Meta-Megastudies

James Myers
National Chung Cheng University

Published as:
Myers, J. (2016). Meta-megastudies. *The Mental Lexicon, 11*(3), 329-349.

Address for correspondence:
James Myers
Graduate Institute of Linguistics
National Chung Cheng University
Min-Hsiung, Chiayi 62102
Taiwan
Phone: 886-5-242-8251
Fax: 886-5-272-1654
Email: Lngmyers@ccu.edu.tw

Abstract

Cross-linguistic data have always been of interest to mental lexicon researchers, but only now are technological developments beginning to make it possible to treat language as a random variable, in an approach we dub meta-megastudies. A meta-megastudy uses regression techniques to tease apart not just factors that are partially confounded across items within languages, as in traditional megastudies, but also factors partially confounded across languages. While large-scale meta-megastudies will be logistically challenging, they promise great theoretical benefits and are becoming ever more feasible via Web-based coordination between independent research groups.

Keywords: megastudies; 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 sense (where the "language" being "fixed" is actually a set of test items): 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 should one be able to generalize about human word processing by testing a large sample of human languages. I call this a meta-megastudy.

The rest of this essay is devoted to unpacking this simple idea. Befitting the theme of this special issue, my goal is to raise questions for the next decade, not to answer them. I first show how a more sophisticated approach to cross-linguistic research would benefit psycholinguistics by providing an increased ability to disentangle partially confounded variables, and then show how such an approach could be made practical by exploiting the Web to coordinate the running of independent experiments and the sharing of results.

The Need for Meta-Megastudies

The core logic is simple: if megastudies are good for the study of individual languages, then meta-megastudies are good for the study of cross-language differences. By now the benefits of megastudies to lexical research are well known (Balota et al., 2012; Keuleers & Balota, 2015). A lexical experiment can be thought of as an attempt to study how processing is affected by lexical variables, under various conditions. The problems are that lexical items already exist and so cannot be experimentally manipulated, that the variables describing them may not be categorical (e.g., grammatical vs. content words) but gradient (e.g., lexical frequency), and that they may also be partially confounded with each other (e.g., grammatical words tend to have higher frequencies than content words). Forcing a factorial design by cherry-picking subsets of matched items risks experimenter bias (Forster, 2000) without eliminating confounds with uncontrollable or unknown variables (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, particularly regression, to tease apart partial confounds. Megastudy databases exist in an ever-growing number and variety of languages, including English (Adelman et al., 2014; Balota et al., 2007; Hutchison et al., 2013; Keuleers, Lacey, Rastle, & Brysbaert, 2012), Dutch (Ernestus & Cutler, 2015; Keuleers, Diependaele, & Brysbaert, 2010), French (Ferrand et al., 2010), Chinese (Sze, Liow, & Yap, 2014; Myers, 2015; Tse, Yap, Chan, Sze, Shaoul, & Lin, forthcoming), and Malay (Yap, Liow, Jalil, & Faizal, 2010).

Nevertheless, within any particular megastudy, the language itself remains a preexisting, nonmanipulable independent variable. Thus if we want to study how processing is affected by experience with a language as a whole, cross-linguistic comparisons seem like a relevant source of information. Psycholinguists have long recognized the importance of cross-linguistic studies (Bates, Devescovi, & Wulfeck, 2001; Costa, Alario, & Sebastián-Gallés, 2007; Cutler, 1985; Evans & Levinson, 2009; Jaeger & Norcliffe, 2009; Norcliffe, Harris, & Jaeger, 2015); at the time of this writing, ProQuest's Language and Linguistics Behavior Abstracts shows over 9,500 hits for psycholinguistic* AND (cross-linguistic* OR typolog*). Yet many cross-linguistic studies are concerned less with typology than with within-speaker multilingualism (including some that use megastudy methods, e.g., Keuleers, Stevens, Mandera, & Brysbaert, 2015; Lemhöfer et al., 2008), and of the studies that do compare separate speech communities, virtually all involve just two languages, presumably for practical reasons. The challenges of cross-linguistic psycholinguistics are illustrated by the unusually ambitious study of Bates et al. (2003): collecting naming responses on 520 object pictures from speakers of seven different languages (Bulgarian, English, German, Hungarian, Italian, Mandarin, Spanish) required an astonishing 22 authors, affiliated with 10 different institutions.

Just like the lexical items within a language, languages differ from each other in many different variables, some categorical (e.g., whether the language is written with an alphabetic orthography) and some gradient (e.g., the size of the syllable inventory), and these variables may also be partially confounded (e.g., syllable-based orthographies seem to be restricted to languages with relatively small syllable inventories: DeFrancis, 1989). The key justification for meta-megastudies is thus essentially arithmetical: just as for single-language megastudies, running a regression across multiple data points is a much more effective way to disentangle partial confounds than testing two data points at a time.

Suppose there are ten cross-linguistic variables at play in some psycholinguistic debate. According to a common rule of thumb (e.g., Vittinghoff & McCulloch, 2007), data from around 100 languages (ten times the number of variables) may suffice to tease them apart in a regression analysis, even if correlated (O'Brien, 2007). While collecting data from so many languages is clearly beyond the powers of a conventional research team, the burden could be split across 100 separate teams, each generating its own single-language database for its own purposes, and the researchers who use the entire cross-linguistic database to study the ten variables may not even be any of the original data collectors. This is in fact the standard procedure in linguistic typology, where researchers use compilations of existing grammars to test hypotheses that the grammar writers themselves may never have dreamed of (e.g., Haspelmath, Dryer, Gil, & Comrie, 2005). This procedure has proven so reliable that many of the observations first made tentatively on the basis of small cross-language samples (Greenberg, 1963) still hold up in more sophisticated analyses of much larger databases (e.g., Dryer, 1992, 2011).

By contrast, a conventional pairwise exploration of this same language sample would be both more burdensome and less satisfying. While roughly the same number of researchers would be required (at least one native speaker per language), they would have to be in intimate collaboration throughout, from design through write-up; otherwise we would merely have the current status quo of ad hoc two-language studies with incommensurate goals and methods. Using pairwise language comparison to test one cross-linguistic variable also requires keeping all other variables constant, which is impossible except, perhaps, for closely related dialects, or for particular subpopulations of a speech community, like literate versus illiterate participants in an orthographic study. Treating language as a categorical factor, as the traditional approach does, would also force artificial discretization on naturally gradient variables (e.g., dividing languages into those with small vs. large syllable inventories).

Of course, traditional cross-linguistic methods continue to provide rich insights into language differences and similarities, and the testing of universalist processing claims (i.e., about how the human mind naturally responds to particular language stimuli) would also benefit from greater attention to other approaches, including searching for defaults in first language acquisition or artificial grammar learning (Myers, 2012). The goal of this essay is merely to highlight the potential benefits of adding meta-megastudies to an ever-expanding toolkit.

This tool is not even entirely new. The aforementioned Bates et al. (2003), with its uniform picture-naming task applied to seven languages, already provides a database ready-made for meta-megastudies, as is the set of lexical decision and written word naming megastudies cited above (five languages and counting). Although the currently available cross-language sample is still small and typologically limited, the potential is already there.

The Challenges of Meta-Megastudies

Meta-megastudies aim to deal with one challenge, confounding in cross-linguistic variables, by taming another, cross-language sampling. In this section I address each in turn; challenges regarding infrastructure are saved for a later section.

Confounded Variables

One sign of the tendency to neglect confounding in cross-linguistic studies is that even though many such studies involve the same two languages, they ascribe their results to different variables, trusting the experimental design to make all other variables irrelevant. Studies on Chinese, for example, often compare it with English, but the observed processing differences are variously ascribed to differences in orthography (e.g., Feng, Miller, Shu, & Zhang, 2001), the phonemic status of tone (e.g., Klein, Zatorre, Milner, & Zhao, 2001), syllable inventory size (e.g., O'Seaghdha, Chen, & Chen, 2010), or the dominance of compounding in word formation (e.g., Myers, 2007). Of course it is not unreasonable to ascribe specific processing differences to specific language characteristics, but this reasoning itself depends on universalist assumptions. The idea that orthographic experience might influence tone perception may be implausible a priori, but language samples diverse enough to pull apart these variables (e.g., by adding Japanese, a non-tonal language with a Chinese-like orthography) would help settle the point empirically. For other variables, universalist processing models are not yet articulated enough to tell us what to expect; for example, to find out if compounding affects syllable processing or vice versa, our best bet may be to head into the (cross-linguistic) lab.

In some cases there may be several equally plausible explanatory variables. In the remainder of this subsection I drive home this point with a single case study: the apparent cross-linguistic differences in the decomposition of syllables into phonemes (extending the discussion of Myers, 2012). Namely, 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 perception (e.g., Benkí, 2003; Norris & Cutler, 1988) and in production (e.g., Fromkin, 1971; O'Seaghdha et al., 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 classified syllables in terms of onset consonants and rimes for around 1,500 years (Malmqvist, 1994), Chinese poets have long used alliteration (Cai, 2008), and children in modern Chinese-speaking communities are taught syllable-decomposing orthography (e.g., Pinyin). The fact that Chinese speakers are capable of syllable decomposition does not show that they do so readily or routinely, however. Chinese characters themselves represent monosyllabic (and monomorphemic) units, not phoneme-sized units, and while a child's ability to detect word onsets helps when learning to read English, this is not true for children learning to read Chinese (McBride-Chang et al., 2008).

Adult Chinese speakers also show a surprising indifference towards the phoneme. Chen, Chen, & Dell (2002) and O'Seaghdha et al. (2010) found no reaction time effects in Mandarin due to onset consonant priming in a form preparation (implicit priming) task, as compared with the robust effects in English also reported by O'Seaghdha et al. (2010). Similarly, in a picture-word interference naming study on Cantonese, Wong and Chen (2008) failed to observe the facilitation due to shared onset consonants found in languages like Dutch (Levelt, Roelofs, & Meyer, 1999). Qu, Damian, & Kazanina (2012) also failed to find onset priming in a task naming pictures of colored objects, where the monosyllabic color name and the disyllabic object name did or did not share the initial consonant, and Yu, Mo, & Mo (2014) found none in a picture naming task where disyllabic object names did or did not have the same onset consonant in both syllables.

As noted earlier, phonemes are not entirely ignored in Chinese, and Yu and Shu (2003) buck the trend by reporting onset priming in a Mandarin picture-word interference task. Moreover, two of the studies mentioned above, Qu et al. (2012) and Yu et al. (2014), both found significant phoneme priming in event-related potentials (ERP). In an fMRI study dispensing with behavioral measures entirely, Yu, Mo, Li, & Mo (2015) had participants read aloud pairs of nonsense disyllables that shared syllable onsets, entire syllables (but not characters), or nothing, finding brain activation for the onset repetition condition, in different areas from those for the syllable repetition condition. The authors of these three neuroimaging studies ascribe the paucity of behavioral evidence for onset priming to inhibition at a late self-monitoring stage that is intended to reduce the risk of consonant errors. Nevertheless, since they see such inhibition as universal, they must also explain why overt segment priming occurs in English anyway, ascribing it to intrinsically stronger phoneme activation in this language. Thus despite disagreements in detail (cf. O'Seaghdha, Chen, & Chen, 2013; Qu, Damian, & Kazanina, 2013), the consensus remains that Chinese word production differs from that of English because the former is dominated by syllables, not phonemes.

In spoken Chinese word recognition, listeners clearly rely on subsyllabic information (e.g., Gandour, Xu, Wong, Dzemidzic, Lowe, Li, & Tong, 2003; Malins & Joanisse, 2010), but the methods used so far are not designed to distinguish phonemic from acoustic processing. Relevant techniques are available; for example, Benkí (2003) quantified the perceptual independence of subsyllabic components in English by analyzing the patterns of identification errors in noise, but his method does not seem to have been used yet in Chinese. Currently all we have are intriguing hints from independent studies using different methods: while Tseng, Huang, and Jeng (1996) report that Mandarin listeners were *faster* to detect onset consonants in lexical than in nonce syllables, Fox (1984) found, in a categorical perception study on English onset consonants, that the influence of syllable lexical status increased with *slower* response times.

Finally, wordlikeness judgment experiments suggest that Chinese speakers are much less comfortable than English speakers with expanding their existing syllable inventory via novel combinations of phonemes. The spoken and written nonword wordlikeness judgments collected from English speakers by Bailey and Hahn (2001) imply an overall mean acceptance score around .42 (on a zero-to-one scale, transformed from a nine-point Likert scale) for 272 randomly generated nonce English monosyllables (differing from real syllables by one or two phonemes), as if the participants recognized that many English-like syllables can be generated by freely combining English segments. By contrast, in a wordlikeness megastudy on Mandarin speakers, Myers (2015) reports an overall acceptance rate of only .11 (in a binary judgment task) for 3,274 nonlexical syllables generated by combining Mandarin onsets, prevocalic glides, rimes, and tones (96% differing from real syllables by just one of these elements), presented in phonological transcription (Taiwan's equivalent of Pinyin) to avoid misperception. Methodological differences across these studies seem insufficient to explain such a large difference in results; Bailey and Hahn (2001) and Myers and Tsay (2005) found very little effect of modality (speech vs. writing) on wordlikeness judgments in English and Mandarin, respectively, and Bader and Häussler (2010) and Weskott and Fanselow (2011) show that conclusions drawn from linguistic acceptability judgments do not depend on response scale (whether binary, Likert, or continuous magnitude estimation).

Moreover, neighborhood density, which involves comparisons with entire words, and phonotactic probability, which depends on decomposition into phonemes and phoneme strings (Luce & Large, 2001), affect wordlikeness roughly equally in English; Bailey and Hahn (2001) report that phonotactic and orthotactic probabilities together accounted for about 17% of response variance for their auditory nonce syllables, with neighborhood density accounting for an additional 14% of the variance (as computed in a multiple regression taking both variables into account simultaneously; personal communication, T. Bailey, 14 September, 2016). By contrast, for spoken Cantonese wordlikeness judgments of 270 nonce syllables on a seven-point Likert scale, Kirby and Yu (2007) found that neighborhood density accounted for around 33% of the response variance, while phonotactic probability accounted for less than 2% of additional variance (again computed in a multiple regression taking both variables into account simultaneously; personal communication, J. Kirby, 14 September, 2016). A similar asymmetry can be seen in the Mandarin wordlikeness database used by Myers (2015) (available at http://lngproc.ccu.edu.tw/MWP): a mixed-effects logistic regression model with random slopes predicting binary response choice from log neighborhood density and log mean bigram conditional probability (both phoneme-based, ignoring tone) yields a larger standardized coefficient for neighborhood density ($\beta = 0.98$, $SE = 0.05$, $z = 17.86$, $p < .001$)

than for phonotactic probability (β= 0.16, *SE* = 0.02, *z* = 6.00, *p* < .001). This difference in coefficients is significant by a likelihood ratio test comparing the above additive model with one assuming a single coefficient for the sum of the fixed variables, algebraically the same as an additive model with identical coefficients for both variables ($\chi^2(4) = 2086.2$, *p* < .001).

The reason for cross-language differences in syllable decomposition may seem obvious: English orthography spells out consonants and vowels, but Chinese orthography does not. While the evidence that experience with an orthographic system affects speech processing is mixed (Alario, Perre, Castel, & Ziegler, 2007; Rastle, McCormick, Bayliss, & Davis, 2011), it may be more relevant for certain tasks than for others; in Chinese in particular, Bi, Wei, Janssen, and Han (2009) found that characters affected spoken word form preparation only when the characters were actually read aloud. Nevertheless, focusing exclusively on orthography 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 a possible explanation for cross-linguistic differences in syllable decomposition.

These two explanations, in turn, relate to a third and a fourth: morpheme size and syllable inventory size. Homophones arise so readily in Chinese because virtually all Chinese morphemes are monosyllabic, and there just are not that many Chinese syllables to go around: 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 et al., 1999). This makes it ten times more feasible for Mandarin speakers simply to memorize their syllable inventory, even if they also weakly compose syllables out of, or decompose them into, phonemes, and even if Dutch speakers also memorize their top few hundred most common syllables, as Levelt et al. (1999) suggest.

Meanwhile, Kirby and Yu (2007) explain their Cantonese wordlikeness findings (where phonotactic probability played a much smaller role than in English) in terms of a fifth difference: the Cantonese lexicon uses much more of the syllable space defined by its subsyllabic components (36%) than does English (6%). The same applies to Mandarin: the 3,274 nonlexical syllables tested in Myers (2015) were chosen from all 4,516 logically possible combinations of onset, medial, rime, and tone, resulting in a 28% probability of hitting a real Mandarin syllable through random combinations of these subsyllabic components alone. If logically possible syllables are too likely to be lexical, decomposition into phoneme strings becomes an unhelpful strategy in making wordlikeness judgments (aside from the trivial benefit of detecting non-native phonemes, which could be done via low-level acoustic processing anyway).

Chinese also differs from English in banning consonant clusters and complex rimes; Chinese traditional phonological theory even treats the rime as a whole, an analysis supported by the highly restricted phonotactics of VX sequences (Light, 1977). More generally, as Chen, Dell, and Chen (2007) show, phoneme sequences are statistically more predictable in the Mandarin lexicon than in the English lexicon, making it more parsimonious to process them together, as they demonstrate in a connectionist model.

This makes six confounded variables (orthography, homophony, morpheme size, syllable inventory size, ratio of lexical to possible syllables, phoneme predictability), though for thoroughness we might also want to add the less obviously relevant variables alluded to earlier (lexical tone and compounding). It matters for psycholinguistic theory which of these variables prove to be truly explanatory for particular processing differences; even among the phonological variables, recall that Yu et al. (2015) found that onset repetition and syllable repetition activate distinct brain regions. These observations make it unlikely that cross-linguistic confounding can be avoided simply by recoding all of the variables as a single predictor (cf. Moscoso del Prado Martín, Kostić, & Baayen, 2004, in their replacement of lexical variables like type and token frequencies by a single measure of informational complexity).

Fortunately, the intrinsic distinctness of these six variables also implies that their tight confounding in English and Chinese should show cracks in a sufficiently large and varied cross-linguistic sample. 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 *et al.*, 2005) includes the parameters of 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 related to these.

I have no 51-language meta-megastudy to report here, but the shape that such a study may take is suggested by an analysis by Cohen-Goldberg (2012) of the seven-language picture naming database of Bates et al. (2003). Using the publicly available by-item means (Szekely et al., 2004: https://crl.ucsd.edu/experiments/ipnp/7lgpno.html), Cohen-Goldberg found that the presence of phonologically similar onset and coda consonants slowed down monosyllabic word naming in a data sample combining Bulgarian, English, German, and Hungarian responses. Clearly the phenomenon of onset-coda competition can shed light on cross-linguistic variation in syllable decomposition, but this was not the focus of Cohen-Goldberg' study. His selection criteria led him to exclude the other three languages in the database: Spanish and Italian (too few monosyllabic words) and unfortunately also Mandarin (its phonemically contrastive aspiration did not fit with his uniform segment coding scheme). The four languages that he did test were also treated as a fixed rather than random variable and no interactions with this variable were tested, so we cannot say whether the onset-coda competition effect varies systematically across these languages. Nevertheless, this data set has potential. For example, if trial-level data were also made available, cross-trial onset priming could be quantified and tested for interactions with language; the literature review above would lead us to expect particularly weak priming in Mandarin. It may even be possible to detect correlations between the strength of onset priming and language-level variables like syllable inventory size, though of course seven languages are far too few to tease apart all six of the variables discussed above.

Typological Language Sampling

Even after acknowledging the reality of cross-linguistic confounds, it may seem that the hard work has only just begun: choosing a sample of languages to compare, from the around 7,000 (Lewis, Simons, & Fennig, 2014) on the planet. After all, typological linguists expend considerable effort on sampling to ensure that inferences about the human language faculty are not skewed by historical descent or borrowing from geographical neighbors (Bickel, 2015; Cysouw, 2005).

Fortunately, meta-megastudies of processing are sufficiently different from typological studies of grammar to make language selection far simpler in the former case. First and foremost, unlike grammars, psycholinguistic processes are essentially automatic, not learned social conventions. Typologists only worry about sampling because the learning of grammatical features is enforced by speech communities, and thus can spread by descent or borrowing. By contrast, it is hard to see how psycholinguistic processes per se could be subject to learned social conventions (e.g., requiring members of a speech community to show a particular kind or degree of lexical frequency effect). Thus if some linguistic feature affects processing, we expect speakers of all languages with this feature to show roughly the same processing effect (aside from interactions with other features), whether or not the languages are related or geographical neighbors.

Moreover, even typologically rare language features (e.g., Chinese orthography) are revealing about general human cognition through their mere existence (see Newmeyer, 2005, for related notions). As long as the cross-language sample provides enough information about the independent variables of interest, typologically unbalanced distributions should not undermine the inferential power of regression analyses (Baayen, Davidson, & Bates, 2008).

Finally, the vast size of single-language megastudies already shows that psycholinguists consider diversity and comprehensiveness more important than representativeness. Due to history and borrowing, any extant lexicon provides as skewed a picture of a native speaker's powers of lexical productivity as extant human languages do of the human language faculty. This lexical skew bothers nobody; as Clark (1973) advised, items are routinely treated as a random variable.

Even with advances in technology and education, the scope of meta-megastudies will likely remain limited by cultural or economic factors, including access to laboratory-quality equipment and familiarity with arcane notions like quiz-taking (cf. Rice, Libben, & Derwing, 2002, and their still too futuristic suggestion that non-invasive neurolinguistic methods may help). Diversity and comprehensiveness also require taking sign languages into account. 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 is intrinsically 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).

Infrastructure for Meta-Megastudies

While the advent of meta-megastudies may be as inevitable as the steady spread of megastudies across ever more languages, this spread would be more efficient, and the component studies more useful for meta-analysis, with improved infrastructure. To put it crudely, I think the key to speeding the growth of meta-megastudies is to exploit the Web 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, 2016) run experiments on the Web, collecting hundreds or thousands of responses from hundreds or thousands of participants. Technological advances are making megastudies ever easier to run, with Web apps like Science XL (http://www.sciencexl.org/home; Dufau et al., 2011) and Tatool (http://www.tatool.ch; von Bastian, Locher, & Ruflin, 2013) supplementing private lab-written Web systems or proprietary services like Amazon Mechanical Turk (http://www.mturk.com) and Cambridge Brain Sciences (http://www.cambridgebrainsciences.com).

Citizen science, by contrast, 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 given to the contributors. Meta-megastudies must distribute responsibility because they only become feasible when not only the raw response data, but the experiments themselves, are contributed by large numbers of people. The consequence is that meta-megastudies require a much less centralized working style than scientists are generally familiar with (Nielsen, 2012, suggests that this is the future of science more generally).

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; Heylighen, 2007), albeit with 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 Web experimentation with those underlying Web databases. Meta-megastudy infrastructure would aim at achieving four key goals. First, it should make it easier for new language data to be added while maintaining methodological consistency. This could be achieved via a Web app with built-in functions for designing and running only a limited number of experiment types (e.g., lexical decision and picture naming). Many languages have never been studied experimentally at all; if ready-made tools embolden novice experimenters to contribute data on their native language, typological psycholinguistics would benefit enormously, whether or not the individual contributions are large enough to constitute megastudies themselves.

Second, the system should ensure data quality and ethical standards. This could be achieved using the tried-and-true Web approach of peer vetting. For example, the system could require experimenters to identify their credentials (e.g., via a link to a university homepage), visible to other experimenters and their own participants, and there could also be a feedback system whereby experimenters are publicly rated by their participants and fellow experimenters for ethical behavior and linguistic competence. If all else fails, transparently appointed moderators could lay down the law on particularly egregious offenders.

Third, the system should assist non-native-speaking typologists in their analyses. This could be achieved by encouraging experimenters to share not only their stimuli and results, but also key lexical variables, both item-level and language-level. For example, experimenters who upload pronunciation dictionaries could be rewarded with the automated calculation of phonological neighborhood densities and phonotactic probabilities. The words could also be parsed into syllables using universal algorithms (as in the Sylli tool of Iacoponi & Savy, 2011: http://sylli.sourceforge.net), which, besides providing the language's syllable inventory, would allow the generation of nonwords for tasks like lexical decision and wordlikeness via a syllable-based bigram-chain grammar (as in the Wuggy tool of Keuleers & Brysbaert, 2010: http://crr.ugent.be/programs-data/wuggy); generating written nonwords might be further simplified with phoneme-to-grapheme conversion (Rentzepopoulos & Kokkinakis, 1996). Automated morphological parsing may also be possible (e.g., Durrett & DeNero, 2013) for generating morphologically complex nonwords. Other algorithms could compute cross-linguistic variables like phoneme frequencies, starting with public databases (e.g., that built for the Automated Similarity Judgment Program by Brown, Holman, &

Wichmann, 2013: http://asjp.clld.org) but gradually incorporating the lexicons contributed by other experimenters in the meta-megastudy system.

Finally, even though the system could be used solely as a private psycholinguistic cloud service, experimenters who choose to share their materials and results would provide an extra contribution visible to all, hopefully inspiring other experimenters to follow suit. Altruism might receive a further nudge through technical means, such as by making an experimenter's quota for running new experiments dependent on sharing previous ones.

To test the feasibility of such a meta-megastudy system, Tsung-Ying Chen and I have been developing a Web app aimed at implementing as many of the above goals as we can. We call it Worldlikeness, since it was originally designed for the collection and sharing of wordlikeness judgments across languages. Built in the Javascript-based Meteor platform (Coleman & Greif, 2013), the open beta version of Worldlikeness is currently hosted at http://www.worldlikeness.org/; it runs in any modern browser, including on mobile devices. As a Web experiment system, it helps experimenters design and run simple non-factorial experiments, and rewards participants with colorful graphic comparisons of personal statistics with group results. Experimenters need to register and provide contact information for their participants, who do not need to register at all, and both types of users control access to their data. As a Web database system, Worldlikeness facilitates the sharing of experimental materials and results, including previously collected data uploaded to the system. Depending on the privacy settings of the data providers, results may be available only to the original experimenter and the experimenter's collaborators, all other registered experimenters, or the public at large. Data sharing is encouraged via a quota system like that sketched above, among other devices. Researchers wanting to perform typological analyses across many languages can download all of the data available to them, or just search for a particular subset of interest; typological researchers thus need not be experimenters themselves. Among our key goals, automating tools for typological research has proven to be the most challenging; we are still working out the kinks in our algorithms for stimulus generation and language type marking.

Nevertheless, Worldlikeness has already proven itself in experiments on Mandarin and Taiwan Southern Min (Myers & Chen, 2016), and on Taiwan Sign Language (Lee, 2016), with the results available for download from the Worldlikeness database. As repeatedly emphasized in this essay, no single research team can contribute enough languages to create a meta-megastudy database on its own, so we are actively soliciting contributions. While wordlikeness is not a particularly popular task in psycholinguistics, the Worldlikeness app is already set up to be used for the lexical decision task as well. Not much modification would be necessary to adapt it for other common megastudy tasks, like written word naming or picture naming: Worldlikeness is not only free but open-source, and readers are more than welcome to repurpose its code.

Conclusions

Are phonemes processing units in spoken word processing? What determines when readers access word meanings via phonological recoding or directly from the written form? How is morphological decomposition affected by semantic transparency? Quantitatively precise answers to theoretically significant psycholinguistic questions depend not just on the particular experimental design or particular test items, but also on the particular language: psycholinguistics is not psychophysics, and language experience matters. Yet two-language studies are not enough: languages differ from each other in too many ways to be sure that any cross-language result must be due to one factor and not another.

Meta-megastudies simply put a fancy name on a simple idea: generalizing the big data approach of single-language megastudies, where the variable confound problem has long been recognized, to cross-linguistic psycholinguistics. If the meta-megastudy approach could be made feasible, the benefits would clearly be enormous: claims about processing universals could be tested the same way typological linguists test claims about grammatical universals. Rather than observing a stimulus-response pattern in one language and assuming things work exactly the same way in all languages unless proven otherwise, stimuli and responses could be studied across a wide variety of languages and the general relationship induced via statistical modeling, with fewer sampling headaches than in typological studies of grammar.

Fortunately, technological advances are indeed making meta-megastudies feasible, the same advances that sparked the megastudy revolution itself. The computing and networking powers of the Web enable the development of systems that link experimenters not just with their participants, but with each other, using automation to ensure methodological consistency and to reward contributors with perks like stimulus creation, and Web-mediated social pressure to enforce ethics and quality. Psycholinguists know that cross-linguistic research is essential to understanding the full richness of the human mind, but the vast majority of the world's languages are still missing from this research. Give citizen scientists the right tools, and watch what happens.

Author note

This study was supported by Ministry of Science and Technology (Taiwan) grant 103-2410-H-194-119-MY3. Many thanks to my lab assistants, the editors and anonymous reviewers, and Tsung-Ying Chen.

References

Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., Perry, J. R., Lupker, S. J., Forster, K. I., Cortese, M. J., Scaltritti, M., Aschenbrenner, A. J., Coane, J. H., White, L., Yap, M. J., Davis, C., Kim, J., & C. J. Davis. (2014). A behavioral database for masked form priming. *Behavior Research Methods, 46*(4), 1052-1067.

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.

Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics, 46* (2), 273-330.

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.

Bi, Y., Wei, T., Janssen, N., & Han, Z. (2009). The contribution of orthography to spoken word production: Evidence from Mandarin Chinese. *Psychonomic Bulletin & Review, 16*(3), 555-560.

Bickel, B. (2015). Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In B. Heine & H. Narrog (Eds). *The Oxford Handbook of Linguistic Analysis, 2nd edition* (pp. 901-923). Oxford: Oxford University Press.

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

Brown, C. H., Holman, E. W., & Wichmann, S. (2013). Sound correspondences in the world's languages. *Language, 89* (1), 4-29.

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.

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*. URL: 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.

DeFrancis, J. (1989). *Visible speech: The diverse oneness of writing systems*. Honolulu: University of Hawaii Press.

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language, 68* (1), 81-138.

Dryer, M. S. (2011). The evidence for word order correlations. *Linguistic Typology, 15*(2), 335-380.

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F. X., Balota, D. A., Brysbaert, M., Carreiras, M., Ferrand, L., Ktori, M., Perea, M., Rastle, K., Sasburg, O., Yap, M. J., Ziegler, J. C., & Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PloS One, 6*(9), e24974.

Durrett, G., & DeNero, J. (2013). Supervised learning of complete morphological paradigms. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 1185-1195.

Erlewine, M. Y., & Kotek, H. (2016). A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory, 34*(2), 481-495.

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *The Quarterly Journal of Experimental Psychology, 68*(8), 1469-1488.

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.

Feng, G., Miller, K., Shu, H., & Zhang, H. (2001). Rowed to recovery: the use of phonological and orthographic information in reading Chinese and English. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(4), 1079-1100.

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.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*(2), 488-496.

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

Fox, R. A. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance, 10*(4), 526-540.

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

Gandour, J., Xu, Y., Wong, D., Dzemidzic, M., Lowe, M., Li, X., & Tong, Y. (2003). Neural correlates of segmental and tonal information in speech perception. *Human Brain Mapping, 20*(4), 185-200.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328-331.

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.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.) *Universals of language* (pp. 73-113). Cambridge, MA: MIT Press.

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.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., Yap, M. J., Bengson, J. J., Niemeyer, D., & Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods, 45*(4), 1099-1114.

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.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology, 68*(8), 1665-1692.

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.

Klein, D., Zatorre, R. J., Milner, B., & Zhao, V. (2001). A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. *Neuroimage, 13*(4), 646-653.

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.

Lee, H.-H. (2016). A comparative study of the phonology of Taiwan Sign Language and Signed Chinese. Unpublished National Chung Cheng University Ph.D. thesis.

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34* (1), 12-31.

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.

Lewis, M. P., Simons, G. F., Fennig, C. D. (Eds.). (2014). *Ethnologue: Languages of the world, Seventeenth edition*. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com/17/.

Light, T. (1977). The Cantonese final: An exercise in indigenous analysis. *Journal of Chinese Linguistics, 5*(1), 75-102.

Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language & Cognitive Processes, 16*(5/6), 565-581.

Malins, J. G., & Joanisse, M. F. (2010). The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. *Journal of Memory and Language, 62*(4), 407-420.

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. (2007). Generative morphology as psycholinguistics. In G. Jarema & G. Libben (Eds.), *The mental lexicon: Core perspectives* (pp. 105-128). Amsterdam: Elsevier.

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), 791-818.

Myers, J., & Chen, T-Y. (2016). The time course of sociolinguistic influences on wordlikeness judgments. In A. Botinis (Ed.), *Proceedings of the 7th Tutorial and Research Workshop on Experimental Linguistics* (pp. 119-122). International Speech Communication Association.

Myers, J., & Tsay, J. (2005). The processing of phonological acceptability judgments. Proceedings of Symposium on 90-92 NSC Projects (pp. 26-45). Taipei, Taiwan, May.

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.

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.

Norcliffe, E., Harris, A. C., & Jaeger, T. F. (2015). Cross-linguistic psycholinguistics and its critical role in theory development: early beginnings and recent advances. *Language, Cognition and Neuroscience, 30*(9), 1009-1032.

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

O'Brien, R. M. (2007). A caution regarding rules of thumb for Variance Inflation Factors. *Quality & Quantity, 41*, 673-690.

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.

Rentzepopoulos, P. A., & Kokkinakis, G. K. (1996). Efficient multilingual phoneme-to-grapheme conversion based on HMM. *Computational Linguistics, 22*(3), 351-376.

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.

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.

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., 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., Arévalo, A., Vargha, A., Butler, A. C., Buffngton, R., & Bates, E. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language, 51*(2), 247-250.

Tse, C.-S., Yap, M. J., Chan, Y.-L., Sze, W.-P., Shaoul, C., & Lin, D. (forthcoming). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*.

Tseng, C.-H., Huang, K.-Y., & Jeng, J.-Y. (1996). The role of the syllable in perceiving spoken Chinese. *Proceedings of the National Science Council, Part C: Humanities and Social Sciences, 6* (1), 71-86.

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology, 165*(6), 710-718.

von Bastian, C. C., Locher, A., & Ruflin, M. (2013). Tatool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods, 45*(1), 108-115.

Weskott, T., & Fanselow, G. (2011). On the informativity of different measures of linguistic acceptability. *Language, 87* (2), 249–273.

Wong, A. W.-K., & Chen, H.-C. (2008). Processing segmental and prosodic information in Cantonese word production. *Journal of Experimental Psychology: Learning, Memory and Cognition, 34* (5), 1172-1190.

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, L., & Shu, H. (2003). Hanyu yanyu chansheng de yuyin jiagong jizhi [Phonological processing mechanism in Chinese speech production]. *Xinli Kexue, 26* (5), 818-822.

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

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.