Muchos están mixing but few are mezclando: A data-driven analysis of AUX-V switching

Blake Holman, Almeida Jacqueline Toribio, and Barbara E. Bullock

The University of Texas

Linguists have long sought to identify and characterize the syntactic limits of Spanish-English code-switching, employing diverse methodologies and analytical lenses. Early observations of small data sets of U.S. Spanish-English bilingual speech indicated a general absence of switching between auxiliaries and main verbs; sequences such as *has seen, ha visto*, *is walking,* and *está camindo* were encountered only as monolingual units in the naturalistic conversations analyzed (Gumperz & Hernández-Chávez 1969, Lance 1972, Timm 1975). Drawing on intuitions of grammatical well-formedness, the prohibition on switching at this (**has visto*) and other junctures was formalized in the Functional Head Constraint (Belazi et al. 1994), which proscribed switching between functional elements and their complements. However, experimental studies uncovered differential responses for switching of *estar* 'be'+ present participle (1) versus *haber* 'have' + past participle (2) in reading times (Dussias 2003, Guzzardo & Dussias 2015) and in acceptability ratings (Giancaspro 2015). Such facts were central to promoting null theories of code-switching, by which switching was proposed to be constrained solely by the properties of lexical items; on the constraint-free approach promoted by MacSwan's (2000, 2009), for instance, the asymmetry (1 vs. 2) is attributed to the fact that *estar* selects and enters into a hierarchical relationship with a present participle complement via Merge, but *haber* bears features that trigger movement and reanalysis with the past participle, resulting in a 'bilingual' Xº category that will not converge at the phonological interface (PF).

1. √ Los ciudadanos están supporting the program.
2. * Los ciudadanos habían supported the program.

The increased accessibility of bilingual data on social media offers new avenues for detecting patterns of code-switching in the context of complex verbs. The present study reconsiders switching at the juncture of the present participle (1), pursuing a data-driven approach that benefits from the methods and tools of corpus and computational linguistics (Guzmán et al. 2016, 2017, Bullock et al. 2017).

As the initial step, we employed the TwitterSearch API in Python to fetch tweets that contained bilingual AUX-V bigrams; the bigrams were generated by taking the Cartesian products of the sets of *estar* or *be* and the present participle of the most common words ending in *-ing* in English and *-ndo* in Spanish. The heuristic yielded over 18,000 unique tweets from which a notable inequality in distribution emerges; while examples of *estar*+V_{ENG} (3) were abundant (~4,000 tweets) in tokens and type, sequences of *be*+V_{SPAN} (4) were rare and restricted (~100 tweets). This directional asymmetry is not predicted by the syntactic accounts proposed for code-switching at this juncture and has not, to our knowledge, been explored in the language contact literature.

3. *estar*+V_{ENG} (n= ~4,000)

   a. Estoy working, no me da la vida para más.

b. No sé si llorar porque estoy feeling o porque estos niños no saben escribir bonito

4. *be*+V_{SPAN} (n=~100)
   What if Cookie Cat was hablando en español

What is the nature of these mixed complex verbs, and what factors might account for the diminished observations of *be*+V_{SPAN} relative to *estar*+V_{ENG}?

We propose that these VP-internal switches reflect borrowing rather than code-switching, and we test this premise quantitatively. Towards that end, each word in the corpus was tagged for language, which allowed for calculation of the length of each monolingual span. We hypothesize that borrowing should be reflected in short, 1-2 word, other-language spans (Bullock et al. 2018). Results bear this out; of the tweets that contain a phrase in the form *estar*+V_{ENG}, the average English span length is 1.96, while it is 3.72 for Spanish. This indicates that Spanish is the matrix language (Myers-Scotton 1993, 2002) into which English participles are inserted (Muysken 2000). In other words, there is no alternation between grammars, and hence syntactic constraints on code-switching are not relevant for this construction.

In taking account of the differential attestation of observations of *be*+V_{SPAN} relative to *estar*+V_{ENG} in the bilingual tweets, we adopt the view that socio-cultural factors, not merely cognitive factors, may be at play (Bhat et al. 2016). Support for this claim is found in the directional asymmetries in DP-internal switching reported by Blokzijl et al. (2017) and Bullock & Toribio (forthcoming) for Spanish and by Bhat et al. (2016) for Hindi: switching at the Det-N boundary, when it occurs, is in the direction of the socially prestigious language—English.

The heuristic that we used to compile our data selected specific tweets that comprised the compound verbal construction. But a review of 310,000 tweets of U.S. Spanish (Solorio & Liu 2008) that did not explicitly target the relevant construction corroborates the trends reported here; while there were only two instances of the *estar*+V_{ENG} complex (5), *be*+V_{SPAN} was unattested.

5. Nada más chévere que cuando te llega el email de que lo que compraste por internet ya está shipping en camino.

To summarize, previous work on Spanish-English bilingual phenomena has shown that language alternation can occur at the juncture of auxiliary and present participle (Dussias 2003); however, our analysis of Twitter data, which is profuse and easily accessible, demonstrates that switching within the compound verb is infrequent (and unidirectional) and does not reflect code-switching, but rather a form of verbal borrowing, operationalized as low average span length. Furthermore, there is an asymmetry in the direction of switching, as the Spanish-English (*estar*+V_{ENG}) appeared 40 times more often than its English-Spanish counterpart (*be*+V_{SPAN}), an asymmetry that can be explained by appeal to the socio-cultural dominance of English, rather than linguistic or cognitive considerations. Overall, the mass retrieval of twitter data shows that bilingual compound verbs are illustrative of English verb insertion into Spanish.