Widening the Net: Challenges for Gathering Linguistic Data in the Digital Age

Pranav Anand, Sandra Chung, & Matthew Wagers
Department of Linguistics, UC Santa Cruz

Abstract

Reliable scientific data from the full diversity of the world's languages is required to validate current views of the human capacity for language. The current methodologies of linguistic investigation—fieldwork, experimentation, the mining of large corpora—have inherent limitations. We raise the challenge of how these methodologies can be transformed to overcome their limitations. Meeting this challenge will require new versions of these methodologies that are simpler, more portable, and less culturally entrenched than those currently in use. Such methodologies should generalize cleanly to diverse languages, communities, and settings and should generate types of data that can be compared across languages more efficiently than can be done now. We consider solutions that maximize the potential of the native speaker as scientific investigator. Micro-tasks embedded in games and deployed on widely accessible electronic mediums, such as mobile devices, illustrate a promising means for realizing this goal, and a viable system for expanding the diversity of data in other behavioral sciences.

The challenge

Linguistic sciences have relied upon three principal methodologies for data-gathering: investigation of speakers' judgments, through interviews in fieldwork settings or introspection; psycholinguistic research conducted in experimental settings; the mining of large corpora. But most linguistic research to date has been conducted on a small circle of languages associated with socio-economic power: English, other Western European languages, Chinese, and Japanese. The question is whether the view of the human capacity for language that emerges from this relatively narrow linguistic and cultural domain is robust enough to account for the properties that all human languages share.

The vast majority of the world's languages are spoken by small populations that have fewer than a million speakers, lack socio-economic power, typically are not literate, and do not share Western cultural presuppositions. Many of the world's languages are also endangered. It is an important, continuing intellectual concern to document as many endangered languages as possible while they are still vibrant. Although the human capacity for language is fundamentally genetic, it allows for systematic differences; this property renders language unique among human cognitive systems. Language documentation is crucial to our understanding of this unique property.

The challenge raised here is different: to validate our views of the human linguistic capacity, we need reliable scientific data from the full diversity of the world's languages. The current *methodologies* of linguistic investigation—fieldwork, experimentation, the mining of corpora—have inherent limitations. How can these methodologies be transformed to overcome their limitations?

Meeting this challenge will require radically new versions of these methodologies that are simpler, more portable, and less culturally entrenched than those currently in use. Such methodologies should generalize cleanly to diverse languages, communities, and settings; they should be cost-effective; they should be user-friendly enough that they could be put to use by speakers of understudied languages in their own communities. They should generate types of data that can be compared across languages more efficiently than can be done now. The ultimate goal would be for such methodologies to reach standards of effectiveness sufficient to replace existing methods completely, whether investigating better studied or undocumented languages. The question of what findings hold outside our cultural matrix is one of general concern to the behavioral sciences. The initiatives we suggest below, particularly those regarding infrastructure

and analysis, have broad applicability to the issue of how behavioral sciences could treat diversity more systematically.

Data-gathering traditions in linguistics

To see what might be involved in such a transformation inside linguistics, consider that field's major data-gathering traditions.

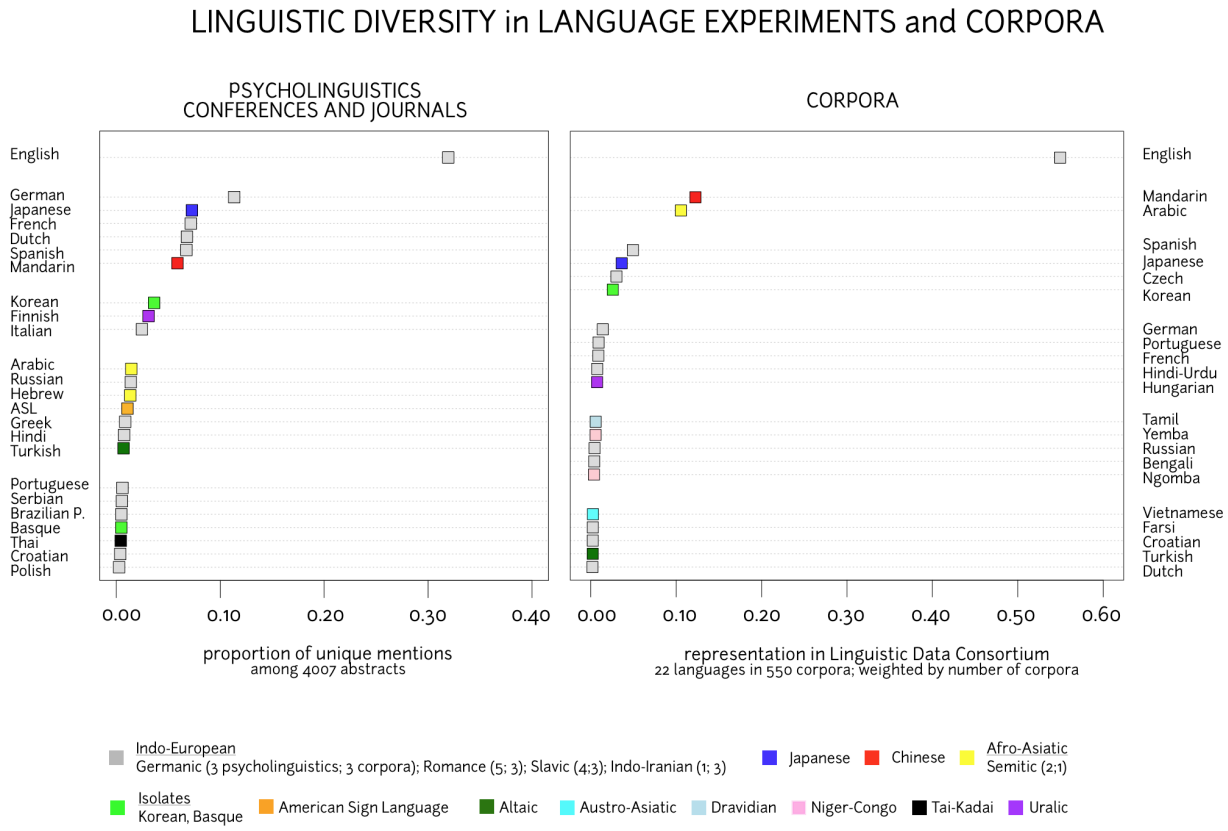
Traditional fieldwork involves working one-on-one with speakers of a language to record their words, sentences, narratives, and linguistic intuitions. Linguists who introspect about their own language are conducting a self-directed version of fieldwork. Successful fieldwork can produce sophisticated data, in a way that is unfeasible or costly with other paradigms. Nonetheless, traditional fieldwork has well-known limitations. It is labor-intensive and individual-centered. Data gathered from one or two individuals depend heavily on the individuals' partnership with the linguist and might not generalize to the larger community. In short, fieldwork is not optimized to scale up.

Psycholinguistic and neurolinguistic experimentation involves measuring the behavioral and physiological responses of participants as they complete language-related tasks, such as judging whether or not a string of characters is a legal word. Measurements include the speed and accuracy with which the task or its subparts are accomplished, data about eye-movements, or physiological indices of brain activity (EEG, MEG, fMRI, NIRS). These measures can be combined to build models of language processing that capture dynamic aspects of language that are absent from grammatical description. The difficulty of discriminating two speech sounds, the moment-by-moment complexity of recognizing a grammatical construction, the time it takes to construct a particular interpretation—these help to deepen our understanding of the human language capacity.

Such research can require considerable data to draw reliable conclusions: studies of sentence understanding typically involve 20-60 participants, each reading 100-200 sentences. Sophisticated, costly instrumentation is often required. A more serious challenge, not often recognized, is that the experimental method is heavily culturally circumscribed. It relies upon specific societal norms: the importance of test-taking, willingness to maintain exclusive focus on unnatural tasks, and an abstract social contract with the experimenter. Additionally, most experimental tasks are solitary, and require responses to linguistic material presented out of context, often by a machine. These 'cultural felicity' conditions are typically satisfied in university settings, so—unsurprisingly—the database of observations is heavily skewed to the cognitive make-up of 18-21 year old university students in highly industrialized societies. Our survey of over 4,000 psycholinguistics abstracts from leading conferences and journals found that ten languages accounted for at least 85% of the research (Figure, left panel). Overall, only 57 languages were represented, an exceedingly narrow slice of the world's linguistic diversity.

Corpus building involves the collection of naturally-occurring text and spoken language. The strength of corpus analysis is its ability to uncover linguistic tendencies, especially those dependent on the larger context in which language is embedded. It also enables comparisons across language, genre, modality, and time. For example, the study of conversational corpora has led to insights into how speakers use rhythmic structure to plan an utterance on the fly. However, the strength of such claims is often correlated with the size of the corpus. There are few accessible data sources drawn from informal, every-day language, and their manufacture has proven labor- and cost-intensive. For reasons of literacy, documentation, and digital access, studies of large corpora have drawn heavily on a very few languages; among 550 corpora available from the Linguistics Data Consortium, five languages accounted for 85% of data

sources: English, Chinese, Arabic, Spanish, and Japanese (Figure, right panel). While the growth of digitized communication should lead to more diverse corpus data in the future, challenges remain. The media produced on-line are skewed towards highly-studied languages; the typical goal—a 1-million word corpus—is unrealistic for languages with small populations of speakers.



Realizing data diversity in diverse languages

Each data-gathering tradition faces the same research challenges. To move to the next stage of research and teaching, scientists must radically scale up the range and diversity of languages investigated. How can this transformation be accomplished in a way that maintains scientific rigor and engages the diverse human populations involved? Language endangerment imposes a time limit on solutions to this problem. Moreover, current linguistic methodologies do not *already* provide a solution. Data-gathering on understudied languages is unlikely to increase radically through conventional fieldwork, given fieldwork's labor-intensive character. The cost, sample size, and cultural circumscription of current experimental research present formidable barriers to the extension of this methodology to understudied languages. Collection of corpora in understudied languages can be difficult when speakers are not literate. These challenges provide the opportunity for linguists to rethink their methodologies and recast them to resolve these issues.

We envision research initiatives that will address these issues in each data-gathering tradition. These initiatives will be motivated by intertwining goals; namely, to:

(a) import the individual-centered, culturally sensitive ethos of fieldwork into the larger-

scale methodologies and bring the consistency of the larger-scale methodologies into fieldwork;

(b) maximize the native speaker's potential as scientific investigator;
(c) develop flexible protocols for language-related tasks that are engaging across a range of cultures;
(d) advance data gathering and analysis practices to make inferences from small data samples robust;
(e) construct applications, tools, and platforms that are usable, intuitive, and fault-tolerant, drawing upon advances in cyber-infrastructure.

Research initiatives in fieldwork should bring together theoretical linguists, typologists, and field linguists to develop in-depth, standard protocols for collecting sophisticated linguistic data in a consistent format. Building on results of the E-MELD project (particularly, GOLD), these protocols should improve on previous typological questionnaires. They should exploit a decision-tree model to maximize opportunities for exploring systematic patterns of cross-linguistic variation. They should involve collecting spontaneous narrative (Bird, 2010), in addition to words and sentences later translated into a reference language (e.g. English, Mandarin, Swahili). The implementation should be piloted in communities that have already partnered with linguists on documentation projects.

Research initiatives in experimental methods should bring together experimental linguists, field linguists, anthropologists, and psychologists to create and evaluate language-related tasks that subjects from many cultures will want to complete. Games, broadly speaking (e.g. card games, strategy games, video games), offer a promising model. Games create expectations of repetitive micro-tasks, contain built-in incentivization, decrease subjects' skepticism about the naturalness of the task and, as a by-product of competition, emphasize rapidity and accuracy of response.

Games are also remarkably flexible. Word games are prevalent worldwide, engage metalinguistic awareness, and have already been used to investigate prosody and word formation. Research should identify ways such games can be extended to other domains of linguistic inquiry. For example, the popular word-building game Ghost has a sentence-building variant ("Cheddar Gorge") that could be used to investigate constraints on question formation and other sentence types. A game like Pictionary could be adapted to investigate how and when sentences are ambiguous. Finally, language-related tasks could be embedded in a variety of games as simple puzzles or brain teasers.

Research initiatives in data analysis should determine how best to approach tasks in which participants (or players) respond only to a few key linguistic tokens or generate one or two pieces of data. Preliminary research on linguistic phenomena suggests that intelligent sampling of small populations can lead to results comparable to those achieved by data-intensive sampling of much larger populations (Munro et al., 2010). The range of measures to which this generalizes is an important research question, and should be addressed by developing systems for pooling data from many experiments and researchers. Such systems would improve the quality of research already being conducted. Standard data analysis often proceeds without making any assumptions about reasonable bounds on the data. If such bounds were known in advance, smaller data sets could be informative. Pooled data should be used to develop benchmarks for common psycholinguistic tasks, which would enable better inference from new sparse data sets.

Complementary research initiatives should partner linguists with computer engineers to implement the computational infrastructure for these new data gathering techniques. The systems developed must be cheap, portable, and easily modifiable. They must allow straightforward

recording of linguistic data and behavioral measures, and provide graceful methods of reporting these measures to the research community. Two strategies are worthy of further attention: Crowdsourcing websites like Mechanical Turk, which provide a marketplace for cheap micro-task completion, provide one potential data-gathering method; research should investigate the viability of such mechanisms for gathering data from diverse populations. The worldwide proliferation of mobile devices (e.g. cell phones) offers a complementary strategy, one which makes recording large amounts of spoken-language conversation a realizable goal. Attending to speakers' desires for recording conversation (e.g. capturing a moment, preserving cultural heritage) will be invaluable for the development of compelling, fool-proof applications across cultures and computational proficiencies. It will also help foreground the intellectual property issues involved in disseminating language materials, an issue of real concern for vulnerable populations.

The initiatives sketched above offer a glimpse of where linguistic practice could be in the next 20-50 years: a science as firmly engaged in large-scale data gathering and analysis as applied computational work on language, but with a focus on understanding both the common aspects and full diversity of human linguistic expression.

Thanks to Scott Seyfarth for assistance.

References

Bird, Steven, 2010. A scalable method for preserving oral literature from small languages. *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries*, 5-14, Gold Coast, Australia.

Electronic Metastructure for Endangered Languages Data (E-MELD). http://emeld.org/index.cfm

Munro, Robert, et al. 2010. Crowdsourcing and language studies: The new generation of linguistic data. *NAACL 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*, June 6, Los Angeles.