Meta-Megastudies

James Myers
National Chung Cheng University
DRAFT
2015/10/15

Abstract

Cross-linguistic data have always been of interest to mental lexicon researchers, but only now are technological developments making it possible to coordinate experiments across languages in a typologically sophisticated way, in an approach we dub meta-megastudies. A meta-megastudy not only tests a large sample of items or speakers but also a large sample of languages, so that partially confounded language-specific variables can be teased apart the way traditional megastudies tease apart partially confounded word-specific variables. Meta-megastudies are made possible by Web-based tools that allow independent research groups to run standardized experiments and share the results.

Keywords: megastudies; phonology; cross-linguistic; typology; Chinese

The language-as-fixed-effect fallacy occurs when a psycholinguist runs experiments on one or two languages, then draws inferences about human language processing that go beyond the specific languages being tested. At least this is what the term should mean. The solution to the problem is similar to that recommended by Clark (1973) for the more familiar use of the term: treat language as a random variable. In other words, just as a megastudy (Balota, Yap, Hutchison, & Cortese, 2012) can generalize about word processing in some specific language by testing a large sample of words from that language, so too can one generalize about human word processing by testing a large sample of human languages. We will call this a meta-megastudy.

The rest of this essay is devoted to unpacking this simple idea. I first show how a typologically more sophisticated approach to cross-linguistic research would benefit psycholinguistics, and then I show how such an approach might be made practical. I end by sketching the outlines of Worldlikeness, a Web-based system for wordlikeness judgment meta-megastudies, currently under development.

The Need for Meta-Megastudies

Psycholinguists are deeply interested in universal principles, and they already have

conventions for discovering and testing them: (a) focus most research on a well-studied base language, and (b) only turn to other languages when competing hypotheses cannot be discriminated within the base language alone. This is why readers of psycholinguistics papers must often wait until the methods section before learning what language was tested (in which case it was probably English). These conventions may seem designed to irritate linguists, but both have some justification. If a psycholinguist argues for a new universal claim about semantic processing on the basis of a reading task in an obscure language, skeptics may rightly ask how we can rule out mere orthographic effects without a paper trail of previous reading studies in this language. As for convention (b), scientists rarely collect data blindly, but rather search systematically for observations that do or do not conform to some hypothesis: when comparing alternative theories of how reading is affected by orthographic transparency, Chinese suddenly becomes a crucial test case.

Linguists have some right to be irritated by these conventions anyway. Convention (a) seems to assume that we can say nothing about the higher levels of a complex system until we have thoroughly understood all of the lower levels, but if psychologists truly believed this, they would all give up their button boxes for brain scanners. While discoveries about high-level processes in exotic languages should still be linked to those in more familiar languages, this may be possible at the high level itself, and if not, the lower levels can be filled in later.

Convention (b) is even more problematic. Aside from arbitrarily marginalizing almost all human languages as exotic, it risks missing crucial information about the mind. Psycholinguists know that languages differ (e.g., Bates, Devescovi, & Wulfeck, 2001; Costa, Alario, & Sebastián-Gallés, 2007; Cutler, 1985; Jaeger & Norcliffe, 2009), and may even know that they differ more than we can yet explain (Evans & Levinson, 2009). This situation, however, means that convention (b) has a serious blind spot. If we do not know ahead of time how languages can differ, we cannot search for new test languages by manipulating parameters derived solely from familiar languages. Rather, like astronomers sweeping the sky for theory-challenging surprises, we should make cross-linguistic study part of the routine.

Convention (b) also faces the same inferential problems faced by factorial designs in lexical research more generally. Lexical items already exist and so cannot be experimentally manipulated. Forcing a factorial design by cherry-picking subsets of matched items risks experimenter bias (Forster, 2000) without eliminating confounds with uncontrollable or unknown variables (Balota et al., 2012; Cutler, 1981). Megastudies attempt to ameliorate these problems by treating experiments as database-generation machines, permitting future researchers to analyze the results in terms of any variable they can think of, using statistical techniques (clustering, regression) to tease apart partial confounds. Megastudy databases have already become mainstream, both for "base" languages like English (Balota et al., 2007; Keuleers, Lacey, Rastle, & Brysbaert, 2012) and Dutch (Keuleers, Diependaele, & Brysbaert,

2010) and for "exotic" languages like Chinese (Sze, Liow, & Yap, 2014; Myers, 2015) and Malay (Yap, Liow, Jalil, & Faizal, 2010).

Nevertheless, within any given megastudy, the language itself is not a manipulable independent variable either. Nor does testing two languages help much, since no language pair differs in just one way. Run ever-larger meta-megastudies, however, testing ten, twenty, one hundred languages, and it should become increasingly possible to tease apart language-specific confounds as well.

To make the discussion a bit more concrete, consider just a single theoretical issue: the apparent cross-linguistic differences in the decomposition of syllables into phonemes (extending the discussion of Myers, 2012). While there is good evidence that spoken word processing depends crucially on phonemes in English (and the other European languages that have been tested), both in speech perception (e.g., Benkí, 2003; Norris & Cutler, 1988) and in production (e.g., Fromkin, 1971; O'Seaghdha, Chen, & Chen, 2010), the status of phonemes in Chinese (i.e., Sinitic languages like Mandarin and Cantonese) is much less clear than that of syllables.

Chinese linguists have known about onset consonants and rimes for around 1,500 years (Malmqvist, 1994), and Chinese poets have long used alliteration (Cai, 2008), but Chinese characters themselves represent monosyllabic (and monomorphemic) units, not phoneme-sized units. Moreover, ordinary Chinese readers seem to be much less sensitive to onset consonants than linguists and poets: a child's ability to detect word onsets helps when learning to read English but not when learning to read Chinese (McBride-Chang et al., 2008).

Based on wordlikeness judgment experiments, Chinese speakers also seem to be less comfortable than English speakers with expanding their existing syllable inventory via novel combinations of phonemes. While the results of Bailey and Hahn (2001) imply a mean acceptance score around .42 (on a zero-to-one scale, transformed from a nine-point Likert scale) for their nonlexical test syllables, Myers (2015) reports an overall acceptance rate of only .11 (in a binary judgment task) for nonlexical syllables. Moreover, (holistic) neighborhood density and (analytical) phonotactic probability affect wordlikeness roughly equally in English, each account for roughly 15% of the response variance in Bailey and Hahn (2001), but in a nonlexical syllable wordlikeness judgment task in Cantonese (on a seven-point Likert scale), Kirby and Yu (2007) found that neighborhood density accounted for around 33% of the response variance, compared with just 2% for phonotactic probability.

The evidence that phonemes play a role in Mandarin word production is similarly weak. Chen, Chen, & Dell (2002) and O'Seaghdha et al. (2010) found no reaction time effects due to onset consonant priming in a form preparation (implicit priming) task (cf. the robust effects in English reported by O'Seaghdha et al., 2010). Qu, Damian, & Kazanina (2012) and Yu, Mo, & Mo (2014) also found no behavioral evidence for onset priming in other production tasks, though they did find priming in event-related potential components (Yu, Mo,

Li, & Mo, 2015, found onset priming in an fMRI study, but they report no behavioral results). Thus despite disagreements in detail (O'Seaghdha, Chen, & Chen, 2013; Qu, Damian, & Kazanina, 2013), the consensus from production studies supports the general conclusion that phonemes are much less important for Mandarin speakers than for English speakers.

The question is why. The answer may seem obvious: English orthography spells out consonants and vowels, Chinese orthography does not. While the evidence that orthography affects speech processing is mixed (Alario, Perre, Castel, & Ziegler, 2007; Rastle, McCormick, Bayliss, & Davis, 2011), the idea is not intrinsically implausible. Focusing exclusively on orthography, however, begs the question of how Chinese manages to get away with a non-phonemic orthography in the first place. A common answer is that an alphabetic orthography would be impractical because the Chinese lexicon is rife with homophones (e.g., Sampson, 2015). Thus in addition to orthography, we must also consider homophony as an explanation for the processing difference between English and Chinese.

These two explanations, in turn, relate to a third and fourth: homophones arise so readily in Chinese because virtually all Chinese morphemes are monosyllabic, and the syllable inventory in Mandarin (approximately 1,300, even taking tone into account; Myers, 2012) is around ten times smaller than that in English or Dutch (approximately 12,000 each; Levelt, Roelofs, & Meyer, 1999). This makes it ten times more feasible for Mandarin speakers simply to memorize their lexical syllables, even if they also weakly decompose them, and even if Dutch speakers also memorize their top few hundred most common syllables, as Levelt et al. (1999) suggest.

Kirby and Yu (2007) do not explain their Cantonese wordlikeness results directly in terms of syllable inventory size, but rather (in a fifth difference) by pointing to the fact that the Cantonese lexicon uses much more of the syllable space defined by syllable components (36%) than does English (6%). If logically possible syllables are too likely to be lexical, phonemic decomposition becomes an unhelpful strategy in making wordlikeness judgments, and perhaps in other experimental tasks as well.

Syllable inventory size is itself related to what is phonotactically permissible in a given language. For example, unlike Chinese, English permits consonant clusters and complex rimes, and more generally, as Chen, Dell, and Chen (2007) show, phonemes are statistically less predictable in the English lexicon than in the Mandarin lexicon. This leads them to speculate that this (a sixth difference between English and Mandarin) is the ultimate explanation for the weaker activation of phonemes in Chinese processing.

Orthography, homophony, monosyllabicity, syllable inventory size, ratio of lexical to possible syllables, and phoneme predictability are related, but not identical. Indeed, these variables seem to be associated with quite distinct aspects of processing (most obviously in the case of the orthography-, syllable-, and phoneme-based variables). This makes it unlikely that confounding can be side-stepped by combining them into a single information-theoretic

predictor (cf. Moscoso del Prado Martín, Kostić, & Baayen, 2004). Fortunately, their intrinsic distinctness also implies that cross-linguistic correlations in these variables are likely to be noisy, and so it should be possible to separate them statistically. For example, phoneme predictability depends not just on the size of the syllable inventory, but also on the size of the phoneme inventory.

Since English and Chinese differ in all of these variables, what other sorts of languages should we test? Many languages are traditionally unwritten or have large illiterate populations, so it should not be too difficult to find speakers of languages with large syllable inventories who nevertheless have as little orthographic training in syllable decomposition as Mandarin speakers. As for the purely phonological variables, the WALS database (World Atlas of Language Structure; Haspelmath, Dryer, Gil, & Comrie, 2005) includes the parameters consonant inventory (five levels), consonant-vowel ratio (five levels), and syllable structure (three levels). Crossing these in the online interface generates a table with 51 non-empty cells, a result that bodes well for a meta-megastudy dependent on there being sufficient variability in factors similar to these.

There remains, however, yet another possible explanation for the observed processing differences across English and Chinese: the observations are wrong. Convention (a) is still valid to the extent that we know a lot more about processing in English than in Mandarin. Until Norris and Cutler (1988) fixed experimental confounds that beset earlier studies (e.g., Savin & Bever, 1970), researchers had seriously considered the claim that syllables were more important than phonemes in English spoken word recognition, just as is now claimed about Mandarin spoken word production (albeit with better evidence). The methodological diversity of the studies I reviewed above casts further doubt on tentative cross-linguistic conclusions; the only controlled cross-linguistic comparison in my survey is O'Seaghdha et al. (2010).

Mental lexicon research is thus in need of an approach capable of testing a wide variety of languages in a unified way. That this is more than mere pipe dream is demonstrated in the next section.

The Logistics of Meta-Megastudies

To put it crudely, the key to bringing meta-megastudies from theory into reality is to take crowdsourcing, that favorite buzzword of the megastudy literature, and add another: citizen science. Crowdsourcing is when an elite group has the masses do menial tasks for them, as when psychologists (e.g., Graham et al., 2011), psycholinguists (e.g., Keuleers & Balota, 2015) or even theoretical linguists (e.g., Erlewine & Kotek, 2015) run experiments on the Web. Citizen science is when the masses themselves do science, making small-scale observations (e.g., cataloging visitors to home bird feeders) that can be compiled and studied

for large-scale patterns (Silvertown, 2009; Bishop, 2014). Of course like many buzzwords, crowdsourcing and citizen science blur together; the difference I want to highlight is the degree of responsibility taken by the contributors. Meta-megastudies must distribute responsibility because they only become feasible when not only the raw response data, but the experiments themselves, are generated by large numbers of people. The consequence is that meta-megastudies require a much less centralized working style than scientists are generally familiar with, via a widely distributed but uniformly maintained network (Nielsen, 2012, suggests that this is the future of science more generally).

This is the only way to run meta-megastudies because language-specific experts are needed to maintain quality control, but research groups (even large ones) are orders of magnitude smaller than the number of human languages (often estimated at around 5,000). Thus language sampling by a single research group must either be small and skewed, or riddled with errors. The great burden of getting the data right is illustrated by the seven-language picture naming megastudy of Bates et al. (2003), which involved an astonishing 22 authors, affiliated with 10 different institutions. Even with all this work, their choice of languages (English, Spanish, Italian, German, Bulgarian, Hungarian, Mandarin) seems to be more of a convenience sample than a typologically motivated survey.

The Web already has a smattering of decentralized information compilation projects where contributors are treated like scholars rather than data points. The most famous is Wikipedia, which despite being written and edited by thousands of anonymous people, is often lauded for its accuracy (e.g., Clauson, Polen, Boulos, & Dzenowagis, 2008), though not without caveats (e.g., Kupferberg & Protus, 2011). Within linguistics, WALS (created by a large but fixed group of experts) is being supplemented by projects like Terraling (http://www.terraling.com), which allows linguists to upload and share more detailed linguistic descriptions than is possible in WALS, with a particular focus on syntax. Even more interesting is the Endangered Languages Archive (ELAR: http://elar.soas.ac.uk/; Nathan, 2013), which explicitly encourages interactions between the creators of its language databases and the Web visitors who use them, including speakers of the archived languages themselves.

What I propose, then, is to merge the technology and philosophy underlying projects like Wikipedia with those underlying Web experimentation and database publication. Given that I have suggested that quality control is the main reason to distribute rather than centralize cross-linguistic experimentation, it is reasonable to ask why users would trust data on a language they do not know from an experimenter they have never met, or why experimenters would want to contribute high-quality data in the first place. Similar questions have been researched with respect to open access development more generally (e.g., Heylighen, 2007), and further ideas about how to improve the quantity and quality of meta-megastudies are available from all across the world of social media. Perhaps experimenters could be enticed

by a meta-megastudy system designed to serve their own selfish purposes. For example, the system may not only streamline online experimentation, but also act as a password-protected psycholinguistic cloud service, with the gentle reminder that the results of other experimenters are accessible only to users who reciprocate by sharing theirs. To ensure reputability, the system could require experimenters to identify their scholarly credentials, and there could even be tools for experimenters to be publicly rated by peers and participants for research ethics and linguistic competence. If all else fails, moderators (transparent and community-approved) may also be helpful.

The key job of such a system would be to facilitate the testing of as wide a variety of languages as possible, while still maintaining methodological consistency. We have to keep our expectations realistic: obviously any Web-based cross-linguistic sample will be skewed by cultural or economic factors, and even with the worldwide spread of smart phones, psycholinguistic experiments necessarily assume familiarity with arcane notions like quiz-taking (see, e.g., Rice, Libben, & Derwing, 2002, and their still too futuristic suggestion that non-invasive neurolinguistic methods may help).

There are also intrinsic limitations on typological research more generally. Fortunately, the historically-driven skew of the set of attested human languages (Cysouw, 2005) is not a serious challenge, since what is important in cross-linguistic experimentation is the mere fact that languages differ in type, not that these types differ in frequency (see Newmeyer, 2005, for related notions). If the cross-linguistic sample is large enough, typological redundancies can be handled by regression analysis, which is less sensitive to unbalanced distributions than are factorial analyses (Baayen, Davidson, & Bates, 2008).

A more serious challenge for quantitative typology is posed by sign languages. While many important lexical variables are just as definable in sign languages as in spoken languages, like lexical frequency (Fenlon, Schembri, Rentelis, Vinson, & Cormier, 2014) and neighborhood density (Caselli & Cohen-Goldberg, 2014), the sign/speech parameter still has to be treated categorically, and it is confounded with many other lexical variables. Whether due to the visual-manual modality (Meier, 2002) or creolization (Singleton & Newport, 2004), even historically unrelated sign languages differ much less from each other than do spoken languages (Sandler & Lillo-Martin, 2006), with very similar constraints in both phonology (e.g., Sandler, 1999) and morphology (e.g., Aronoff, Meir, & Sandler, 2005).

When conducting a meta-megastudy, we also have to be realistic about methodological consistency, since even among quiz-taking cultures, there is no single task that works identically in all languages. The picture-naming task of Bates et al. (2003) may seem ideal, since it has high ecologically validity and the same stimulus set could be used for all speaker and signer populations. Nevertheless, its uniformity is illusory. The target words' phonological forms, and often also their morphological structures, differ dramatically across languages, and even their semantics may differ unexpectedly; according to their online

database (http://www.crl.ucsd.edu/~aszekely/ipnp/7lgpno.html), the picture named "artichoke" by English speakers tended to be called 花 *huā* "flower" by Mandarin speakers (the vegetable is not sold in Taiwan, where they were tested). Such differences limit the statistical benefits of grouping trial-level responses by item.

Similar challenges exist for all tasks where target items are real words, including (one half of) the lexical decision task. Tasks like nonword naming or wordlikeness judgments, however, make it possible to present precisely the same targets to speakers of any language (though illiterates cannot be tested with nonword naming, and signers would have to be tested separately for either task). If the test languages were widely varied, as they should be in a meta-megastudy, the vast majority of stimuli would of course be rejected as unwordlike or named incorrectly, but with a large enough sample, even low response rates can yield enough data for meaningful analysis. For example, given the large number of participants and test items, even the measly 11% acceptance rate in the wordlikeness judgment task of Myers, (2015) represented over 37,000 positive responses.

Scaling Bates et al. (2003) up to true meta-megastudy size would still be worth the trouble, caveats and all. When Cohen-Goldberg (2012) reanalyzed their data to test whether phonologically similar onset and coda consonants slowed down monosyllabic word naming, he found that they did so in English, Hungarian, German, and Bulgarian, even though phonological restrictions made it impossible to test Spanish, Italian, and Mandarin. While his goal was not to compare onset-coda inhibition effects across languages, it would be relatively straightforward to do so in an expanded data set, and the results would clearly be relevant for exploring the decompositionality of syllables. We could also address this issue by quantifying cross-language differences in the relative influence of phonotactic probability or in cross-trial onset priming. All such analyses would look for interactions between fixed item-level and language-level variables, with item and language themselves treated as random variables (the former nested within the latter if cross-language stimulus matching is impossible).

Expanding Bates et al. (2003) would require detailed lexical statistics on each of the test languages, but in a properly Web-coordinated meta-megastudy all such information could be supplied by the language-specific experimenters themselves. Self-interested motivation could be inspired by stimulus-generation tools in the meta-megastudy system, whereby experimenters who upload dictionaries or corpora would be rewarded not just with the frequencies, neighborhood densities, and phonotactics probabilities of all test items, but also with the automatic extraction of stimuli with particular properties (e.g., monosyllables) and the generation of nonwords. Computations like these could be done via universal algorithms; for example, nonwords can be generated from dictionaries by first parsing words into syllables (as in Iacoponi & Savy, 2011), and then generalizing them (as in Keuleers & Brysbaert, 2010).

Worldlikeness

The issues discussed in this essay are not idle speculations: I am currently involved in the development of an actual meta-megastudy Web-based system. We call it Worldlikeness, since it is designed for the collection and sharing of wordlikeness judgments, but the basic architecture would be suitable for any number of common psycholinguistic tasks. The current version (programmed almost entirely by Tsung-Ying Chen) is in the form of a Web app built in Meteor (Coleman & Greif, 2013), which uses JavaScript to coordinate smoothly between servers and browsers. Worldlikeness is designed with three distinct types of users in mind: experimenters, participants, and researchers. The modules relating to experimenters and participants contain familiar tools for designing experiments (including automated stimulus generation), running experiments (including written, auditory, and video displays), and rewarding participants (including colorful graphic comparisons of personal statistics with group results), but crucially they also facilitate the sharing of experiments and results with other experimenters. The researcher module allows non-experimenters to explore a database of all results that experimenters have made maximally open (i.e., not restricted just to themselves or just to other experimenters). The major novelty of Worldlikeness, then, is its combination of three proven technologies: online experiments (including stimulus generation), online databases, and online researcher interaction.

Developing an actual, functional system continues to reveal unexpected challenges. For example, in our initial piloting of Worldlikeness on college-aged Mandarin speakers in Taiwan, almost 20% chose to permit only us (the original experimenters) to have access to their data. While it is heartening to know that even Web natives value their privacy, this kind of response it is somewhat problematic for the scientific goals of meta-megastudies. Whatever method is found to encourage greater participant data sharing, it must of course be consistent with the multitude of independent review boards that would have to approve a meta-megastudy as it slowly grows across the experimenter network.

One might object that a Web-based meta-megastudy system merely replace one form of centralization with another; after all, we currently have sole control over the Worldlikeness server. A meta-megastudy system could be readily decentralized, however, for example by mirroring or even torrenting it. More importantly, given that Worldlikeness is open-source, we would be satisfied if it ended up living on primarily through the future meta-megastudy systems that it helps inspire.

References

Alario, F. X., Perre, L., Castel, C., & Ziegler, J. C. (2007). The role of orthography in speech production revisited. *Cognition, 102*(3), 464-475.

Aronoff, M., Meir, I., & Sandler, W. (2005). The paradox of sign language morphology. *Language, 81*(2), 301-344.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory & Language, 44*, 569-591.

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.

Balota, D. A., Yap, M. J., Hutchison, K.A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed). *Visual word recognition, Vol. 1* (pp. 90-115). London: Psychology Press Psychology Press.

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., Herron, D., Lu, C-C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., Tzeng, A., & Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review, 10* (2), 344-380.

Bates, E., Devescovi, A., & Wulfeck B. (2001). Psycholinguistics: A cross-language perspective. *Annual Review of Psychology, 52*, 369-96.

Benkí, J. R. (2003). Quantitative evaluation of lexical status, word frequency, and neighborhood density as context effects in spoken word recognition. *The Journal of the Acoustical Society of America, 113*(3), 1689-1705.

Bishop, S. (2014). Science exposed. *Scientific American, 311*(4), 46.

Cai, Z.-Q. (Ed.) (2008). *How to read Chinese poetry: A guided anthology*. New York: Columbia University Press.

Caselli, N. K., & Cohen-Goldberg, A. M. (2014). Lexical access in sign language: a computational model. *Frontiers in Psychology, 5*, 428. http://doi.org/10.3389/fpsyg.2014.00428.

Chen, J.-Y., Chen, T.-M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language, 46*(4), 751-781.

Chen, T.-M., Dell, G., & Chen, J.-Y. (2007). A cross-linguistic study of phonological units: Syllables emerge from the statistics of Mandarin Chinese, but not from the statistics of English. *Chinese Journal of Psychology, 49*(2), 137-144.

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

Clauson, K. A., Polen, H. H., Boulos, M. N. K., & Dzenowagis, J. H. (2008). Scope, completeness, and accuracy of drug information in Wikipedia. *Annals of*

Pharmacotherapy, 42(12), 1814-1821.

Cohen-Goldberg, A. M. (2012). Phonological competition within the word: Evidence from the phoneme similarity effect in spoken production. *Journal of Memory and Language, 67*(1), 184-198.

Coleman, T., & Greif, S. (2013). *Discover Meteor*. http://www.discovermeteor.com/

Costa, A., Alario, F. X., & Sebastián-Gallés, N. (2007). Cross-linguistic research on language production. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 531-546). Oxford: Oxford University Press.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition, 10*, 65-70.

Cutler, A. (1985). Cross-language psycholinguistics. *Linguistics, 23*, 659-667.

Cysouw, M. (2005). Quantitative methods in typology. In R. Kohler, G. Altmann, & R. G. Piotrowski (Eds.) *Quantitative Linguistik: Ein internationales Handbuch* [Quantitative linguistics: An international handbook] (pp. 554-578). Berlin: Walter de Gruyter.

Erlewine, M. Y., & Kotek, H. (2015). A streamlined approach to online linguistic surveys. *Natural Language and Linguistic Theory*. doi:10.1007/s11049-015-9305-9

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32* (5), 429-492.

Fenlon, J., Schembri, A., Rentelis, R., Vinson, D., & Cormier, K. (2014). Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. *Lingua, 143*, 187-202.

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition, 28*(7), 1109-1115.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language, 47*(1), 27-52.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101*(2), 366-385.

Haspelmath, M., Dryer, M.S., Gil, D., & Comrie, B. (Eds.) (2005). *The world atlas of language structure*. Oxford: Oxford University Press.

Heylighen, F. (2007). Why is open access development so successful? Stigmergic organization and the economics of information. In Lutterbeck, B., Bärwolff, M., & Gehring, R. A. (Eds.) *Open Source Jahrbuch 2007*. Berlin: Technical University of Berlin.

Iacoponi, L., & Savy, R. (2011). Sylli: Automatic phonological syllabification for Italian. *INTERSPEECH 2011*, 641-644.

Jaeger, T. F., & Norcliffe, E. J. (2009). The cross-linguistic study of sentence production. *Language and Linguistics Compass, 3/4*, 866-887.

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology, 68* (8), 1457-1468.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*(3), 627-633.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Frontiers in Psychology, 1*, 174.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44*(1), 287-304.

Kirby, J. P., & Yu, A. C. L. (2007). Lexical and phonotactic effects on wordlikeness judgments in Cantonese. *Proceedings of the International Congress of Phonetic Sciences, 16*, 1389-1392.

Kupferberg, N., & Protus, B. M. (2011). Accuracy and completeness of drug information in Wikipedia: An assessment. *Journal of the Medical Library Association, 99*(4), 310-313.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(01), 1-38.

Malmqvist, G. (1994). Chinese linguistics. In G. Lepschy (Ed.), *History of linguistics: Volume I: The Eastern traditions of linguistics* (pp. 1-24). London: Longman.

McBride-Chang, C., Tong, X., Shu, H., Wong, A. M. Y., Leung, K. W., & Tardif, T. (2008). Syllable, phoneme, and tone: Psycholinguistic units in early Chinese and English word recognition. *Scientific Studies of Reading, 12*(2), 171-194.

Meier, R. P. (2002). Why different, why the same? Explaining effects and non-effects of modality upon linguistic structure in sign and speech. In R. P. Meier & K. Cormier (Eds.) *Modality and structure in signed and spoken languages* (pp. 1-25). Cambridge, UK: Cambridge University Press.

Moscoso del Prado Martín, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition, 94*(1), 1-18.

Myers, J. (2012). Chinese as a natural experiment. In G. Libben, G. Jarema, & C. Westbury (Eds.), *Methodological and analytic frontiers in lexical research* (pp. 155-169). Amsterdam: John Benjamins.

Myers, J. (2015). Markedness and lexical typicality in Mandarin acceptability judgments. *Language & Linguistics, 16* (6).

Nathan, D. (2013). Access and accessibility at ELAR, a social networking archive for endangered languages documentation. In M. Turin, C. Wheeler & E. Wilkinson (Eds.) *Oral literature in the digital age: Archiving orality and connecting with communities*.

Cambridge, UK: Open Book Publishers. Website URL: http://www.elar-archive.org/index.php.

Newmeyer, F. J. (2005). *Possible and probable languages: A generative perspective on linguistic typology*. Oxford: Oxford University Press.

Nielsen, M. (2012). *Reinventing discovery: The new era of networked science*. Princeton, NJ: Princeton University Press.

Norris, D., & Cutler, A. (1988). The relative accessibility of phonemes and syllables. *Perception & Psychophysics, 43*(6), 541-550.

O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition, 115*, 282-302.

O'Seaghdha, P. G., Chen, J. Y., & Chen, T. M. (2013). Close but not proximate: The significance of phonological segments in speaking depends on their functional engagement. *Proceedings of the National Academy of Sciences, 110*(1), E3.

Qu, Q., Damian, M. F., & Kazanina, N. (2012). Sound-sized segments are significant for Mandarin speakers. *Proceedings of the National Academy of Sciences, 109*(35), 14265-14270.

Qu, Q., Damian, M. F., & Kazanina, N. (2013). Reply to O'Seaghdha et al.: Primary phonological planning units in Chinese are phonemically specified. P*roceedings of the National Academy of Sciences, 110*(1), E4.

Rastle, K., McCormick, S. F., Bayliss, L., & Davis, C. J. (2011). Orthography influences the perception and production of speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1588-1594.

Rice, S., Libben, G., & Derwing, B. (2002). Morphological representation in an endangered, polysynthetic language. *Brain and Language, 81*(1), 473-486.

Sampson, G. (2015). A Chinese phonological enigma. *Journal of Chinese Linguistics, 43*, 679-691.

Sandler, W. (1999). Cliticization and prosodic words in a sign language. In T. A. Hall and U. Kleinhenz (Eds.) *Studies on the phonological word* (pp. 223-255). Amsterdam: John Benjamins.

Sandler, W., & Lillo-Martin. (2006). *Sign language and linguistic universals*. Cambridge, UK: Cambridge University Press.

Savin, H. B., & Bever, T. G. (1970). The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior, 9*(3), 295-302.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution, 24*(9), 467-471.

Singleton, J. L., & Newport, E.L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*

49, 370-407.

Sze, W. P., Liow, S. J. R., & Yap, M. J. (2014). The Chinese Lexicon Project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods, 46*(1), 263-273.

Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods, 42*(4), 992-1003.

Yu, M., Mo, C., Li, Y., & Mo, L. (2015). Distinct representations of syllables and phonemes in Chinese production: Evidence from fMRI adaptation. *Neuropsychologia, 77*, 253-259.

Yu, M., Mo, C., & Mo, L. (2014). The role of phoneme in Mandarin Chinese production: Evidence from ERPs. *PloS one 9* (9), e106486.