An Exploration into Word Order in Learner Corpora: The WOSLAC Project

Gema Chocano,[1] Rocío Jiménez,[1] Cristóbal Lozano,[2]
Amaya Mendikoetxea,[1] Susana Murcia,[1] Michael O'Donnell,[1]
Paul Rollinson[1] and Iván Teomiro[1]

1. Introduction

This paper reports on work in progress under the framework of a research project[3] investigating the acquisition of word order in Second Language Acquisition (SLA), based on two written learner corpora of L2 English and L2 Spanish. We will discuss (i) the motivation and objectives of the project (ii) data collection (iii) query software and (iv) data analysis. The purpose of the three-year project is to determine the properties which constrain word order in the interlanguage of L2 learners of English (with L1 Spanish) and L2 learners of Spanish (with L1 English). We examine both **lexicon-syntax** and s**yntax-discourse** properties in the analysis of non-canonical word order structures in learner English and learner Spanish (right-periphery of the clause, left-periphery and 'special' constructions: passives, subject inversion and so on).

Word order in English and Spanish differs significantly: in English word order is often said to be 'fixed', while Spanish allows for what is often referred to as 'free order'. In languages with free word order, information structure properties and discourse properties in general play a crucial role in the position occupied by constituents in sentences. The two languages differ in the devices they employ to order constituents in the sentence. An in-depth investigation into word order in advanced learners of L2 English and L2 Spanish will thus offer answers to questions regarding the relative difficulty of acquiring lexical-syntactic and syntactic-discursive properties, as well as general issues related to L1 transfer and the occurrence of constructions which cannot be attributed to the L1 nor to the target language. Some of these issues have been explored in a preliminary analysis on the production of postverbal subjects in learner English (see Lozano & Mendikoetxea, forthcoming).

2. Data collection: learner corpora

Learner corpora are an invaluable tool to explore these issues. Our target is for our two corpora to contain 1,000,000 words at the end of the three-year project . Data collection for WriCLE (Written Corpus of Learner English) began in October 2005.

[1] Universidad Autónoma de Madrid
e-mail: amaya.mendikoetxea@uam.es [project leader]
[2] Universidad de Granada
[3] Funded by Research grant HUM2005-01728/FILO (*The lexicon-syntax and discourse-syntax interfaces: Syntactic and pragmatic factors in the acquisition of L2 English and L2 Spanish*) from the Spanish Ministry of Education) and Research Grant 09/SHD/016 (*The acquisition of word order in English and Spanish as second languages (L2): syntactic and pragmatic factors)*, from the Autonomous Region of Madrid and UAM. Both grants are gratefully acknowledged.

To date (July 2007) we have around 400 academic essays (approx. 600,000 words) in electronic format ranging from around 500 words up to 2,000 words, written by first year and third year Spanish students in an academic writing course on a degree in English Philology at the Universidad Autónoma de Madrid. The texts have been normalised, stripping out references, quotations, footnotes/endnotes, translations and bibliographies, leaving simply the raw text for analysis.

The basic procedure for gathering the data is as follows. Each writer is given a code number, and asked to fill out a *Learner Profile* sheet, giving some basic information about his or her English language background and proficiency in other languages. With each essay (up to three per student) the writer includes a completed *Essay Profile* sheet, providing details about the writing of the essay itself: what kinds of reference tools were used, the time spent on the essay, and so on. Additionally, students fill in consent forms giving their permission for their writing to be used for the purposes of research. Finally, every student completes the Oxford Quick Placement Test, so that we have a standardised measure of general proficiency in English. All this information will be put in a database which will allow the selection of texts according to certain criteria to carry out different types of studies.

As for the corpus of Spanish texts written by English speakers (CEDEL2, Corpus Escrito del Español como L2), it contains at the moment around 250,000 words). Collection for this essay is done online, as texts are written by university students all around the world. Learner data is collected and participants have to complete a Spanish Placement Test (see http://www.uam.es/proyectosinv/woslac/cedel2.htm).

3. Data analysis: software

In regards to analysing the corpus, we will use software for text annotation, *UAM CorpusTool* (Michael O'Donnell) and is freely available (downloadable from http://www.wagsoft.com/CorpusTool/index.html) The tool allows a analyst to select a text from the corpus, and annotate it in various ways. For instance, presented with a student essay, the analyst can highlight a segment (e.g., an *it*-cleft) and then assign features to that segment. The tool produces an XML-encoded version of the text file, including the features assigned to the segments. The annotation scheme for the WOSLAC project includes, among other things, the type of word order phenomena we are interested in analysing. The tool selects a segment, typically a clause, and provides different possibilities for the analysis. For, in Figure 1, the tool provides the possibility of annotating the highlighted segment as a passive sentence:

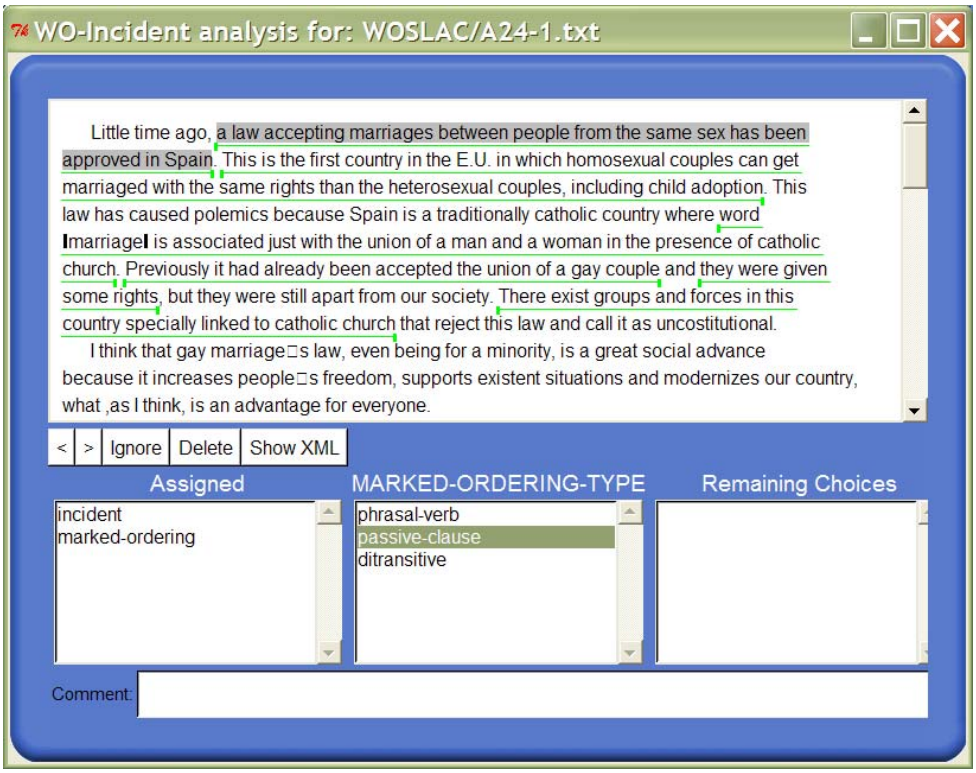

Figure 1: CorpusTool annotation

The tool allows for searches across different levels. Figure 2, for instance, shows subjects in passive sentence structures in two essays of the corpus:

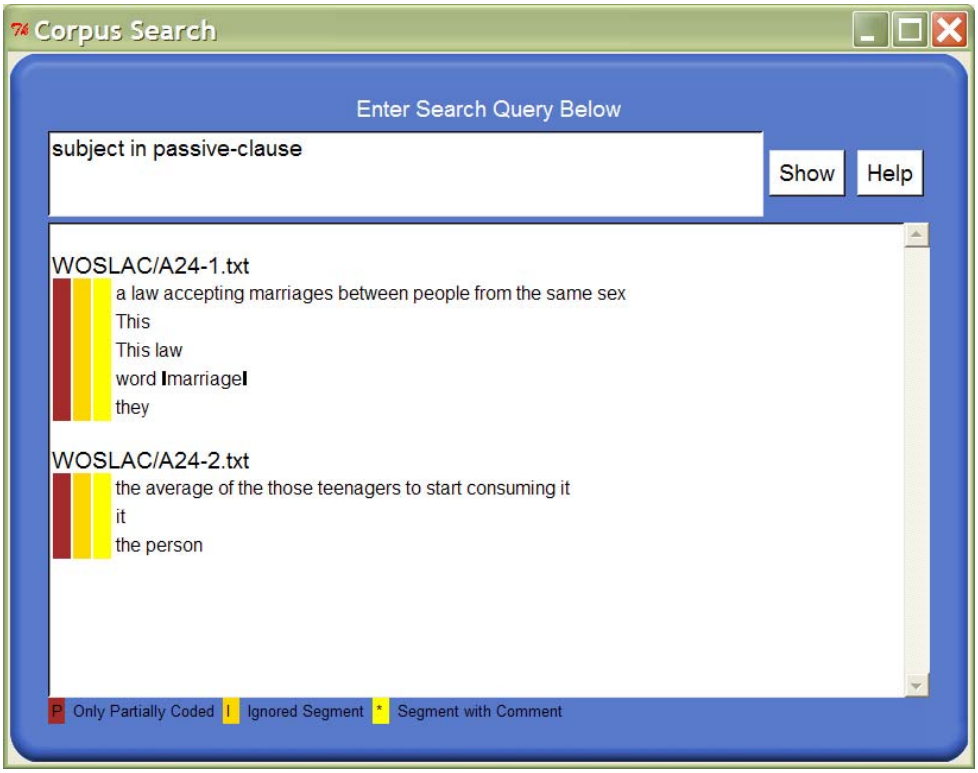

Figure 2: Search queries in CorpusTool

Because hand-annotation is slow, the tool will allow the analyst to associate lexico-syntactic patterns with each feature, allowing the tool to automatically detect instances of the pattern. For instance, a pattern like: "it be# NP that" would match sentences in the corpus like "It was John that we saw", and tentatively mark them with the feature *it-cleft*. The tool would then ask the user to eliminate false matches. This approach eliminates much of the corpus annotation effort.

4. Data analysis: statistics

CorpusTool also includes a simple statistical package as illustrated in Figure 3 (for another project).

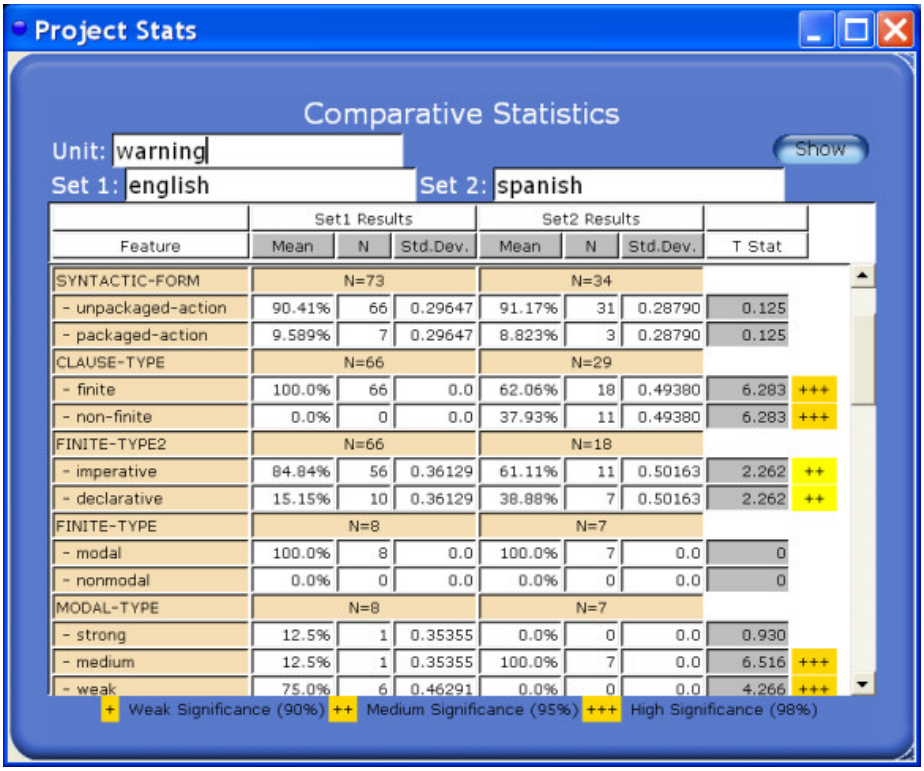

Figure 3: Statistics in CorpusTool

4. Contributions of the project

The project is still in an early phase, but we expect several valuable contributions from the study: (1) A contribution to linguistic theory in the identification of syntactic and pragmatic factors constraining word order in natural languages; (2) A better understanding of processes involved in the acquisition of word order and of the language learning process in general, revealed by the relative difficulty of interlanguage transfer of various phenomena; and (3) A third more concrete benefit will be the two corpora (English L2 and Spanish L2) that we collect , which we intend to make available to the community.

References

O'Donnel, M. (2007) CorpusTool 1.0. Available online from http://www.wagsoft.com/CorpusTool/index.html

Lozano, C. and A. Mendikoetxea (forthcoming), Postverbal subjects at the interfaces in Spanish and Italian learners of L2 English: a corpus study, in G. Gilquin, B. Díez and S. Papp (eds). Linking up Contrastive and Learner Corpus Research. Amsterdam: Rodopi.