Bilingualism: Language and Cognition 22 (4), 2019, 682–684 © Cambridge University Press 2018 doi:10.1017/S1366728918000755

COMMENTARY

# Scaling up: How computational models can propel bilingualism research forward

PING LI
*The Pennsylvania State University*
ANGELA GRANT
*Concordia University*

(Received: May 08, 2018; final revision revised: May 24, 2018; accepted: May 31, 2018)

The Multilink model that Dijkstra, Wahl, Buytenhuijs, van Halem, Al-jibouri, de Korte, and Rekké (2018) present is an excellent example that connects empirical patterns obtained from behavioral studies with mechanisms that can be implemented in computational models. We have previously argued that implementation of computational models is important because it forces the researchers to be explicit about assumptions and to specify parameters and variables that may be absent in verbal models. The Multilink model, along with BIA/BIA+ and many other models, provides concrete hypotheses regarding the role of variables such as word frequency, word length, orthographic similarity, and phonological neighborhood for researchers to test and verify against empirical data (see examples in the special issue on computational modeling published in this journal; Li, 2013).

A key attraction that Multilink has over previous models is that it scales up to a more realistic lexicon (e.g., over 1500 words from both languages in the enriched lexicon). Bilingual models should indeed aim at scaling up to the experience of real language users, and computational models can propel us forward in this direction. Empirical work may often be restricted to small sets of limited vocabulary from the bilingual's two languages due to resource and time constraints in any given experiment, but computational modeling has the flexibility and freedom to include large sets of lexical items, modeled over large-scale datasets or databases (e.g., the English Lexicon Project and the Dutch Lexicon Project, as used by Multilink; Balota, Yap, Cortese, Hutchison, Kessler, Loftis, Neely, Nelson, Simpson & Treiman, 2007; Brysbaert, Stevens, Mandera & Keuleers, 2016). In our view, there are two dimensions that the current version of Multilink has not considered sufficiently, where future work could be done to further scale up the modeling effort.

First, Multilink lacks real semantic representations in the lexicon. Although Multilink has the advantage of deriving accurate orthographic similarity between languages via the Levenshtein distances, the model approximates semantic representations only indirectly through localist connectionist node representations. A more direct approach could be attained through vector representations of the type used in semantic space models such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), Hyperspace Analogue to Language (HAL; Burgess & Lund, 1997), and many other semantic vectors that can be generated by popular software such as Word2vec (Mikolov, Chen, Corrado & Dean, 2013) or GloVe (Caliskan, Bryson & Narayanan, 2017). It is unclear why efforts elsewhere in language and text research cannot be incorporated more easily into models of bilingualism, and the lack of true semantic representations has already clearly impacted the ability of Multilink in modeling cross-language homographs, homophones, and words that do not match neatly across two languages (but see, for example, Fang, Zinszer, Malt & Li, 2016; also Malt, Pavlenko, Zhu & Ameel, 2015; Zinszer, Malt, Ameel & Li, 2014). Instead, Multilink's lexicon is highly restricted to words that share semantics (e.g., cognates) or words that have translation equivalents across two languages (Dutch and English). This will severely limit the model's generalizability to other language pairs where the correspondence between word pairs is much less transparent.

Second, Multilink only provides an account of the proficient adult bilingual speaker. Although Dijkstra and colleagues attempt to model differences in proficiency by adjusting the resting level activation of the lexicon, such an approach is overly simplistic, as they themselves have noted (p. 42). The lack of formal models of bilingual language acquisition has been a long-standing problem that one of us has pointed out in several other contexts (Li, 2002; Li & Farkas, 2002).

In order to move forward, future models will need to consider language proficiency with more nuance (see an example in Thomas, 1997). Although word frequency is certainly related to proficiency, assuming that the two are

Address for correspondence:
Ping Li, 452 Moore Building, Department of Psychology, The Pennsylvania State University, University Park, PA 16802
*pul8@psu.edu*

equivalent is problematic. First, consider the differences between word frequency as assessed by SUBTLEX and word frequency as experienced by a classroom L2 learner. For a classroom learner, items like "desk" and "stapler" are likely to be much more highly frequent than would be estimated by SUBTLEX, and vice versa. Thus, any predictions made using a database constructed from native speakers' productions may not accurately capture the experience of L2 learners. Second, even modeling a highly proficient, simultaneous bilingual (as in Multilink) could be problematic when based on the data from two monolingual corpora. After all, "The bilingual is not two monolinguals in one person" (Grosjean, 1989) and this is reflected by slower speech processing in each language compared to monolinguals, which may be due to overall reduced frequency of use (e.g., Gollan, Montoya, Cera & Sandoval, 2008). Third, even if a model were to use frequency information based on a learner corpus, using frequency alone as a measure of estimating proficiency ignores the modality of L2 use, another critical factor. Many learners report higher proficiency in speech comprehension than in production, and in some cases, such as heritage language learners, may have an extensive comprehension lexicon while rarely producing those same items. These differences in exposure to language comprehension and production may have meaningful effects on proficiency (e.g., Hopman & Macdonald, 2018; Kang, Gollan & Pashler, 2013) that are ignored in the current manifestation of Multilink.

Another factor to consider when attempting to model proficiency, and one perhaps more easily accommodated within Multilink's existing structure, is word association. Although the authors propose that lexical links between translation equivalents could be implemented in the model, they contest that these links are not necessary "at least in the case of the more proficient bilinguals whose data we simulated" (p. 40). However, such an assumption misses the opportunity to potentially improve the model's simulation of lower proficiency bilinguals, which is the same population that the word-association route was meant to accommodate in the RHM (Kroll & Stewart, 1994). Moreover, part of the appeal of formalized models is the ability to explicitly test predictions in a way that verbal models (such as the RHM) lack.

In conclusion, Multilink represents a significant step forward in the modeling and understanding of bilingual lexical processing. However, there are places where the model could be scaled up further. We have identified two potential next steps – improving the modeling of semantic representations and proficiency – and we look forward to seeing where research using Multilink will lead.

## References


Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39,* 445–459.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance, 42,* 441–458.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language & Cognitive Processes, 12,* 177–210.

Caliskan, A., Bryson, J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356,* 183–186.

Dijkstra, A., Wahl, A., Buytenhuijs, F., van Halem, N., Al-jibouri, Z., de Korte, M., & Rekké, S. (2018). Multilink: a computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition,* doi:10.1017/S1366728918000287.

Fang, S., Zinszer, B., Malt, B., & Li, P. (2016). Bilingual object naming: A connectionist model. *Frontiers in Psychology: Cognitive Science, 7,* Article 644. doi: 10.3389/fpsyg.2016.00644

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language, 58,* 787–814. https://doi.org/10.1016/j.jml.2007.07.001

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language, 36*(*1*), 3–15.

Hopman, E. W. M., & Macdonald, M. C. (2018). Production Practice During Language Learning Improves Comprehension. *Psychological Science*. Advance online publication. https://doi.org/10.1177/0956797618754486

Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review, 20*(*6*), 1259–65. https://doi.org/10.3758/s13423-013-0450-z

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33,* 149–174.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013. (arXiv preprint arXiv:1301.3781).

Li, P. (2002). Bilingualism is in dire need of formal models. *Bilingualism: Language and Cognition, 5,* 213.

Li, P., & Farkas, I. (2002). A self-organizing connectionist model of bilingual processing. In R. Heredia & J. Altarriba (Eds.),

*Bilingual sentence processing* (pp. 59–85). North-Holland: Elsevier Science Publisher.

Li, P. (2013). Computational modeling of bilingualism: How can models tell us more about the bilingual mind? *Bilingualism: Language and Cognition, 16,* 241–245.

Malt, B., Li, P., Pavlenko, A., Zhu, H., & Ameel, E. (2015). Bidirectional lexical interaction in late immersed Mandarin-English bilinguals. *Journal of Memory and Language, 82,* 86–104.

Thomas, M. (1997). Connectionist networks and knowledge representation: The case of bilingual lexical processing. Unpublished dissertation, Oxford University, UK.

Zinszer, B. D., Malt, B. C., Ameel, E., & Li, P. (2014). Native-likeness in second language lexical categorization reflects individual language history and linguistic community norms. *Frontiers in Psychology: Language Sciences, 5,* Article 1203. doi: 10.3389/fpsyg.2014.01203