A Grand Challenge for Linguistics: Scaling Up and Integrating Models

Emily M. Bender
Department of Linguistics
University of Washington
ebender@uw.edu

Jeff Good
Department of Linguistics
University at Buffalo
jcgood@buffalo.edu

Abstract

The preeminent grand challenge facing the field of linguistics is the integration of theories and analyses from different levels of linguistic structure and aspects of language use to develop comprehensive models of language. Addressing this challenge will require massive scaling-up in the size of data sets used to develop and test hypotheses in our field as well as new computational methods, i.e., the deployment of cyberinfrastructure on a grand scale, including new standards, tools and computational models, as well as requisite culture change. Dealing with this challenge will allow us to break the barrier of only looking at pieces of languages to actually being able to build comprehensive models of all languages. This will enable us to answer questions that current paradigms cannot adequately address, not only transforming Linguistics but also impacting all fields that have a stake in linguistic analysis.

I. Overview

The preeminent grand challenge facing the field of linguistics is the integration of theories and analyses from different levels of linguistic structure and aspects of language use to develop comprehensive models of language, encompassing cognitive representations and processing, the deployment of language in social interaction and dynamics of historical linguistic change. We argue in this white paper that addressing this challenge will require massive scaling-up in the size of data sets used to develop and test hypotheses in our field as well as computational methods for handling both these large data sets and the integration of different analyses. In short, the challenge can only be met through deployment of cyberinfrastructure on a grand scale.

II. Challenges

II.a. The Data Acquisition Challenge

Too many linguistic analyses are based on small data sets. The examples in the data sets may be well chosen, and may be believed to represent large classes of utterance types, but this is often not substantiated. The problem is particularly acute in subfields concerned with linguistic structure, but even interdisciplinary fields such as sociolinguistics and psycholinguistics could benefit from working with larger data sets from a wider variety of languages.

The data acquisition challenge thus concerns the problems of collecting, curating and sharing data sets on a grand scale. These data sets include not only raw data, but also various kinds of metadata, annotation and analyses: As Bender and Langendoen (2010) note, one linguist's

analysis/annotation is the next linguist's data. The data should include naturally occurring speech (audio, video, transcription) as well as findings from more controlled interactions (e.g., psycholinguistic experiments). Most importantly the data should include as many of the world's languages (spoken and signed) as possible, and soon: Time is getting critically short for many languages as evidenced by NSF's efforts in the Documenting Endangered Languages program (see also Abney and Bird 2010).

Another facet of the data acquisition challenge is data sharing: both convincing linguists to share data and creating infrastructure that makes it feasible to do so. The infrastructure issues are taken up below. Here we note that funding agencies have an important role to play in encouraging data sharing by requiring data collected through funded research to be made available, ideally publicly, and in supporting the researchers' efforts to improve data sharing infrastructure.

II.b. The Data Mining Challenge

Once we have more data to work with we are faced with the data mining challenge: How do we find the data that is relevant for a particular analysis or hypothesis from within the much larger set? This is particularly clear when the data is naturally occurring text or speech and the analysis or hypothesis concerns a relatively rare construction. In some cases, the solution involves searching over previously provided annotations. In others, finding relevant examples requires automated (perhaps noisy) analysis of the data.

II.c. The Complexity Challenge

The field of linguistics studies the structure, use, acquisition and processing of human languages. The field has made much progress in understanding these aspects of language by seeking out generalizations and using them build models that are significantly simpler than the complex objects they represent. However, this is largely achieved by modeling only a single dimension (e.g., level of linguistic structure) or at most two (e.g., studies of interfaces, or studies of variation or processing within a level) at a time. Even within these restricted domains, models typically only focus on a small set of phenomena. These are clearly poor approximations of the way people acquire, produce and understand language.

The time is right to test whether or not our models can scale up and, if they cannot, to develop new ones that can. Building large-scale models requires the assistance of computational methods, allowing us to test models of the required complexity against representative datasets. In structural subfields of linguistics, such as phonology and syntax, this takes the form of grammar engineering, where the linguist encodes analyses in a machine-readable notation. Those analyses are then used to assign structures to linguistic data, with manually annotated data used to verify their appropriateness.

III. Opportunities

Though these challenges are big, recent developments both within linguistics and in other fields have put us in a good position to address them. General computational infrastructure, including ever increasing computational power, networks (particularly the Internet), organizational models (e.g., crowdsourcing) and electronic recording devices provide the foundations for systems that allow linguists to collect, annotate and share data.

Particularly noteworthy are developments in the last decade which have resulted in significant convergence among descriptive and computational linguists in their approaches to encoding linguistic data. On the one hand, as primary data collection and annotation, even for understudied languages, has increasingly taken on digital forms, best practices have emerged to ensure data longevity and interoperability.[1] On the other hand, computational linguistics has come to make use of a diverse portfolio of data processing methods, relevant to corpus processing, psycholinguistic modeling and formal grammar construction, which have ready applications to more traditional domains of language analysis.

IV. Key Aspects of the Solution

The Cyberling 2009 Workshop[2] and other similar discussions have mapped out the following as key strategies in approaching the grand challenge.

IV.a. Standards

Meeting the sub-challenges enumerated above will require building a cyberinfrastructure for linguistics, but that effort in turn will necessarily be decentralized and come about through contributions of many independent research groups. In order for those contributions to add up to a cohesive cyberinfrastructure, we need interoperability. Interoperability, in turn, requires standards: for data encoding, for annotations, for storage, retrieval and search of data and for tracking provenance of both data and annotations.

IV.b. Tools

Standards alone won't do the job. In order to engage Ordinary Working Linguists (OWLs) in collecting and contributing data and annotations, we need standards-compliant tools. These tools will make it easier for OWLs to carry out research they would otherwise want to carry out, while producing data in standards-compliant formats as a side effect and making it trivial for them to share their data when they are ready to do so. On the other end, the field will need tools for data discovery, search and aggregation.

IV.c. Coordination and Collaboration

As noted above, a successful cyberinfrastructure for linguistics will be the result of combined efforts from many separate groups. Integrating the work of these separate groups will require coordination and communication among all stakeholders and will require the creation of new opportunities for collaboration. Some of these opportunities will be primarily technology-driven (e.g., through the adoption of interoperable formats and tools). Others will be more social: Computational linguists, for example, will need to actively work with descriptive linguists. Funding agencies can encourage this by prioritizing such collaborative work in their award decisions.

[1] See, for example, http://emeld.org/.
[2] http://cyberling.elanguage.net/

IV.d. Data Sharing

The best cyberinfrastructure in the world is still useless until it is populated with data. Here the field of linguistics is in need of culture change. At present, most linguists fail to share their data for reasons ranging from the difficulties involves in curating it into a distributable form, to concerns regarding speaker privacy, to a desire to be finished working with it on their own before giving others access. This is clearly an unsustainable state of affairs. Data publication is key to replicability and verification, and the field must, therefore, learn to value data publication independent from the research done using that data.

Such culture change must be accompanied by dedicated cyberinfrastructure for data sharing in the form of archives like the Endangered Languages Archive which blends long-term preservation of data with near-term access, while respecting the needs and sensitivities of the diverse stakeholders of language data. After all, the audience for language resources is vast, especially when their content is revealing of a given language group's customs and traditions. The data sharing must be done sustainably and sensitively, but at the same time, the broader impacts of increased sharing of language data are potentially enormous.

IV.e. Computational Methods

Finally, we need computational methods for assisting linguists, whether alone or in teams, in performing analysis of large-scale data sets and building and testing models. The methods must be designed to support a diverse set of conditions, from refinement of our understanding of well-studied languages like English to arriving at a more basic picture of languages, often on the verge of extinction, which have been barely studied. Bird (2010) offers an interesting proposal for social and technological models for the latter scenario, but this is just a start. In addition, new methods also require rigorous schemes to evaluate their relative effectiveness.

When scaling analyses of particular phenomena to large data sets, we need tools for extracting relevant data out of them. Here there are two approaches, both valuable. One is large-scale annotation of corpora, where the volume of data processed can be scaled up because the annotations are applied by hand-built grammars (as in the Redwoods Treebank).[3] The second is to use machine learning (e.g., the active learning methods deployed by the EARL project at UT Austin[4]) to generalize from a small set of hand-annotated data to larger sets of otherwise unannotated data. The results of both kinds of approach are promising, but further research is needed to expand their effectiveness.

The grand challenge is ultimately about building comprehensive models incorporating what we learn from analyses of constrained domains. This is a relatively unexplored problem, and we need software support in the form of development environments and formalisms in which multiple alternative models can be tested against the requisite datasets. For structural aspects of language, such development environments exist. There are also multilingual grammar engineering projects (including ParGram[5] and the LinGO Grammar Matrix[6]) which facilitate the development of such models of new languages by leveraging what has been learned in the

[3] http://wiki.delph-in.net/moin/RedwoodsTop
[4] http://comp.ling.utexas.edu/earl/
[5] http://pargram.b.uib.no/
[6] http://www.delph-in.net/matrix/

development of integrated resources for well-studied languages. What we do not yet have is a means to create truly comprehensive models of human language, including cognitive processing, language use in social interaction and language change over time which, moreover, are consistent with the attested range of linguistic diversity.

V. Contributing Fields and Synergies

Though this challenge is framed as a challenge for Linguistics, meeting it will certainly involve other fields. First and foremost there is Computational Linguistics (an interdisciplinary area involving Linguistics, Computer Science and Electrical Engineering), which stands to benefit from the increased availability of linguistic data as well as increased scalability of computational methods to lesser-studied languages. Similarly, the comprehensive models we call for require input from Psychology, Anthropology and Sociology, as well as aspects of Computer Science not specifically geared towards language analysis such as Social Computing. More broadly, any field of study which requires data in the form of natural language speech or text has a synergy with Linguistics in these efforts: While we may require different annotations over the data, the initial encoding, curating and archiving of the texts can be a shared effort.

VI. Conclusion

VI.a. Foundational aspect

This challenge is at the heart of Linguists and also central to Cognitive Science and Anthropology. It speaks to the foundations of all language research by seeking to reorient modes of exploration away from the dominant trend of "siloed" research on narrow problems towards an "additive" approach that constructs the pieces into a comprehensive model. Language, after all, functions as a coherent whole; research paradigms should reflect this.

VI.b. Transformative potential

Dealing with this challenge will allow us to break the barrier of only looking at pieces of languages to actually being able to build comprehensive models of all languages. This will enable us to answer questions that current paradigms cannot adequately address, not only transforming Linguistics but also impacting all fields that have a stake in linguistic analysis.

References

Abney, Steven and Steven Bird. 2010. "The Human Language Project: Building a Universal Corpus of the World's Languages." Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Pp.88-97. Uppsala, Sweden. Association for Computational Linguistics.

Bender, Emily M. and D. Terence Langendoen. 2010. "Computational Linguistics in Support of Linguistic Theory." Linguistic Issues in Language Technology 3(2). Pp.1-31.

Bird, Steven. 2010. "A Scalable Method for Preserving Oral Literature from Small Languages." Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries. Pp.5-14.