Mechanistic formal approaches to language acquisition

Yes, but at the right level(s) of resolution

Stefan Th. Gries
University of California, Santa Barbara

I am coming to Yang's target article (Yang, 2018) as a usage-based quantitative corpus linguist with interests, though not the greatest degree of expertise, in first/second/foreign language acquisition. There is a lot to like in the keynote paper both in terms of theoretical and methodological orientation: I am quite sympathetic to how Yang defines his formalist ('mechanistic') perspective; I appreciate his plea for methodological rigor and its emphasis over mode of explanation; I agree with his view that a (formal) theory of language acquisition ought to be more than a description of patterns and correlations but needs to include a mechanistic and causal account. Somewhat more specifically, I also particularly welcome the notion to develop mathematical models of language acquisition/learning especially when such models are probabilistic, can accommodate individual variation, and are compatible with a lesser role of innate parameters than much more generative work has been arguing for.

All that being said, it is a truism that the devil is in the details. Space does not permit me to discuss a variety of (admittedly smaller) issues so I will focus here on Yang's characterization of the state and standards of research on usage-based learning. In particular, Yang argues that "the evidence for an initial item-based stage of child language has been overstated" (p. 674), that claims regarding 'statistical dominance' are insufficient and claims regarding item-based learning 'require more work', that at a minimum, one needs to show that, say, unanalyzed collocations such as *give me* are conspicuously more frequent than a productive rule account would predict (p. 674).

But there are two reasons why his points of critique are not all compelling. First, Yang's terminology is just as imprecise as that of the usage-based work he discusses critically: what does 'statistically predominant' mean, what is 'conspicuously more

https://doi.org/10.1075/lab.18050.gri
Linguistic Approaches to Bilingualism 8:6 (2018), 733–737.
ISSN 1879–9264 | E-ISSN 1879–9272

frequent'? Even, or maybe especially, in a target article, we need more precise criteria – if only to agree with them or to admit that usage-based linguistics has not reached those thresholds. Relatedly, the choice of works cited is selective: Unless I missed something, with the exception of Pine et al. (2013), none of the usage-based original research articles in Section 3 'Rules vs. Storage', for instance, is less than 15 years old and even Pine et al. is only mentioned in a footnote but not where the issue of determiner imbalance is discussed in the main text. While I myself have issues with some of the usage-based acquisition work, this does not seem to compare both approaches on the same level of evolution; I would be very interested to see how Yang's variational model and equations stack up against, or even just compare and relate to, the kinds of studies discussed in Chapter 5 of Christiansen and Chater (2016) that highlight the power of distributional information or, for instance, work using the Vapnik-Chervonenkis dimension, which Christiansen and Chater (2016, 155f.) state "establishes an upper bound for the number of examples needed by a learning process that starts with a set of hypotheses about the task solution" and which Yang is of course familiar with.

The second reason why I feel that some of Yang's discussion falls a bit short is more important, however, and it is actually a reason for me to also be somewhat unhappy with some of the usage-based corpus-linguistic work on first/second language acquisition. Simply put, much like several of Yang's arguments and mathematical models, a lot of corpus-linguistic work severely underutilizes the huge amount of probabilistic information corpora have to offer by restricting itself to absolute and relative frequencies of (co-)occurrence of types and tokens (and maybe their ratios). However, as especially Nick Ellis and collaborators have argued repeatedly (e.g. Gries & Ellis, 2015; Ellis, 2016), much more information needs to be included: *frequency*, yes, but also *dispersion* of elements in a corpus (because dispersion is related to learning, Ambridge et al., 2006; Gries, 2008), *association/contingency* (Ellis, 2006), *predictability/surprisal* (Smith & Levy, 2013), *entropy, salience, prototypicality* (Ellis, Römer, & O'Donnell, 2016), and more. Thus, while I am genuinely sympathetic to the general idea of the variational model and the Tolerance/Sufficiency principles, these, too, try to explain something as complex as learning a first/second language – with all the multidimensional information that entails – on the basis of little more than type and token frequencies, which is certain to underestimate both the complexity of the task at hand and disregards any information regarding the salience or surprisal of exceptions to the hypothesized rules/grammars.

As a brief example, consider *gimme*, which Yang discusses in Section 3.1. First, it is not only just the high frequency of *give me* that led usage-based linguists to hypothesize that this might be a single unit but precisely the fact that it is frequently pronounced in the reduced version of *gimme*. Second, whether

or not *give me*/*gimme* is a unit requires more 'methodological rigor' and clarity than a mere comparison of ratios of *give me*: *give him*: *give her* vs. *me*: him: *her*; there is much corpus-linguistic work on assessing association/contingency and I briefly checked the associations of the words following *give* in the Brown (1973) corpus (downloaded from <https://childes.talkbank.org/data/Eng-NA/Brown.zip> 10 June 2018).

These collocates were identified (rather heuristically) by loading and conflating each file into one string, breaking it up at utterance tiers, retrieving utterances containing v|give in the %mor tier, and finding sequences of non-spaces after *give* separately for children and caretakers. Then, I determined for each type attested after *give* its overall frequency in the corpus (separately for children and everyone else). Then, to keep effects of frequency and association separate (see Gries & Ellis, 2015; Gries, 2018), I collected for each collocate its frequency after *give* and the log odds ratio, an association measure ranging from $-\infty$ to $+\infty$, where the sign of the log odds ratio indicates whether the collocate is more often (positive) or less often (negative) observed after *give* than expected by chance; the absolute value of the log odds ratio quantifies the strength of the effect. To illustrate this approach for the word *her*, the data amount to those shown in Table 1, from which one can compute an odds ratio of $(^{12}/_{300} / {}^{380}/_{89144}) \approx 9.3836$, leading to a log odds ratio of ≈ 2.239.

Table 1. 2 × 2 co-occurrence matrix for *give* and *her* in the Brown (1973) corpus

	Give	**Other words**	**Sum**
her	12	380	392
other words	300	89144	89444
Sum	312	89524	89836

The two panels of Figure 1 show the results for all collocates following *give*. As indicated in the *x*-axis labels, the left panel shows the results for all three children combined whereas the right one shows the results for the caretakers. In each panel, the *x*-axis represents the co-occurrence frequency of the collocate with *give* (logged to the base of 2) and the *y*-axis represents the log odds ratio of the collocate and *give* as computed above; lemmas printed in green exhibited a significant attraction/repulsion to the slot after *give*, lemmas in purple did not.

Clearly, the data are at least compatible with the notion that *gimme* might be a unit. Not only do we find that it is often transcribed with the phonological reduction from *give me* to *gimme*, but we also see in both the children's and the caretakers' output that *me* is the most frequent collocate after *give* and, more importantly, it is also the first or second most strongly attracted collocate to the

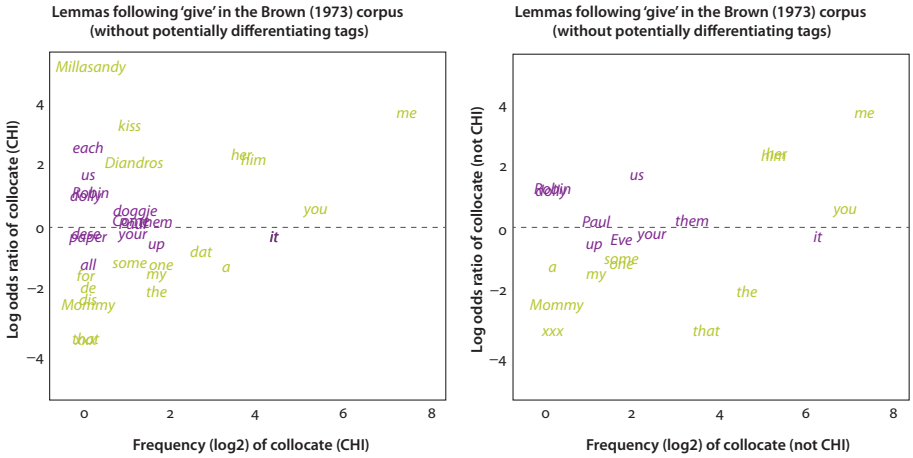

Figure 1. Frequencies and association measures for lemmas following *give* in the Brown (1973) corpus of language acquisition (left panel: children, right panel: caretakers)

position after *give* when the collocates overall frequencies in the corpus are taken into consideration. Similarly, *give me* is also the most evenly dispersed *give+x* collocation in the corpus (as measured by the *DP* index, see Gries, 2008), which means children's chance to encounter it (often) are highest; this is particularly relevant given how little of the actual input to and output of the child we actually have (Tomasello & Stahl, 2004).

None of this reduces the onus on usage-based work in general to provide "rigorous" testing of their L1/L2 acquisition data – I have been making similar points for years – but it shows that the situation may not be as clear cut as Yang makes it appear to be. While the above analysis is not certainly not the most sophisticated corpus-linguistic analysis one can do (and of course it's also no proper linguistic analysis), it shows that a mere comparison of frequency ratios is neither. A position paper that criticizes usage-based work for 'overstating their evidence,' not doing 'enough work', and lacking proper hypotheses/rigor needs to be more compelling even if much usage-based work has also not done the ideal corpus statistics: All sides of the debate – formalists, usage-based linguists, but also general corpus linguists like myself – need to step up their quantitative corpus-linguistic game.

References

Ambridge, B., Theakston, A. L., Lieven, E. V. M., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development*, 21(2), 174–193. https://doi.org/10.1016/j.cogdev.2005.09.003

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/harvard.9780674732469

Christiansen, M. H. & Chater, N. (2016). *Creating language: integrating evolution, acquisition, and processing*. Cambridge, MA: The MIT Press.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1), 1–24. https://doi.org/10.1093/applin/ami038

Ellis, N. C. (2016). Cognition, corpora, and computing: triangulating research in usage-based language learning. *Language Learning*, 67(51), 40–65. https://doi.org/10.1111/lang.12215

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing*. New York: Wiley-Blackwell.

Gries, St. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, St. Th. (2018). Operationalizations of domain-general mechanisms cognitive linguists often rely on: a perspective from quantitative corpus linguistics. In S. Engelberg, H. Lobin, K. Steyer, & S. Wolfer (Eds.), *Wortschätze: Dynamik, Muster, Komplexität* (pp. 75–90). Berlin & Boston: De Gruyter. https://doi.org/10.1515/9783110579963-005

Gries, St. Th. & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, 65(Supplement 1), 1–28. https://doi.org/10.1111/lang.12119

Pine, J. M., Freudental, D., Krajewskyi, G., & Gobet, F. (2013). Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition*, 127(3), 345–360. https://doi.org/10.1016/j.cognition.2013.02.006

Smith, N. J. & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Tomasello, M. & Stahl, D. (2004). Sampling children's spontaneous speech: how much is enough? *Journal of Child Language*, 31(1), 101–121. https://doi.org/10.1017/S0305000903005944

Yang, C. (2018). A formalist perspective on language acquisition. *Linguistic Approaches to Bilingualism*, 8(6), 665–706.

Address for correspondence

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106–3100
United States of America

stgries@linguistics.ucsb.edu
https://orcid.org/0000-0002-6497-3958

Publication history

Date received: 11 June 2018
Date accepted: 15 August 2018