Data before models

SHANA POPLACK
University of Ottawa
RENA TORRES CACOULLOS
The Pennsylvania State University

New theories are a constant of the now vast literature on code-mixing (CM). The *Gradient Symbolic Computation* model proposed by Goldrick, Putnam and Schwartz (Goldrick, Putnam & Schwartz) will appeal to many, especially those who already espouse constraint-based approaches to grammar. As variationist sociolinguists, we particularly welcome the model's incorporation of "relative probabilities of certain structures", a feature we believe can enhance our chances of capturing actual CM behavior. We also applaud Goldrick et al.'s efforts to integrate experimental findings on co-activation with grammatical principles. Our questions concern the utility of "doubling constructions" to showcase the model, and by extension, the degree to which it can account for bilinguals' SPONTANEOUS PRODUCTION of CM. A historical perspective on the field shows that none of the myriad theories of CM, often inspired by competing sets of grammatical principles, has yet achieved broad acceptance. In the absence of any widely endorsed evaluation metric – still sadly lacking -– how are we to decide amongst them?

Sociolinguists would require that a model be tested against, and supported by, the data of actual spontaneous bilingual production. We first observe that in this kind of data, the doubling constructions the model is claimed to account for are exceedingly rare. Goldrick et al. acknowledge this, but imply that such constructions are nonetheless a CONSISTENT feature of CM corpora (emphasis ours). Empirical research suggests otherwise: in over a dozen bilingual data sets (involving thousands of tokens of CM) systematically studied by us, only a very few (e.g., Poplack, Wheeler & Westwood, 1987; Sankoff, Poplack & Vanniarajan, 1990) featured any instances at all of such constructions, and rarely did these exceed a handful in any one. This makes the phenomenon not only rare but sporadic, explaining why Sankoff et al. (1990) referred to it (as would we) as an "ad-hoc processing mechanism". While doubling may nonetheless serve the purpose of model development, its very exceptionality raises the question of whether and how its analysis can be generalized to the bulk of the data, surely a desideratum of any predictive account of CM.

Even if doubling WERE a robust phenomenon, other key elements of the proposed model are also at odds with the facts of CM on the ground. Goldrick et al. maintain that "speakers are not only uttering lexical items from both languages but are also integrating grammatical principles from each linguistic system" and, further, that such integration is reflected in "the transfer of grammatical patterns from one language to another". But "blend representations" (p. 6) (already proposed by Weinreich, 1963), should not be conflated with grammatical convergence in production. Actual bilingual behavior, whether viewed in diachronic or synchronic perspective, fails in the aggregate to support blending; instead, it reveals bilinguals' knowledge and application of independent, language-particular, grammatical principles.

Diachronically, grammatical convergence between languages in contact, although widely assumed, is seldom satisfactorily demonstrated once accountable methodology and appropriate benchmarks are employed (e.g., Poplack, Zentz & Dion, 2012; Silva-Corvalán, 2008; Torres Cacoullos & Travis, 2015). Synchronically, empirical study of many bilingual speech communities has confirmed that the vast majority of CM data is constituted by LONE other-language items, which in turn, tend to be morphologically and syntactically integrated into RECIPIENT-language grammar (e.g., Poplack & Meechan, 1998). This is consonant with – and identical to – lexical borrowing. Here, although the lexical item derives from one (donor) language, only the grammar of the other language is at play, not a blend of both. MULTIWORD CM, on the other hand, is indeed compatible with co-activation of both of the speakers' grammars: intra-sententially, code-switching is strongly preferred where the word orders of both languages are homologous, a fact modeled by the Equivalence Constraint (Poplack, 1980). But switching at points of congruence between two languages does not entail amalgamation of their grammatical principles. On the contrary, *pace* coincidental inter-linguistic similarities, the multiword strings are internally consistent ONLY with the grammatical principles of the language from which

Address for correspondence:
Prof. Shana Poplack, University of Ottawa, Department of Linguistics, 401–70 Laurier Avenue East, Ottawa, ON K1N 6N5 *spoplack@uottawa.ca*

they are drawn. Thus, neither of the major manifestations of CM involves blending: lexical borrowing because only ONE (recipient-language) grammar is involved, code-switching because two grammars are INDEPENDENTLY at play.

Goldrick et al. adduce further evidence for blending from cross-language priming and cognate effects. But the only study of bilingual syntactic priming outside the lab (Travis, Torres Cacoullos & Kidd, 2015) reveals that it is weaker across than within languages, and that its relative strength varies according to contextual features and particular constructions of the language of the target, not the prime. As to cognate effects on phonetic variation, these are shown to emerge from the cumulative effect of usage patterns: when bilinguals' use of both languages is considered, cognate words occur less often than non-cognates in the relevant contexts (Brown, 2015). These findings from spontaneous speech, showing that the conditioning factors are language-specific, are consistent with "gradient co-activation" but not with blending of grammars.

Even CO-ACTIVATION, though clearly pertinent in at least some kinds of CM, is difficult to quantify. We agree that linguistic experience is crucial. But in bilinguals' experience, language-internal inherent variability, for example in word order (unacknowledged here, but see e.g., Poplack, Sayahi, Mourad & Dion, 2015; Sankoff et al., 1990), is inescapable. Given variability, what would constitute appropriate input training data for learning algorithms? And how can activation values be meaningfully estimated in the absence of information about speakers' actual exposure to, relative proficiency in and frequency of use of the languages, not to mention the prevailing norms of their bilingual community? These crucial predictors of the selection, form and placement of CM cannot simply be surmised nor are they typically available from lab-based studies of university-student L2 learners.

In sum, a minor phenomenon (if it qualifies as that) has served as the foundation for a model at odds with the patterns suggested by the behavior of major phenomena. Goldrick et al. assure us that their probabilistic approach can accommodate "sociolinguistic factors"; we look forward to this eventuality. In the interim, we commend the authors for bringing to the fore several

of the many challenges confronting the modeling of CM.

References

Brown, E. L. (2015). The role of discourse context frequency in phonological variation: A usage-based approach to bilingual speech production. *International Journal of Bilingualism, 19*, 387–406.

Goldrick, M., Putnam, M., & Schwartz, L. Coactivation in bilingual grammars: A computational account of code mixing. *Bilingualism: Language and Cognition*. doi: dx.doi.org/10.1017/S1366728915000802.

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en Español. *Linguistics, 18*, 581–618.

Poplack, S., & Meechan, M. (eds.). (1998). Instant loans, easy conditions: the productivity of bilingual borrowing. *International Journal of Bilingualism 2*, 2.

Poplack, S., Sayahi, L., Mourad, N., & Dion, N. (2015). An exception to the rule? Lone French nouns in Tunisian Arabic. *U. Penn Working Papers in Linguistics 21,2: Selected papers from NWAV 43*, 175–186.

Poplack, S., Wheeler, S., & Westwood, A. (1987). Distinguishing language contact phenomena: evidence from Finnish-English bilingualism. In P. Lilius & M. Saari (eds.), *The Nordic Languages and Modern Linguistics 6*, pp. 33–56. Helsinki: University of Helsinki Press.

Poplack, S., Zentz, L., & Dion, N. (2012). Phrase-final prepositions in Quebec French: An empirical study of contact, code-switching and resistance to convergence. *Bilingualism: Language and Cognition, 15*, 203–225.

Sankoff, D., Poplack, S., & Vanniarajan, S. (1990). The case of the nonce loan in Tamil. *Language Variation and Change*, *2*, 71–101.

Silva-Corvalán, C. (2008). The limits of convergence in contact. *Journal of Language Contact, 2*, 213–224.

Torres Cacoullos, R., & Travis, C. E. (2015). Two languages, one effect: Structural priming in spontaneous code-switching. *Bilingualism: Language and Cognition*, doi: 10.1017/S1366728914000406. Published online by Cambridge University Press, April 30, 2015.

Travis, C. E., Torres Cacoullos, R., & Kidd, E. (2015). Cross-language priming: A view from bilingual speech. *Bilingualism: Language and Cognition*, 10.1017/S1366728915000127, Published online by Cambridge University Press, 03 July 2015.

Weinreich, U. (1963). *Languages in contact: Findings and problems*. The Hague: Mouton.