Issue Statement for Digital Tool Summit in Linguistics

Lori Levin
Language Technologies Institute, Carnegie Mellon University

Issue: We need fast ways to create tools for more languages: Treebanks (corpora annotated with phrase structure or dependency structure trees) have been used as research tools in computational linguistics since the 1990s. Many treebanks include search tools. For example, TGREP and TGREP2 allow users to describe tree templates and then find treebank examples that match the templates. Such searches are valuable for studying many syntactic phenomena including coordinate structures, subcategorization patterns, long distance dependencies, and many kinds of special constructions. Treebanks, however, are of limited size. More recently, robust parsers have become available for some languages. These parsers can be used to parse large corpora (anything that can be found on the Web) with reasonable accuracy, thus expanding the amount of annotated text that can be searched. Unfortunately, treebanks and/or robust parsers are available for only about a dozen languages (English, Chinese, Dutch, Arabic, Czech, and a few others).

1. **What are the most pressing needs among possible cyberinfrastructure and/or digital tools for linguistics?** Basic tools for more languages, including morphological analyzers, lexicons, corpora, annotated corpora and parsers. A broader range of corpus annotations (syntactic, semantic, and pragmatic) is also needed.

2. **What are some enduring challenges in creating cyberinfrastructure and/or digital tools for linguistics?** The time consuming process of adapting parsers and annotating corpora for new languages. The challenge is to develop tools that will partially automate and speed up the process of developing resources for new languages.

3. **Which existing resources can be leveraged to create digital tools for linguistics?** Linguists should learn the capabilities of machine learning techniques for automatically learning grammars and other linguistic structure. Machine learning techniques for NLP seem initially repugnant because many of them are lingusitics-free. However, they produce analyses of data that can support linguistic research. Linguistis don't need to understand how they work, just what they can do for you. They can speed up the creation of resources for new languages.

4. **How can documentation tools make language resources (e.g. text, lexical or morphological corpora) more readily available for historical, typological, and other theoretical analyses?** Annotated corpora are obviously useful, but computational linguists should build easy-to-use interfaces for non-computational linguists. A good example is Philip Resnik's Linguist's Search Engine.

Biography Lori Levin is an Associate Research Professor in the Language Technologies Institute in the School of Computer Science at Carnegie Mellon University. She received a B.A. degree in Linguistics at the University of Pennsylvania in 1979 and a Ph.D. in Linguistics from MIT in 1986. Her thesis research was on the lexical mapping theory of Lexical Functional Grammar. After teaching Linguistics at the University of Pittsburgh from 1983 to 1988, Levin took a job at CMU in the Center for Machine Translation, which later became the Language Technologies Institute. Since 1988, she has conducted research on Machine Translation of spoken and written language, information extraction from email, and computer assisted language learning. She has supervised projects on at least 12 languages including European, Asian, Semitic, and Native American languages.

In 2001 Levin became a co-PI of an NSF ITR project, AVENUE along with Jaime Carbonell (PI) and Alon Lavie (co-PI). AVENUE automatically learns rules for machine translation from small parallel corpora. AVENUE includes a set of tools that are useful outside of the field of machine translation including an ontology of functional/communicative meanings and an elicitation tool that allows informants to translate sentences and align words.