The challenges and benefits of annotating bilingual corpora: The SpinTX Corpus Project

Barbara E. Bullock, Almeida Jacqueline Toribio, Arthur Wendorf

The University of Texas

Linguistic outcomes of bilingualism and language contact have long invited empirical scrutiny and inspired theoretical debates within all branches of the language sciences: Are there syntactic constraints on code-switching? Is an innovative variant contact-induced or internally motivated? Such questions persist, in part, because of differences among subfields over methods and data, and, more critically in our view, because rich, representative, accessible speech samples from bilingual speakers are simply lacking. In this paper we introduce our efforts to provide such a data set and search tools; in addition, we address the specific annotation processes that we are designing to resolve the problems posed by bilingual corpora and demonstrate how corpus techniques shed new light on long-standing issues in bilingualism.

Bilingual speech phenomena can be broadly divided into two types: *code-switching*, with morphemes from two or more languages overtly manifested in an utterance, and *convergence*, where the influence of one language on the other is manifested without an overt language switch (Haugen 1956, Clyne 1987). Examples of each type appear below.

(1) Code-switching
Mi abuelita murió cuando yo tenía <u>fourteen months old</u>.
'My grandmother died when I was 14 months old.'

(2) Convergence
Como no quiso <u>agarrar ayuda</u> del gobierno …(<conseguir)
'Since he didn't want to get help from the government…'

These examples, produced by the same speaker, are drawn from the Spanish in Texas corpus of SpinTX, a video corpus repository that currently holds 134 video-taped sociolinguistic interviews with Spanish speakers in Texas. The corpus, now ~500,000 words, is tagged for part of speech (POS) and includes comprehensive metadata for each speaker. Although the interviews are conducted primarily in Spanish, there is a great deal of English language usage, in the form of code-switching and convergence, which present unique problems for annotation.

Computational linguists and engineers have begun to address the challenge of annotating code-switched forms as in (1). This is generally resolved via a language identification process that proceeds sequentially, tagging in one language and then the other (Solorio & Liu 2008a,b for Spanish-English; Li et al. 2012 for Mandarin-English; and Diab & Kamboj 2011 for Hindi-English). In SpinTX, we POS tagged the entire corpus and created dictionaries of Spanish and English for the 5,000 most frequent English words from *A Frequency Dictionary of Spanish: Core Vocabulary for Learners* (Davies 2006) and *A Frequency Dictionary of American English* (Davies & Gardner 2010), respectively. We coded as English most words that matched up with words in our English dictionary, and coded all others as Spanish by default. Problems that arise, and that we are attempting to resolve via automation, include the presence of many interlingual homographs (e.g., *me*, *a*, *Texas*, *files* (<fields)) that are often incorrectly tagged as English, and the failure to tag infrequent English words (e.g., *Tex-Mex*, *Spanglish*) and loanwords (e.g., *troquita* 'small truck', *pushándome* 'pushing me); in addition, the procedure (correctly) tags each occurrence of the hedge "um" as English, thus overestimating the amount of code-switching or percentage of English used.

To date, no one has attempted to annotate the forms in (2) as possible covert bilingual variants, although identifying such expressions across corpora would provide valuable information for linguists who do not agree on whether such forms are contact-induced or internal semantic extensions (Otheguy & Stern 2010, Slva-Corvalán 1994). This question can be resolved quantitatively by examining the probability of occurrence of particular collocations from the Spanish in Texas Corpus with their probability of occurrence in reference corpora. We conducted a regular search of *agarrar* + OBJECT, with an NP (object) appearing within five words of the {agarrar} lemma in SpinTX and then ranked them according to frequency of appearance. We repeated this procedure on an 80,000-word corpus of Argentine newspaper texts (Larsen Serigos 2012) and a large corpus (~4 million) of oral Spanish (Davies 2006). We wrote an original script to specifically return the words that appear in proximity to *agarrar* in SpinTX that never or very rarely appear in its proximity in other corpora. Results indicate that *agarrar* is used at significantly higher rates in the Spanish in Texas corpus than in the other corpora, indicating semantic extension at the very least. More importantly, the procedure was accurate in returning the collocations that we had manually identified as expressions that appear to be calqued on English "to get+NP" (e.g., *agarrar trabajo* 'get a job')—approximately 40% of the SpinTX *agarrar* uses—pointing to a contact effect. The implication of this study is that corpora can contribute to resolving whether or not an innovation is contact-induced, an enduring concern in bilingualism (see Treffers-Daller).

We conclude the presentation by discussing how corpus approaches allow for a reframing of multiple problems that have been central to the literature on language contact and bilingualism.

References

Clyne, M. 1987. Constraints on code-switching: how universal are they? *Linguistics* 25: 739-764.

Davies, M. 2006. *A frequency dictionary of Spanish: core vocabulary for learners*. New York: Routledge.

Davies, M. & D. Gardner. 2010. *A frequency dictionary of contemporary American English: word sketches, collocates, and thematic lists*. New York: Routledge.

Diab, M. & A. Kamboj. 2011. Feasibility of leveraging crowd sourcing for the creation of a large scale annotated resource for Hindi English code switched data: a pilot annotation. *Proceedings of the 9th Workshop on Asian Language Resources,* pp. 36-40. Chiang Mai, Thailand.

Haugen, E. 1950. The analysis of borrowing. *Language* 26: 210-231.

Larsen Serigos, J. 2012. Loanwords along a continuum: a corpus-based approach to anglicisms in Argentina. Unpublished manuscript, The University of Texas at Austin.

Li, Y., Yue, Y., & Fung, P. 2012 A Mandarin-English code-switching corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2515-2519.

Otheguy, R. & N. Stern. 2010. On so-called Spanglish. *International Journal of Bilingualism* 15:85-11.

Silva-Corvalán, C. 1994. *Language contact and change: Spanish in Los Angeles.* Oxford: Clarendon Press.

Solorio, T. & Y. Liu. 2008a. Learning to predict code-switching points. *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 973–981. Association for Computational Linguistics.

Solorio, T. and Y. Liu. 2008b. Part-of-speech tagging for English-Spanish code-switched text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 1051-1060. Association for Computational Linguistics.

Trefers-Daller, J. 2012. Grammatical collocations and verb-particle constructions in Brussels French: a corpus-linguistic approach to transfer. *International Journal of Bilingualism* 16:53-82.