Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, June 29, 2009

Computational Linguistics in Support of Linguistic Theory

Emily M. Bender
and D. Terence Langendoen

Submitted, June 29, 2009,
Revised, December 29, 2009,
Published by CSLI Publications

LiLT volume 3, issue 2 February 2010

Computational Linguistics in Support of Linguistic Theory

Emily M. Bender, *University of Washington* and D. Terence Langendoen, *University of Arizona*

Abstract

In this paper, we overview the ways in which computational methods can serve the goals of analysis and theory development in linguistics, and encourage the reader to become involved in the emerging cyberinfrastructure for linguistics. We survey examples from diverse subfields of how computational methods are already being used, describe the current state of the art in cyberinfrastructure for linguistics, sketch a pie-in-the-sky view of where the field could go, and outline steps that linguists can take now to bring about better access to and use of linguistic data through cyberinfrastructure.

LiLT Volume 3, Issue 2, February 2010.
Computational Linguistics in Support of Linguistic Theory.

1 Introduction

This paper seeks to answer two related questions. The first is what can computers and the infrastructure in which they are networked together do for linguistics, and the second is what do linguists need to know and do in order to take advantage of computational tools and resources in their work. We would like to emphasize at the outset that we are not advocating making all linguists into computational linguists: computational methodology for linguistic analysis is not the same thing as computational linguistics (though in many instances, it relies on the results of previous work in computational linguistics), nor do we expect all linguists to become computer programmers. Rather, we are arguing that computers can be very effective tools in doing linguistic research, and the field as a whole will benefit to the extent that we can build on general advances in cyberinfrastructure to create a cyberinfrastructure for linguistics.

As is probably already clear from the tone, this paper aims to be explicitly persuasive. In particular, we aim to persuade the reader to try out current computational methodologies, to teach students to use computational methodologies, and to collaborate in building the next generation of infrastructure for computational methods in linguistics. The paper is structured as follows: §2 describes how computational methodologies can advance inquiry in linguistics, in general terms and then through a series of examples of research questions which can only be approached with computer assistance, but which can be approached with existing or near-term technology. §3 surveys the currently existing infrastructure, and sets the stage for §4, a pie-in-the-sky view of a linguistics research environment of the future. We aim there to dream big and then ask what needs to be done to get there. The first steps are described in §5.

2 What Computers Can Do for Us

In this section, we explore how computers can be used as tools in the service of linguistic research, i.e., data collection, analysis and theory development. As with many other sciences, computers and the infrastructure of the Internet are useful in linguistics because they allow us to access and manipulate more data than could be done otherwise, while also collaborating with more people across greater distances. By allowing us to bring more data into consideration, and to manage the resulting complexity, and by allowing us to more effectively combine the efforts of multiple researchers, computational methods allow us to ask questions that would otherwise be completely intractable. We be-

lieve this to be true across all subfields of linguistics, though the state of the existing infrastructure (and, relatedly, existing practice) differs across subfields.

As part of our overall persuasive goal, we review here a range of research questions which are currently being pursued or could be pursued with existing or near-term technology, across a wide variety of subfields. In doing so, we hope to illustrate for our readers the relevance of computational methods. Note that this list of questions (and of subfields!) is not meant to be exhaustive. Our aim here is merely to give a sense of what is possible. The reader is encouraged to dream up further similar questions!

2.1 Descriptive and Documentary Linguistics

The first area we look at is descriptive and documentary linguistics. There has been a lot of effort in recent years to bring computational methods to bear in this area, as time is running out. There aren't enough linguist-hours left to document all of the endangered languages before they lose their last speakers, and so the need for computational tools which enhance the efficiency of linguists working in the field is acute. The E-MELD project[1] was one response to this need, developing resources and collecting and documenting current best practice. Questions that computational methods will soon be able to help answer include the following: Given a transcribed and translated narrative, which is not yet in IGT format,[2] which words are likely to be belong to the same lemma? Or given a collection of texts and a partial morphological analysis, which words are still unaccounted for? The EARL project at UT Austin is an example of the kind of research that is laying the groundwork for such tools. Moon and Erk (2008) present a methodology for clustering words from the source language side of a collection of translated text. The clusters represent words likely to be different inflected forms of the same lemma. Palmer and Erk (2007) present IGT-XML, an XML format for encoding IGT, which is a necessary precursor for semi-automated production of IGT. Palmer (2009) investigates how a machine learning paradigm called "active learning" (Cohn et al., 1994) can be used to speed up the annotation of transcribed texts. In this paradigm, the computer extracts generalizations out of a small number of annotations provided by the human, and then attempts to apply these generalizations to new data. In doing so, it

[1] http://e-meld.org

[2] IGT stands for interlinear glossed text, the familiar three-line format giving source language form, a morpheme-by-morpheme analysis/gloss, and a free translation into some other language.

keeps track of its own uncertainty, and then presents the human with the most difficult (i.e., interesting) cases to annotate next. Of course, in any tool built on this kind of methodology, the annotations would need to include metadata about their provenance (human or machine) and validation (whether they have been human-approved).

Another set of questions that computers can assist with in descriptive and documentary linguistics concern phonetics and the logically prior problem of transcription. As we will explore further below, linguistic analysis always involves layered systems, where analysis at one level becomes annotations (and then simply "data") at the next. In field linguistics, the very first problem is transcribing the data. Taking for now the relatively simple case where the units of interest are phonological segments, the first problem to be solved is the relationship between phones. The sort of distributional analysis that is traditionally used to group phones into phonemes is highly amenable to automation. Thus one could ask, given data in an IPA transcription, which phones are likely allophones, and what are some likely phonological rules? The goal here is not to take the linguist out of the loop, but to present likely possibilities for the linguist to explore. The ELTK project[3] (Farrar and Moran, 2008) is laying the groundwork for such a system, with software that can automatically generate phoneme inventories based on data sets in Praat[4] (Boersma and Weenink, 2009) or Elan[5] transcription files and extract allomorphs.

The final example in this section concerns the problem of searching for cognates across texts or vocabularies in potentially related languages. This would seem straightforward enough, but in the typical case, each language would be associated with its own transcription system or practical orthography. While the transcription systems might all be based on IPA or some other cross-linguistic system, they are typically each developed in the context of different majority-language orthographic traditions and/or different linguistic traditions, and so each have some idiosyncrasies. Moran (2009) presents a software ontology that supports the explicit encoding of transcription systems, which in turn supports cross-linguistic queries over phonological inventories and word forms.

2.2 Phonetics and Phonology

Turning next to phonetics and phonology, there are a variety of interesting questions that can be asked once phonetic and phonological

[3] http://purl.org/linguistics/eltk
[4] http://www.fon.hum.uva.nl/praat/
[5] http://www.lat-mpi.eu/tools/elan/

data and analyses are encoded in interoperable, machine-readable form. Building such a resource is the goal of the PHOIBLE project (Moran and Wright, 2009), which is collecting and digitizing phoneme inventories (including information about allophones) from legacy sources, web-accessible papers, and user input, while also collecting and integrating distinctive feature sets. With this resource, it becomes possible to ask questions such as: How do different feature systems quantify the variation across languages differently? Which feature systems locate differences in historically plausible ways, such that differences among historically or areally related languages are less pronounced?

Similarly, the constraint sets of Optimality Theory (OT) raise questions that are best answered with computational support, and there has been a significant amount of work on computational implementations of OT. Two give just two examples, the Erculator software[6] (Riggle et al., 2007) allows linguists to take a set of OT constraints, and discover the range of language (types) that set of constraints predicts. Looking at the problem of acquisition from an OT perspective, Boersma and Hayes (2001) ask what kind of data is required for learning rankings of a given set of OT constraints.

2.3 Morphosyntax

Just as with phonetics and phonology, computational methods and data aggregation allow linguists researching morphosyntax to look at broad patterns across languages, discover languages instantiating properties of interest, and verify the correctness of formal analyses. The former two functions are supported by large databases, including corpora, databases of linguistic examples, and databases of language properties. The latter function is supported by specialized software for interpreting and applying linguistic formalisms.

Databases in this area include ODIN[7] (Lewis, 2006) and WALS[8] (Haspelmath et al., 2008). ODIN, the On-line Database of INterlinear glossed text, is a collection of linguistic examples harvested from linguistics papers available as pdf documents on the web. These examples are a very rich source of information. The interlinear (IGT) format gives source language text, a morpheme-by-morpheme gloss, and a free translation into some other language. While the level of detail given in the glossing depends on the purpose for which the author is citing the example, they always do include some additional information. By systematizing this information (through building an extension for mor-

[6] http://clml.uchicago.edu/?page_id=11
[7] http://odin.linguistlist.org/
[8] http://wals.info

phosyntactic concepts to the GOLD ontology (Farrar and Langendoen, 2003)), ODIN enables linguists to search across the data to discover, e.g., which languages have ergative-absolutive case-marking and object agreement on the verb, or which languages have anti-passive voice and reflexives expressed through affixes.[9] The site reports that as of October 16, 2009, it had curated 127,306 data instances in 1266 languages from 2013 documents.

WALS, the World Atlas of Language Structures, is a massive typological database, including 142 chapters each studying some linguistic feature and categorizing 120–1370 languages according to that feature. In total, there are 2,650 languages represented in the database, and over 58,000 data points (feature values for languages).[10] In addition, the languages are all associated with geographical location, enabling the mapping of linguistic properties in the world. As with ODIN, but based on a different original set of data sources, WALS allows linguists to search for languages with interesting combinations of typological properties.[11]

Another major strand in computational approaches to syntax is grammar engineering, or the process of encoding syntactic analyses in machine-readable form, so that they can be validated through parsing input strings or generating from input semantic representations. The methodology of grammar engineering allows syntacticians to ask questions such as: How does my new analysis of phenomenon X interact with the rest of the grammar as implemented so far (i.e., with my analyses of phenomena A-W)? How many distinct structures does my grammar assign to this sentence? How many realizations does my grammar assign to this input semantics? Software support for grammar engineering exists in a variety of frameworks (e.g., HPSG (Copestake, 2002), LFG (Crouch et al., 2001) and CCG (Baldridge et al., 2007)), TAG (Crabbé, 2005), P&P (Fong, 1999), and Minimalism (Stabler, 1997)) and has become increasingly practical as computers have gotten faster and parsing algorithms more sophisticated.[12]

2.4 Semantics and Pragmatics

Computational resources can also be brought to bear on the entire range of problems in semantics and pragmatics, such as exploring the

[9] More precisely, ODIN enables linguists to discover languages which other linguists have analyzed in this fashion.

[10] http://wals.info; accessed on April 16, 2009.

[11] WALS includes chapters on phonetics, phonology, and lexical properties, as well as morphosyntax.

[12] For more on computational syntax, see Bender et al. to appear and Bender, this volume.

nature of lexical structure; resolving ambiguity in context; determining the scope of connectives (e.g. conjunction, disjunction and negation), quantifiers, modals and other operators; analyzing the structure of events; tracking antecedent-anaphor relations in discourse; identifying and classifying metaphors and other figurative use of language; and comparing the nature and use of honorific forms across languages. For doing lexical research, many machine-readable dictionaries are available on line that provide useful information for many languages,[13] but with few exceptions these have not been designed to work with computational tools to provide a representation of meaning or semantic structure than can be integrated with other resources, such as a syntactic parser. One resource that can be used with other resources for work on English is WordNet (Fellbaum, 1998); WordNets are also being developed for several other languages, including Czech, Hindi and Japanese.[14]

In order to carry out systematic cross-linguistic computationally-supported semantic investigations, we require a metalanguage that is both capable of uniformly representing the conceptual and structural richness of the semantic properties and relations of the expressions of the world's languages, and that is computationally tractable. We also require a method for carrying out inferences in that metalanguage. The common practice of using English (or another natural language), possibly together with bits of logical notation, is not adequate for this purpose. While developing such a metalanguage and inference engine together with tools for using them is itself a monumental research undertaking, the payoff will be enormous not only for semantic research, but also for addressing perhaps the greatest "grand challenge" problem in computational linguistics: the development of a computational system that can learn to understand and to produce natural language as fluently and appropriately as humans do.[15] Some initial steps are being taken, for example for representing predicate-argument structure (PropBank; Palmer et al. 2005), semantic frames (FrameNet; Fillmore et al. 2002), temporal relations (TimeML; Pustejovsky et al. 2003),

[13]Hundreds are listed at http://www.yourdictionary.com/languages.html.

[14]The Global WordNet Association maintains a list of WordNets here: http://www.globalwordnet.org/gwa/wordnet_table.htm

[15]Halevy et al. (2009) describe the general need for an appropriate metalanguage and inference engine, together with a model of the domain for carrying out computational linguistic research. They point out that there are many ways that these components can be set up, with the choice being dependent on the task at hand, and the computational resources that are available. For example, for doing unsupervised machine translation, the model, metalanguage, and inference engine would be very different than for constructing semantic representations.

discourse structure (Miltsakaki et al., 2004), intended coreference (Denis and Baldridge, 2008, Rahman and Ng, 2009), and intended senses (Mihalcea and Moldovan, 1999, Sinha and Mihalcea, 2009).

2.5 Psycholinguistics and Language Acquisition

The field of language acquisition has a long history of taking advantage of (networks of) computers to share data and promote the incremental accumulation of knowledge. It began when Brian MacWhinney and Catherine Snow organized a group of 20 child language researchers to pool and digitize their data in 1984 and established the CHILDES database[16] (MacWhinney, 2000).[17] In addition to annotation standards and software tools, CHILDES now incorporates and provides open access to over 44 million words, 2 terabytes of media data, from 32 languages, and has inspired more than 3,000 publications. The related, broader data repository TalkBank[18] has over 63 million words and .5 terabytes of media data from 18 languages (MacWhinney, 2008). Child language data is expensive and difficult to collect. The CHILDES organizers and contributors realized that pooling data would allow them to approach questions that no one researcher or research group could efficiently collect enough data investigate alone. For example, Tardif et al. (1997) investigate whether the relative prominence of nouns and verbs in child-directed speech varies across different languages (English, Italian and Mandarin) and whether these differences correlate with the presence or absence of noun-bias in early language production among children learning these languages. A second example is the work of Alishahi and Stevenson (2007), who build a Bayesian model capable of learning general properties of semantic roles (i.e., theta-roles) and their association with syntactic positions (in English) and particular verb types. This model takes as input a corpus of strings paired with semantic representations. In order for the model to mimic human language acquisition, Alishahi and Stevenson construct a corpus on the basis of information about the frequency of particular verbs and particular nouns as arguments of those verbs in child-directed speech from CHILDES.

In psycholinguistics, there are a variety of interesting questions around the relationship between frequency of morphemes, words, and other linguistic structures and the way they are processed in human language production and comprehension. Answering such questions nec-

[16] http://childes.psy.cmu.edu/

[17] Virginia Yip's interview with Brian MacWhinney, September 2008. http://childes.psy.cmu.edu/intro/interview.mov

[18] http://talkbank.org/

essarily involves computational methods, in order to get the frequency measurements. For example, Jaeger et al. (2009) ask how speaker's choices in utterance generation are influenced by various factors such as information density (Cook et al., 2009, Gómez Gallo et al., 2008, Frank and Jaeger, 2008). In order to calculate information density, they need to process a large corpus of naturally occurring text.[19] Similarly, Arnon and Snider (2009) combine corpus-based methodology (leveraging 20 million words for transcribed American English telephone conversations from the Switchboard (Godfrey et al., 1992) and Fisher (Cieri et al., 2004) corpora) with psycholinguistic methods to determine whether speakers are sensitive to the frequency of four-word phrases as distinct from the frequencies of their subparts.

2.6 Language Variation and Change

Researchers looking at language variation and change also have a long history of adopting computational methods to manage the datasets being explored. In sociolinguistics, this includes the Varbrul software package for computing the contributions of both internal and external constraints (Sankoff, 1975), as well as extensive use of Microsoft Excel and similar software packages for tabulating the properties of occurrences of sociolinguistic variables. Sociolinguistic studies of variable production are canonically corpus-based, and in the typical case involve the researcher working directly with data s/he has collected. Those studies that investigate lexical frequency as a conditioning factor, however, also make use of large, external corpora to get frequency counts (e.g., Bybee 2003). The field of "sociophonetics" involves the pairing of instrumental (i.e., computer-aided) phonetic analysis with sociolinguistic investigation (e.g., Wassink 2006 and the papers in Jannedy and Hay 2006).

Linguistic research on historical varieties (aside from that done solely through reconstruction) is necessarily corpus-based, and has been since well before the advent of digital corpora. As in other subfields, digitization makes it possible to deal with much larger quantities of data as well as larger quantities of more elaborate annotations on data. In addition, computational methods can assist in the creation of consistent annotations over interestingly large collections of text. A pioneering example of this methodology is the Penn-Helsinki Parsed Corpus of Middle English (Kroch and Taylor, 2000). Once such a resource is constructed, it can be used to answer a variety of interesting questions, such as how *do*-support spread across grammatical contexts in the history of English

[19] For discussion of using text corpora in psycholinguistic studies, see Roland and Jurafsky 1998, Gahl et al. 2004 and Frank et al. 2008.

(Han and Kroch, 2000).

The comparative method is also amenable to computational treatment. Nakhleh et al. (2005) present a model of language change that allows for both genetic (common source) and borrowing (contact) relationships between varieties, and a methodology for computing the degree to which a data set provides evidence for each type of connection. They then use this model to estimate, on the basis of 294 lexical, phonological and morphological characters for 24 Indo-European languages, the extent to which early Indo-European languages developed in isolation from each other. They find that the primary evidence for contact relationships involves the Germanic branch, the other branches developing largely independently of each other. It is worth emphasizing here that the computer is not supplanting the linguist in doing this work, but rather systematizing the data in such a way that the linguist can more effectively work with it: the initial analysis results in a selection of possible 'phylogenetic networks' which the authors then analyze for plausibility.

2.7 Typology

Finally, we turn typological research, or the study of variation across, rather than within languages. In addition to WALS (mentioned above), we would also like to highlight the innovative computational methodology of the Autotyp project (Bickel and Nichols, 2002). This project combines the methodologies of 'autotypologizing' and 'exemplar-based sampling' to address the dual problems incorporating languages on their own terms in crosslinguistic work and discovering which, if any, properties of (macro) constructions pattern together crosslinguistically (Bickel, 2007). This methodology fundamentally relies on dynamic computer databases, in which the researchers enter the relevant linguistic properties of each language studied, updating the set of options available when a new language is found that does not fit the existing set of values for a given feature. The databases form a linked set, so that the information gathered in one project can be directly incorporated into the next. This includes both baseline information on genetic affiliation or geographical location, as well as grammatical information such as the phonological and morphological properties of various grammatical markers (e.g., flexivity, host restrictions, and exponence).

In addition to assisting in the exploration of typological variation (Autotyp) and presenting the results in human-readable format (WALS), typological databases also open up the possibility of automatic processing of large numbers of typological facts in order to extract potential typological generalizations. Daumé III and Campbell

(2007) develop a Bayesian approach to this task, extracting both familiar implicational tendencies (e.g., Greenberg's 1963 #4: OV order tends to predict postpositions) and new candidates for consideration (e.g., plural prefixes predict noun-genitive order).

2.8 Summary

This section has presented examples across a variety of subfields of linguistics, in an effort to show how computational methods can help us take linguistic research to the next level. They allow us to work with more data, in multiple ways: annotate more data, more efficiently; through machine-mediated collaboration, construct larger, more cross-linguistic datasets; and systematically incorporate more data into analytical tasks than would otherwise be possible. Furthermore, computational methods allow us to verify the interaction of formal rules in the complex systems we model.

Though this overview has been necessarily incomplete—there are subfields we did not address as well as much excellent work in applying computational methodology to linguistic problems that we did not cite—we hope to have included enough to persuade linguists of any stripe that computers (and digitized data sets, computer networks, and specialized software) are useful tools for doing linguistics. This is true whether you are interested in formal or quantitative studies, linguistic competence or language processing and use, well-studied or under-resourced languages. More data is better, but only if we can work with it systematically. Computers provide assistance in maintaining systematicity as we scale up. In the next section, we provide an overview of existing infrastructure to support computational methods in linguistics, before turning to a vision of what that infrastructure could develop into, and a discussion of how we can work towards that vision.

3 Existing Infrastructure

Having illustrated with a variety of examples the ways in which computational methods can support linguistic analysis, we now turn to the issue of the infrastructure that supports those computational methods. As the primary benefit of computational methods is their ability to help us deal with large datasets systematically, the primary purpose of a cyberinfrastructure is to ensure access to data. For data to be accessible to computers, it must come in a standardized format. Accordingly, key components of the cyberinfrastructure include the following:

(i) Standards for data encoding

(ii) Standards for data annotation

(iii) Standards for metadata (discoverability)

(iv) Services for archiving and providing access to data sets

3.1 Annotation Standards

Point (ii) deserves some further comment: Each level of analysis in linguistics relies on previous analyses, down to phonetic transcription; in other words, one linguist's analysis is the next linguist's data. Annotation systems are a means of encoding analyses at one level so that they can be used as data in the next. In small scale work, it is often tempting to use tacit speaker knowledge to short-circuit some of this process, for example, using standard orthography where available instead of phonetic transcription for analysis in syntax, semantics and sometimes even morphology. We often even assume morphological analysis based on speaker intuitions rather than explicit mark up. But this doesn't scale up, as computers don't have access to tacit speaker knowledge, and it can be error-prone even at smaller scales.

Of course, for a cyberinfrastructure to provide access to data, linguists producing analyses must share those analyses so that they can become data for the next iteration. We return to the issue of establishing a culture of data sharing (i.e., of valuing contributions to the empirical base of our field) in §5.1 below. Here, though, we would like to look at the process of sharing from the perspective of annotation standards. In order for annotated data to be maximally useful, we need methods of representing analyses that we trust, that are robust across languages and linguistic theories, and that will scale to many kinds of use. The richer the structure of the data, the more the interesting questions that can be asked of it. However, the annotations have to be consistent (within a data set, and ideally across many similar data sets), in order to support the weight of further analysis. Annotation standards help provide consistency and sustainability within and across resources.[20]

As we develop the cyberinfrastructure for linguistics, new annotation standards will surely be required, but our field already has some: for segmental encoding, the IPA and more generally Unicode; for prosody and intonation ToBI (Silverman et al., 1992); for sublexical annotation, the Leipzig Glossing Rules for interlinear glossed text (Comrie et al., 2003) and the E-MELD "best practice" recommendations (Bow et al., 2003); for supralexical annotation, the various treebanks discussed further below and the Unified Linguistic Annotation (ULA) effort (Puste-

[20] On sustainability, see Rehm et al. 2009.

jovsky et al., 2005, Verhagen et al., 2007); and for resource discovery, the OLAC metadata standards (Bird and Simons, 2001).

3.2 The Development of DILI

In this discussion, we limit ourselves to the digital (computational) infrastructure that supports linguistic inquiry (DILI), putting aside the non-digital infrastructure of books, journals, papers and audio and video recordings, except as it relates to the digital one. DILI overlaps with the digital infrastructure that supports general-purpose inquiry (DIGPI). Each can be abstractly described as a network consisting of nodes, representing computing devices of various sorts, and links connecting them. In addition to these core infrastructures are others that keep them operating, including power systems, cable and wireless networks, and the facilities that manufacture, house and service the equipment.

DIGPI alone is often a sufficient resource for an individual researcher, teacher or student working on a personal computer or PDA connected to the Internet, and interacting with a commercial search engine to get a piece of information about a language they are interested in. However, DIGPI has been developed largely without the specific interests of linguistics communities (or those of most other scholarly communities) in mind, so that most queries of a technical nature are not likely to get an answer, much less a correct one, unless specifically linguistic digital resources and services are available and accessible.[21]

The current state of DILI is the result of largely uncoordinated efforts that have been made by individual researchers and research teams—often but not always with government funding support—linguistics departments and centers, research libraries and archives, private-sector research laboratories, and standards developers, who may or may not be working in collaboration with the International Standards Organization (ISO). Some of the work that has gone into the development of DILI has been, or has the potential to be, incorporated into DIGPI. For example, the Unicode Consortium standards for character encoding provide widely-available general-purpose support for most of the writing systems of the world's languages,[22] though much more work is needed to make many of these character sets avail-

[21] Though one should not underestimate the degree of linguistic sophistication that can be achieved by using general-purpose research tools on the massive amount of text data on the Internet (Halevy et al., 2009), particularly when additional information for interpreting the data is provided on certain websites (Potts and Schwarz, this volume).

[22] http://www.unicode.org/versions/Unicode5.1.0/

able, for example, for text-messaging on hand-held devices.[23] Similarly, the ISO 639-3 standard for the three-character identifiers for the world's languages has the potential of enabling all inquirers, not just specialists, to obtain accurate and up-to-date information about those languages.[24] Finally, it is worth noting that many of the resources that are now available in DIGPI have been developed by computational linguists, such as real-time text-to-speech, speech-to-text (including closed captioning) and machine-translation applications, and tools to support named entity recognition for identifying people, places, corporations, etc. in documents in a variety of languages. The development of these technologies typically requires large quantities of annotated or otherwise curated data. Major clearing houses for such data include the Linguistic Data Consortium (LDC)[25] and the Evaluations and Language resources Distribution Agency (ELDA).[26]

The elements that are specific to DILI can be broadly classified into linguistically enriched data sources, and services (or tools) for linguists to use to discover, aggregate or analyze such data. We have already mentioned in Section 2 some of these data sources and services. Data sources range from text and speech corpora in a variety of languages that have been annotated for features of linguistic interest to databases that have been designed to record the results of linguistic analysis of particular languages, for example the distribution of linguistic properties and relations across languages, and those that have been set up to record the results of experiments using linguistic materials. The use of specific data types characterizes both annotated corpora and linguistic databases, for example in treebank corpora, one may find phrase-structure trees (e.g., the Penn Treebank Project[27] (Marcus et al., 1993)), dependency-structure trees (e.g., the Prague Czech-English Dependency Treebank[28] (Čmejrek et al., 2004)), and directed graphs with feature-structures as nodes (e.g., the LinGO Redwoods Treebank[29] (Oepen et al., 2004)).

One data type that is central to many subfields of linguistics is the structure of interlinear glossed text (IGT) that was developed over the course of several decades to display in an informally agreed-upon

[23]The best-known linguistic standard is the symbol set of the International Phonetic Alphabet (IPA), which was developed long before there was a digital infrastructure. It has now received a Unicode encoding.

[24]http://www.sil.org/ISO639-3/codes.asp

[25]http://ldc.upenn.edu/

[26]http://www.elda.org/

[27]http://www.cis.upenn.edu/∼treebank/

[28]http://ufal.mff.cuni.cz/pcedt/doc/PCEDT_main.html

[29]http://wiki.delph-in.net/moin/RedwoodsTop

human-readable format the alignment of morphological forms with their meanings or grammatical functions in their occurrences in running text (see Section 2.3).[30] Because of the degree of consistency of IGT formatting in linguistic documents, Will Lewis was able to use standard text-harvesting techniques to collect a great deal of the glossed text that appears on the Internet, and with a certain amount of further processing has made much of the collection available in the ODIN database[31] for further research (Lewis, 2006).

When DILI resources were first being created, they were typically set up as self-contained objects without much thought given to integrating them with other resources. For example, digital lexicons for different languages were typically not designed to be comparable, except superficially, even if they were created using the same software package. Increasingly, however, such resources are being developed with the intent that they can be used together with other resources, enabling the data to be aggregated and further analyzed over the combined resources. Also, under certain circumstances it is possible to redesign and rebuild non-comparable resources so that they can be aggregated (Simons et al., 2004). Both of these changes—creation of sharable "born digital" linguistic resources and conversion of stand-alone (legacy) resources to sharable ones—are facilitated by the availability of digital infrastructures that support computation and collaboration with sufficiently large bandwidth and computational speed to make it increasingly seem that all the necessary resources are available in real time for all interested parties. In addition, since machines can perform computations over symbolic representations with the same facility and precision as over numerical ones, all that is required to enable them to do linguistic computations is to design the operations and represent linguistic properties and relations in such a way that the computations are performed as intended. These observations lead to our next topic.

4 A Linguistics Research Environment of the Future

We envision a future DILI that builds on the work that has been done so far, and provides among other things:

1. ready access large amounts of digital data in text, audio, and audio-video media about many languages, which are relevant to

[30] The Leipzig glossing rules (Bickel et al., 2008) constitute a best-practice set of recommendations for formatting IGT in linguistic documents. For discussion of XML representation of the structure of IGT, see Bow et al. 2003 and Palmer and Erk 2007.

[31] http://odin.linguistlist.org/

many different areas of research and application both within and outside of linguistics;

2. facilities for comparing, combining and analyzing data across media, languages and subdisciplines, and to enrich DILI with their results; and

3. services to support seamless collaboration across space, time, (sub)disciplines and theoretical perspectives.

Crucially, by data we mean here much more than "raw data", such as untranscribed sound recordings. For us the concept also subsumes analyses that the relevant communities of practice consider correct, or at least a sufficient basis for further inquiry, what we might call "enriched data". It is the responsibility of the various linguistic and other communities of practice to determine for themselves what results can be considered sufficiently settled to count as enriched data, rather than conjecture.

Since we envision machines being able to compute over all linguistic data, including enriched data, it must be interpretable in the metalanguage of linguistic description. The method by which this has been done, starting with the Brown Corpus (Francis and Kucera, 1964), the first serious effort to make linguistic digital data available to a broad research community, is through annotation, the explicit association of linguistic analysis with segments of data. The Brown Corpus annotations consist simply of part-of-speech "tags" for each word in the corpus, based on an analysis of English morphology that the community using the corpus found acceptable enough for their work. Over subsequent decades, the use of annotation tagsets extended to other linguistic domains and other languages, and what might be called a "theory of annotation" came to be developed, dealing with such issues as whether tags should be considered atomic or molecular (decomposable into elementary "features") in nature (Langendoen and Simons, 1995), and what is an acceptable level of accuracy in the assignment of tags in a corpus. Equally important, especially for cross-linguistic research, are questions of identity; for example, whether the past-tense tag in an English corpus represents the same concept as its counterpart in a Hindi one. However this last question gets settled, we must have a theory of annotation that allows us to say that for certain purposes, the past-tense concepts in various languages are sufficiently similar that they can be treated as the same (e.g. to answer a query like 'What languages mark past tense morphologically?'), while for others they must be distinguished.

As this example should make clear, annotations are to be understood as representing linguistic concepts that relate to one another in a network, so that in effect their meanings depend on their place in the network. It is not required that there be a single overarching network for all the annotations in DILI, but it would be desirable if sense could be made of the relations among conceptual networks for different annotation schemes, particularly those that represent different theoretical perspectives. We suppose, then, that items 2 and 3 above in our future DILI include facilities and services that encode the conceptual networks underlying the annotation schemes developed by linguistic communities of practice, and relate them to one another. This view of the role of conceptual encoding in a future DILI was recently articulated in Farrar and Lewis 2006, along with a plan for how to achieve it. Similarly, the Linguistic Annotation Framework (LAF; Ide and Romary 2004) and its extension, the Graph Annotation Format (GrAF; Ide and Suderman 2007), can be used to support this kind of interoperability, as well as the ability to map the structure of annotations over text between resources.

We would like the reader to imagine, then, a linguistics research environment of the future, with web services through which one could access alone or with partners around the world all of the following: analyzed, annotated texts and examples in all the world's languages, including child language and child-directed speech, associated with sound and video files; quantitative data from psycholinguistic experiments; detailed typological information; all searchable by language, linguistic feature and geographical region. What is the minimum amount of data you, the reader, would like to see on each language in such a system? How useful would it be if we got even only part way there?

Though that may sound too pie-in-the-sky, we note that similarly ambitious research environments exist today in other fields: Biochemistry has the Protein Folding Database,[32] Nanotechnology has the Nanomaterial Database,[33] and Astronomy has the National Virtual Observatory.[34] The astronomers are particularly eloquent in describing the benefits of aggregating data:

> All astronomers observe the same sky, but with different techniques, from the ground and from space, each showing different facets of the Universe. The result is a plurality of disciplines (e.g., radio, optical or X-ray astronomy and computational theory), all producing large volumes of digital data. The opportunities for new discoveries are greatest

[32] http://pfd.med.monash.edu.au/public_html/index.php
[33] http://www.nanowerk.com
[34] http://www.us-vo.org/

> in the comparison and combination of data from different parts of the spectrum, from different telescopes and archives.[35]

Analogously, linguists are all observing the same language faculty. What can we achieve, once we are able to efficiently combine perspectives?

We believe that a DILI along the lines that we have sketched out here will be developed, if for no other reason than that it fits with the kinds of digital infrastructures that are being developed across a wide spectrum of science, humanities and arts communities worldwide; see for example (Erny et al., 2009). However, it will happen more quickly and efficiently, if linguists, including computational linguists, begin to work together to bring it about.

5 What We Can Do Now

The previous section has presented a long term view of where we'd like to be. This section looks to the short term, and discusses what we can do now to take advantage of existing infrastructure and work towards the long-term vision presented above. In addition to participating in the workshops and other venues for discussion around developing the infrastructure, individual linguists can help work towards the linguistics research environment of the future by sharing data, by teaching, and by effecting culture change.

5.1 Share data

As discussed in §2, computers (and associated software and networks) are useful tools for doing linguistics, largely because they allow us to systematically handle much larger data sets. To gain that benefit, however, we need larger data sets. The most efficient way to build them is by pooling resources, i.e., sharing data, and sharing it in such a way that it is discoverable, accessible, and aggregatable. If the data in question are primary data collected from human subjects (as opposed to the addition of annotations on existing data sets, or primary data collected from public behavior, e.g., television shows), then the first step is to seek appropriate human subject permissions and consent to make the data distributable.[36] The second step is to use existing standards for data encoding wherever possible (and provide feedback to and/or

[35] "Harnessing the Power of Digital Data for Science and Society" 2009, p.11, http://www.nitrd.gov/about/Harnessing_Power_Web.pdf

[36] There are some situations in which data cannot be collected unless it is kept completely confidential. In such cases, it is simply not possible to contribute the data afterwards to larger collections. Here, our goal is to urge researchers to share their data whenever it is possible.

join the relevant working groups if the standards are not satisfactory). Finally, the third step is to publish data sets electronically, in existing repositories, or independently (see Simons 2008 for suggestions), but marked up with OLAC[37] metadata for discoverability.

In conversations with fellow linguists, we have heard many objections to or reasons for hesitation in sharing data. We would like to address some of them here, in the hopes of persuading more people to contribute to the field's empirical base.

Free-loading Data collection is extremely difficult, time consuming work, and though it is (often) eventually rewarded through the questions that can be answered once the data set is constructed, that reward can be a long time coming. Once a researcher has put the effort into collecting a data set (e.g., finding speakers, making recordings, doing transcriptions, collecting translations, producing glosses and other annotations), it is quite natural to want to squeeze all possible research results out of that data set before letting other linguists use it (and get the benefit of the hard work without having to do it). This situation is compounded by the fact that academic credit (e.g., in hiring decisions or tenure and promotion cases) is accrued for the research results derived from such data sets and not for the construction of the datasets themselves.

To this quite reasonable objection, we offer the following responses: First, we believe that the field is in need of culture change regarding the recognition of the value of providing data (see §5.3 below). A simple first step towards this culture change that data providers can take themselves is including instructions for citation prominently in the material being provided. Second, we point out that every linguist looks at data sets with different questions in mind: No one of us could think of every question it would be reasonable to ask of a given data set. Thus while it can certainly make sense to keep data private for some period of time, eventually it is beneficial, even for the linguist who did the original data collection, to open that data set up to fresh perspectives that come from different research directions as well as the possibility of aggregation with larger collections of data from the same or other languages. Finally, we would like to offer up the possibility of data-sharing collectives, to which researchers gain access by providing data themselves. In this way, both the hard work and the benefits of collecting new data are shared.

Incomplete/imperfect data sets In many subfields, data collection and annotation (including transcription) proceeds in parallel with data

[37] http://www.language-archives.org/

analysis. A side effect of this is that there is often no point in a research project when the data collection and annotation is finished. It can be difficult for researchers to either find the time to polish up such data sets (such time would be taken away from analysis and/or the beginning of a new project) or to feel comfortable publishing something that is incomplete or imperfect.

Here, we recommend first looking at the situation from the data consumers point of view: in many (if not most!) situations, incomplete or imperfect data sets are nonetheless quite valuable. If the alternative is no information about the language or variety in question (or even just less), we need not demand perfection. Second, we recommend establishing publishing systems that allow for editions of data sets. This would mean that a researcher could publish a preliminary version that could be superseded if/when s/he made corrections later. Thirdly, we recommend that metadata include information about the state of completeness of the annotations, so that data consumers have a sense of which parts of the data set are most reliable (and data publishers won't feel that they are being held responsible for every last detail). Finally, we recommend that relatively senior researchers set the precedent by releasing draft works.[38]

Support A third type of obstacle is the problem of supporting other users of one's own data. It is easy to imagine a linguist who would in principle happily share a set of field notes, recordings, etc., but just doesn't have the time to make copies to distribute or update the media/software to work with modern machines. Fortunately, this is not the role of the linguist, but rather that of archives, such as AILLA, the MPI Language Archive, the DOBES archive, and others.[39] Johnson (2002) presents an overview of AILLA's goals and plans for achieving them. Archives like AILLA merge the goals of preservation and access, both near-term and long-term. This includes the migration of legacy data to modern digital formats, on-going migration to new formats as they emerge, storage of redundant copies of all data at multiple locations, maintenance of metadata for resource discovery, delivery of materials (over the web and/or on digital media through snail mail) to users, and gate-keeping. Regarding gate-keeping, Johnson (2001) presents a graded access system that allows resource depositors to determine who will have access to the materials they deposit with the archive on a resource-by-resource basis.

[38]This idea was suggested by Heidi Johnson, p.c.

[39]OLAC maintains a list of language archives at http://www.language-archives.org/archives.php

In summary, we encourage readers to see these objections not as justifications for not sharing data, but as challenges to be addressed as we work to expand the empirical base of our field.

5.2 Teach

The second thing that can be done now is teaching, i.e., making sure that the next generation of linguists has the skills they need to take advantage of the emerging infrastructure. Once again we would like to emphasize that we are not advocating making all linguists into computational linguists. Rather, there are skills which do not differ much in complexity from the use of word processors, library databases, and search engines, but which are more specific to linguistics and therefore need to be explicitly taught. This could be done as part of a research methods class, or as units within introductory classes in different subfields, or some of each.

At a very high level, we believe that students need to know the following:

What resources exist This would include first an overview of what is available now in terms of corpora, endangered language archives, basic NLP tools (part of speech taggers, morphological analyzers, etc.), collections of IGT,[40] typological databases, etc. In addition, students should learn where new resources are likely to be announced,[41] so that they can stay 'tuned in'.

What standards/best practices exist It's easiest to comply with standards if you know what you're working with from the start, and we can save students lots of time by starting them off with best practices (and avoiding, e.g., recoding data later). Under this heading, we include things like knowing how to enter IPA as Unicode, the Leipzig glossing rules for IGT,[42] and the recommendations for digitizing endangered languages data compiled by the E-MELD project.[43]

Basic corpus manipulation tools There are a handful of very simple command-line text-processing tools, such as Unix 'grep' and 'wc', which can be very powerful aids in understanding what is happening in large collections of text files. Grep is a tool for searching for strings (or more generally, regular expressions) in texts. Wc (for 'word count') counts the number of characters, words (separated by white space), and lines in a file or set of files. Knowing how to handle these and similar utilities (and being comfortable with a command-line interface) allow

[40] e.g., ODIN: http://odin.linguistlist.org/
[41] e.g., LINGUIST List: http://linguist.org
[42] http://www.eva.mpg.de/lingua/resources/glossing-rules.php
[43] http://emeld.org/school/

linguists to do 'reality checks' on data collections when more complex software seems to be misbehaving.[44]

Basic database querying techniques Linguistic databases often have special-purpose user interfaces. Nonetheless, to the extent that they are also available as 'raw' SQL (or other) databases, it will be useful for linguists to know how use general SQL (or similar) queries. Basic familiarity with SQL allows users to ask questions not anticipated by the designers of the database (and its front-end).

Subfield-specific high-level programming languages These can take the form of "linguistic programming languages", i.e., machine-readable versions of linguistic formalisms. For example, Head-Driven Phrase Structure Grammar can be implemented in tdl (Type Description Language (Krieger and Schäfer, 1994)), which is interpreted by the LKB grammar development environment (Copestake, 2002) and other DELPH-IN tools.[45] Similarly, XFST (Beesley and Karttunen, 2003) provides a finite-state implementation of phonological rules in the style of (Chomsky and Halle, 1968). In other subfields, the equivalent might be statistical software packages, such as SPSS[46] or R.[47]

General computational skills There are a set of skills that computer programmers use that are somewhat peripheral to programming, and are helpful in any project using computers as an aid in managing complexity. These include version control (software for backing up various versions of a set of data, as well as integrating changes made by multiple collaborators), debugging (the process of systematically exploring what went wrong), and regression testing (establishing test suites to ensure that existing functionality/analytical coverage is not lost when a system is extended).

Finally, though none of the above entails requiring linguistics students to take a programming class or otherwise become computer programmers or computational linguists, it is important to encourage those who have an interest in that direction to do so. This entails identifying appropriate courses available at the university, and structuring both graduate and undergraduate curricula so that students can discover these avenues relatively early on and find time to explore them. Should there be resources available to teach a programming course tai-

[44] These are originally Unix utilities, but they are available for Windows as well, and of course Mac OS X is built on Unix.

[45] http://www.delph-in.net/

[46] http://www.spss.com

[47] http://www.r-project.org

lored specifically for linguists, the Natural Language Toolkit[48] (NLTK, Bird et al. 2009) and its associated book provide excellent resources for supporting beginning programmers interested in natural language processing.

5.3 Effect culture change

The third thing that linguists can do now to help bring about the linguistics research environment of the future is to work to effect culture change. The vision outlined here requires wide-spread buy-in from the field at large, for several reasons: First, it requires relatively large mobilization of resources, and that will be done more easily with broad support. Second, in order to build effective cyberinfrastructure (and at a smaller level, effective tools) we need linguists to participate in the design process. Third, and most importantly, as noted above, the cyberinfrastructure will only be interesting to the extent that it is populated with useful data. It follows that we need linguists to be motivated to contribute data.

This paper is overtly an attempt to promote culture change. Aside from writing such papers, there is much that can be done: First, we need to work to establish a culture of giving academic credit for creating, curating, and enriching data sets. This includes both small acts like being meticulous about citing the sources for all data that we use that we did not collect ourselves, and larger conversations within the linguistics community and with university administrators about how to give credit for such work in hiring decisions and tenure and promotion cases. Other venues for providing recognition include annual prizes, for which data providers could be nominated.

Second, we need to work to establish a culture of expecting data sets to be published. It is common in too many subfields of linguistics for analytical results to be illustrated with a few key examples, without the rest of the supporting data being available for others to examine. This bodes poorly for replicability of results in our science. As reviewers, of conference and journal submissions or books, we are in a position to ask for the data to be provided. Typically, the most practical means would be as an on-line appendix.[49] Likewise, when reviewing grant proposals, if any data collection is proposed, we should expect that provisions are made for disseminating the resulting data sets. It may not always be feasible or appropriate to do so (see §5.1 above for some discussion),

[48] http://www.nltk.org/

[49] Certain electronic journals, such as the *Journal of Experimental Linguistics* are already explicitly accommodating the publication of supporting datasets, programs, etc. along with articles they accept.

but often it will be; the expectation should be that the supporting data be published, unless there is some compelling reason otherwise.

Finally, we need to establish a culture of expecting claims to be checked against web-available data. Here again, it is as reviewers that we are best equipped to effect this aspect of culture change. If, for example, an author makes a claim about the co-variation of some typological properties, as reviewers we should expect this claim to be tested against the data in resources such as ODIN and/or WALS. Likewise, claims about the non-acceptability of certain sentence patterns should be backed up with corpus-based searches in languages where appropriate corpora are available. This is not because (non)attestation in a corpus necessarily implies (un)grammaticality, but because when considering structures in isolation, it is often difficult to come up with appropriately contextualized examples; corpus-based methods can turn up example types that would otherwise escape attention (see e.g., Baldwin et al. 2005 and van Noord and Bouma 2009). Once again, this is not always possible: There will always be interesting new claims for which further appropriate data are not yet available in the general cyberinfrastructure. But once again, that doesn't mean that when there is data available it can be ignored.

These three aspects of promoting culture change should interact with each other to produce a virtuous circle: The more we accord academic credit to the production of data sets, the more data sets will become available. The more data sets that become available, the more able we will be to check our claims against larger empirical bases. The more we check our claims against larger empirical bases, the more we will cite the original data sets. The more we cite the original data sets, the more academic credit will accrue to their producers, etc.

6 Conclusion

This paper has been written with the intent to persuade. In particular, we hope to have convinced the reader to try current computational methodologies, to teach students to use computational methodologies (and to advocate for inclusion of such instruction in linguistics curricula), and to collaborate in bringing about the next generation of cyberinfrastructure for linguistics. We've described a vision of cyber-enabled linguistics, and exemplified what it will allow us to do through a selection of research questions across a wide variety of subfields. (Along the way, we've emphasized that using computers as tools in doing linguistics is not the same thing as doing computational linguistics.) In order to realize this vision, we, as a field, need to build infrastructure, includ-

ing standards; contribute data; and promote and expect wide-spread use of cyberinfrastructure, as it is now and as new resources and tools become available.

Acknowledgments

We thank the Linguistic Society of America for hosting the Symposium on Computational Linguistics in Support of Linguistic Analysis at its annual meeting in San Francisco in January 2009, the audience at that symposium for stimulating discussion and two reviewers for helpful comments.

References

2009. Harnessing the power of digital data for science and society. Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. http://www.nitrd.gov/about/Harnessing_Power_Web.pdf.

Alishahi, Afra and Suzanne Stevenson. 2007. A computational usage-based model for learning general properties of semantic roles. In *Proceedings of the 2nd European Cognitive Science Conference*. Delphi, Greece.

Arnon, Inbal and Neal Snider. 2009. More than words: Speakers are sensitive to the frequency of multi-word sequences. Paper presented at the 83rd Annual Meeting of the Linguistic Society of America.

Baldridge, Jason, Sudipta Chatterjee, Alexis Palmer, and Ben Wing. 2007. DotCCG and VisCCG: Wiki and programming paradigms for improved grammar engineering with OpenCCG. In T. H. King and E. M. Bender, eds., *Proceedings of the GEAF 2007 Workshp*. Stanford, CA: CSLI.

Baldwin, Timothy, John Beavers, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2005. Beauty and the beast: What running a broad-coverage precision grammar over the bnc taught us about the grammar — and the corpus. In S. Kepser and M. Reis, eds., *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*. Berlin: Mouton de Gruyter.

Beesley, Kenneth R. and Lauri Karttunen. 2003. *Finite State Morphology*. Stanford CA: CSLI Publications.

Bender, Emily M., Stephen Clark, and Tracy Holloway King. to appear. Computational syntax. In T. Kiss and A. Alexiadou, eds., *Handbook of Contemporary Syntax*. Walter de Gruyter.

Bickel, Balthasar. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology* 11:239–251.

Bickel, Balthasar, Bernard Comrie, and Martin Haspelmath. 2008. The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. Max Planck Institute for Evolutionary Anthropology and Department of Linguistics, University of Leipzig.

Bickel, Balthasar and Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics*. Las Palmas.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. Oreilly & Associates Inc.

Bird, Steven and Gary Simons. 2001. The OLAC metadata set and controlled vocabularies. In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources-Volume 15*, pages 7–18. Association for Computational Linguistics Morristown, NJ, USA.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32:45–86.

Boersma, Paul and David Weenink. 2009. Praat: Doing phonetics by computer (version 5.1.05). [Computer program] Retrieved April 16, 2009, from http://www.praat.org/.

Bow, Cathy, Baden Hughes, and Steven Bird. 2003. Towards a general model of interlinear text. In *Proceedings of the E-MELD '03 Language Digitization Project: Workshop on Digitizing and Annotating Texts and Field Recordings*. University of Michigan.

Bybee, Joan. 2003. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14(03):261–290.

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.

Cieri, Christopher, David Miller, and Kevin Walker. 2004. The Fisher corpus: A resource for the next generations of speech-to-text. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal.

Čmejrek, Martin, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. 2004. Prague Czech-English Dependecy Treebank: Syntactically annotated resources for machine translation. In *4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal.

Cohn, David, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine Learning* 15(2):201–221.

Comrie, Bernard, Martin Haspelmath, and Balthasar Bickel. 2003. The Leipzig glossing rules. http://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Cook, Susan W., T.F̃lorian Jaeger, and Michael Tanenhaus. 2009. Producing less preferred structures: More gestures, less fluency. In *The 31st Annual Meeting of the Cognitive Science Society (CogSci09)*, pages 62–67.

Copestake, Ann. 2002. *Implementing Typed Feature Structure Grammars*. Stanford, CA: CSLI Publications.

Crabbé, Benoit. 2005. Grammatical development with XMG. In P. Blache and E. Stabler, eds., *Logical Aspects of Computational Linguistics*, vol. 3492/2005, pages 84–100. Springer.

Crouch, Dick, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. 2001. XLE documentation. On-line documentation, Palo Alto Research Center (PARC).

Daumé III, Hal and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Conference of the Association for Computational Linguistics (ACL)*. Prague, Czech Republic.

Denis, Pascal and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of EMNLP 2008*, pages 660–669. Honolulu, Hawaii.

Erny, Tim, Tim Morris, Cita Furlani, William Turnbull, Helen Wood, et al. 2009. Harnessing the power of digital data for science and society: Report of the Interagency Working Group on Digital Data to the Commitee on Science of the National Science and Technology Council.

Farrar, Scott and D. Terence Langendoen. 2003. A linguistic ontology for the semantic web. *Glot International* 7(3):97–100.

Farrar, Scott and William D. Lewis. 2006. The GOLD community of practice: An infrastructure for linguistic data on the Web. *Language Resources and Evaluation* 41:45–60.

Farrar, Scott and Steven Moran. 2008. The e-linguistics toolkit. In *Proceedings of e-Humanities–an emerging discipline: Workshop in the 4th IEEE International Conference on e-Science*. IEEE/Clarin, IEEE Press.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge MA.

Fillmore, Charles J., Christopher R. Johnson, and Miriam R. L. Petruck. 2002. Background to FrameNet.

Fong, Sandiway. 1999. Parallel principle-based parsing. In *Proceedings of Natural Language Understanding and Logic Programming (NLULP) 1999*, pages 45–57.

Francis, W. Nelson and Henry Kucera. 1964. *Brown Corpus Manual*. Providence: Department of Linguistics, Brown University.

Frank, Austin and T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)*. Washington, D.C.

Frank, Austin F., Celeste Kidd, Matt Post, Benjamin Van Durme, and T. Florian Jaeger. 2008. The web as a psycholinguistic resource. In *5th International Workshop on Language Production*. Annapolis, MD.

Gahl, Susanne, Dan Jurafsky, and Douglas Roland. 2004. Verb subcatergorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. *Behavior Research Methods, Instruments, & Computers* 36(3):432–443.

Godfrey, J.J., E.C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92.*, vol. 1.

Gómez Gallo, Carlos, T. Florian Jaeger, and Roger Smyth. 2008. Incremental syntactic planning across clauses. In *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)*, pages 845–850. Washington, D.C.

Greenberg, Joseph. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg, ed., *Universals of Language*, pages 58–90. Cambridge MA.

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24(2):8–12.

Han, Chung-hye and Anthony Kroch. 2000. The rise of *do*-support in English: Implications for clause structure. In *Proceedings of NELS 30*.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie, eds. 2008. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. http://wals.info.

Ide, Nancy and Laurent Romary. 2004. International standard for a linguistic annotation framework. *Journal of Natural Language Engineering* 10:211–225.

Ide, Nancy and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8. Prague, Czech Republic: Association for Computational Linguistics.

Jaeger, T. Florian, Austin Frank, Carlos Gómez Gallo, and Susan Wagner Cook. 2009. Rational language production: Evidence for uniform information density. In *Proceedings of the 83rd Annual Meeting of the Linguistic Society of America*. San Fransisco, CA.

Jannedy, Stefanie and Jennifer Hay, eds. 2006. *Modeling Sociophonetic Variation*. Elsevier, Inc. Special issue of *Journal of Phonetics* 34(4).

Johnson, Heidi. 2001. Graded access to sensitive materials at the Archive of the Indigenous Languages of Latin America. In *Proceedings of the IRCS Workshop on Linguistic Databases*. Philadelphia.

Johnson, Heidi. 2002. The Archive of the Indigenous Languages of Latin America: Goals and visions for the future. In *Proceedings of the Language Resources and Evaluation*. Las Palmas.

Krieger, Hans-Ulrich and Ulrich Schäfer. 1994. TDL–a type description language for constraint-based grammars. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 893–899. Kyoto, Japan.

Kroch, Anthony and Ann Taylor. 2000. *Penn-Helsinki Parsed Corpus of Middle English*. Philadelphia: University of Pennsylvania, 2nd edn.

Langendoen, D. Terence and Gary Simons. 1995. A rationale for the Text Encoding Initiative: Recommendations for feature-structure markup. *Computers and the Humanities* 29:191–205.

Lewis, William D. 2006. ODIN: A model for adapting and enriching legacy infrastructure. In *Proceedings of the e-Humanities Workshop, Held in cooperation with e-Science*. Amsterdam.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates, 3rd edn.

MacWhinney, Brian. 2008. Talkbank—reintegrating the disciplines. Presentation at the London Linguistics circle, 2/11/2008.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19:313–330.

Mihalcea, Rada and Dan I. Moldovan. 1999. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 152–158. Morristown, NJ, USA: Association for Computational Linguistics. ISBN 1-55860-609-3.

Miltsakaki, Eleni, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The penn discourse treebank. In *In Proceedings of Language Resources and Evaluation Conference 2004*.

Moon, Taesun and Katrin Erk. 2008. Minimally supervised lemmatization scheme induction through bilingual parallel corpora. In *Proceedings of the International Conference on Global Interoperability for Language Resources*, pages 179–186.

Moran, Steven. 2009. An ontology for accessing transcription systems (OATS). In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 112–120. Athens, Greece: Association for Computational Linguistics.

Moran, Steven and Richard Wright. 2009. PHOIBLE: PHOnetics Information Base and Lexicon. Accessed 4/23/2009.

Nakhleh, Luay, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81(2):382–420.

Oepen, Stephan, Daniel Flickinger, Kristina Toutanova, and Christopher D. Manning. 2004. LinGO Redwoods. A rich and dynamic treebank for HPSG. *Journal of Research on Language and Computation* 2(4):575 – 596.

Palmer, Alexis. 2009. *Semi-Automated Annotation and Active Learning for Language Documentation*. Ph.D. thesis, University of Texas at Austin.

Palmer, Alexis and Katrin Erk. 2007. IGT-XML: an XML format for interlinearized glossed texts. *Proceedings of the Linguistic Annotation Workshop* pages 176–183.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles.

Pustejovsky, James, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. In *IWCS-5 Fifth International Workshop on Computational Semantics*.

Pustejovsky, James, Adam Meyers, Martha Palmer, and Massimo Poesio. 2005. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and coreference. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 5–12. Ann Arbor, Michigan: Association for Computational Linguistics.

Rahman, Altaf and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of EMNLP 2009*, pages 968–977. Singapore.

Rehm, Georg, Oliver Schonefeld, Andreas Witt, Erhard Hinrichs, and Marga Reis. 2009. Sustainability of annotated resources in linguistics: A web-platform for exploring, querying, and distributing linguistic corpora and other resources. *Literary and Linguistic Computing* 24(2):193–210.

Riggle, Jason, Maximillian Bane, James Kirby, and Jeremy O'Brien. 2007. Efficiently computing OT typologies. Paper presented at the 81st Annual Meeting of the LSA.

Roland, Douglas and Daniel Jurafsky. 1998. How verb subcategorization frequencies are affectedy by corpus choice. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 1122–1128.

Sankoff, David. 1975. Varbrul version 2.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. ToBI: A standard for labeling English prosody. In *Second International Conference on Spoken Language Processing*. ISCA.

Simons, Gary, William Lewis, Scott Farrar, Terence Langendoen, Brian Fitzsimons, and Hector Gonzalez. 2004. The semantics of markup: Mapping legacy markup schemas to a common semantics. In G. Wilcock, N. Ide, and L. Romary, eds., *Proceedings of the 4th Workshop on NLP and XML (NLPXML-2004)*, pages 25–32. Barcelona, Spain: Association for Computational Linguistics.

Simons, Gary F. 2008. The rise of documentary linguistics and a new kind of corpus. http://www.sil.org/∼simonsg/presentation/doc%20ling.pdf.

Sinha, Ravi and Rada Mihalcea. 2009. Unsupervised graph-based word sense disambiguation. In N. Nicolov and R. Mitkov, eds., *Current Issues in Linguistic Theory: Recent Advances in Natural Language Processing*.

Stabler, Edward P. 1997. Derivational minimalism. In C. Retoré, ed., *Logical Aspects of Computational Linguistics*, pages 68–95. Springer.

Tardif, Twila, Marilyn Shatz, and Letitia Naigles. 1997. Caregiver speech and children's use of nouns versus verbs: A comparison of english, italian, and mandarin. *Journal of Child Language* 24(03):535–565.

van Noord, Gertjan and Gosse Bouma. 2009. Parsed corpora for linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 33–39. Athens, Greece: Association for Computational Linguistics.

Verhagen, Marc, Amber Stubbs, and James Pustejovsky. 2007. Combining independent syntactic and semantic annotation schemes. In *Proceedings*

of the Linguistic Annotation Workshop, pages 109–112. Prague, Czech Republic: Association for Computational Linguistics.

Wassink, Alicia Beckford. 2006. A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties. *The Journal of the Acoustical Society of America* 119:2334–2350.