliente cum laude y la mención de doctorado internacional.
Palabras clave: lingüística de corpus, español como lengua extranjera, expresión oral, análisis de errores, interlengua

1 Aims and motivations

The thesis studied the development of oral expression (oral interlanguage) of forty (N=40) university learners of Spanish as a Foreign Language from A2 and B1 levels (Council of Europe, 2001). The data collection method was a semi-spontaneous interview, with four participants for each mother tongue (henceforth, L1) to obtain comparable data. Learners' L1s were Portuguese, Italian, French, Germanic, English, Dutch, German, Polish, Chinese, and Japanese. With a view to comparing the data, four native speakers in a similar age range were interviewed with the same procedure (control group). Thus, the production of oral errors, vocabulary, and part-of-speech (from here on, PoS) categories could be compared.

The study is based on the methodology of Corpus Linguistics, Second Language Acquisition research, Second Language Teaching, and Error Analysis. The interviews were recorded, manually transcribed, and computationally processed. Then, learners' errors were analysed.

A motivation of this thesis was the need of research on the acquisition of Spanish at A2 and B1 levels. In addition, this project collected a corpus of non-native speech. This data bank may be useful to design or train a speech recognition system for assessing pronunciation or practising oral expression. Results of the analyses may also be taken into account in the design of an automatic error corrector for an
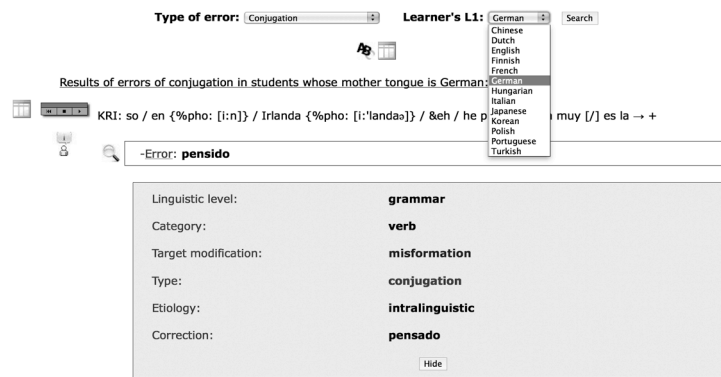
Figure 1: Corpus interface

Intelligent Computer Assisted Language Learning (ICALL) system.

2 Structure of the thesis

The first part introduces the theoretical and methodological framework: background and research approach (Chapter 1), corpus design, criteria for error correction and classification (Chapter 2), and methodology (Chapter 3).

The second part explains both the interlanguage analysis and the error analysis. Firstly, Chapter 4 analyses the formal features of non-native speech (word count, lexical richness, and most frequent vocabulary and categories). To achieve this, the corpus was PoS tagged with the GRAMPAL tagger (Moreno and Guirao, 2006). Afterwards, this tagging was manually verified to correct ambiguities and follow the criteria established according to our definition of lexical unit.

Secondly, the thesis addresses the error analysis of the oral production at different linguistic levels: lexis-semantics (Chapter 5), grammar (Chapter 6), pronunciation (Chapter 7), and pragmatics-discourse (Chapter 8). For this goal, an error typology was designed, the corpus was manually annotated with XML error tags, and a search interface was developed.

3 Contributions

One of the results was the collection of a Spanish learner language oral corpus. The search interface (Figure 1) is available at: http://cartago.lllf.uam.es/corele/index.html. The thesis also presented the results obtained from the corpus data (interlanguage and error analysis). The analyses discussed the evidence of learners' improvement in proficiency level, or the degree of influence of the mother tongue.

Additionally, the thesis included a list of the most frequent lemmas in non-native speech and a set of tables that outline common errors and difficulties in our data for each learner group.

4 References

Council of Europe. 2001. *Common European Framework of Reference for Languages.* Cambridge: Cambridge University Press.

Moreno Sandoval, A., and J. M. Guirao. 2006. Morphosyntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation. In Y. Kawaguchi, S. Zaima, and T. Takagaki (eds.) *Spoken Language Corpus and Linguistic Informatics*, pp. 199-218. Amsterdam: John Benjamins.

List of publications from the thesis

Campillos Llanos, L. 2011. A XML-tagged Spanish learner oral corpus for learner language research. *Learner Corpus Research 2011.* (Poster) 15-17 September 2011. Louvain-la-Neuve, Belgium.

Campillos Llanos, L. 2012. Designing a search interface for a Spanish learner spoken corpus: the end-user's evaluation. In N. Calzolari et al. (eds.) *Proc. of LREC 2012.* 23-25 May 2012. Istanbul, Turkey.

Campillos Llanos, L. 2013a. Oral expression in Spanish by low-intermediate learners: a computer-aided error analysis. *Learner Corpus Research Conference 2013.* 27-29 September 2013. Bergen, Norway.

Campillos Llanos, L. 2013b. Análisis de la producción y de errores en un corpus oral de español como lengua extranjera. *Revista Iberoamericana de Lingüística,* 8, pp. 5-43.