

Toward the Interoperability of Language Resources

July 13-15, in conjunction with the 2007 LSA Summer Institute



Beyond primary documentation: facilitating exchange of linguistic data

Stephen Grimes
Indiana University

I view the issues of interoperability through the lenses of both a computational linguist and a field linguist. In the course of my work as a graduate student, I have interacted with linguists and anthropologists involved with fieldwork, documentation, and analysis of linguistic data. At the American Indian Studies Research Institute (AISRI), since 1985 we have been compiling not just dictionaries and grammars of Siouan and Caddoan languages, but we also have been managing an extensive collection of sound recordings. Since the late 1990s, we have also been developing software to manage dictionary databases and morphologically parse our linguistic data. However, as time passes, the danger exists that our print-based data may outlast our electronic data due to issues of interoperability.

Despite the international trend towards open source code and open source standards, the majority of our linguistic data at AISRI continues to be stored in Visual FoxPro databases. While we have built a tool to export our data into an XML format, there is no easy way to import data, even data that we have exported ourselves. We face the inevitable possibility that our software will fail to compile on future operating systems, although virtual machines may help to alleviate this issue. While the Unicode standard continues to gain acceptance, we must still use in-house fonts for the American Indian languages due to the lack of universal fonts. In the future using Java or other platform independent solutions may be the answer, but at this point at AISRI, we must honestly evaluate whether there is a desire in the greater community for us to continue to develop our software for distribution or whether better partial alternatives exist and can be relied upon.

My personal stance is that many of the issues we face will not be solved by linguists alone, but by the IT community at large. Having said that, it is imperative that linguists and support programmers continue to meet negotiate standards for the exchange of linguistic data. While tagged corpora primarily exist for Indo-European languages, I am particularly interested in developing and refining tagging standards for easy comparison of structurally disparate languages in a manner that is linguistically atheoretical. I also eagerly await the recommendations and designs developed by participants in the Digital Tools Summit in Linguistics in 2006. Furthermore, as a computational phonologist, I am interested in having



conversations about standards with those developing phonological lexicons such as those appearing in CELEX2.

Allow me to emphasize an issue which I feel has been generally less-stressed in the linguistic community: interface design. As our software tools at AISRI have continued to expand in functionality, so too have the number of buttons, radio buttons, checkboxes, dialog boxes, and wizards. I personally feel that the lack of both application-based help and an intuitive interface has prevented our otherwise useful software from being more widely adopted in the community. With an eye towards less technically savvy linguists, we should be developing tools that do not require steep learning curves. Every linguist should not be expected to be a programmer, but those of us who know how to develop GUIs to interface language data should do so with a big tent philosophy. I am specifically thinking of GUIs and web-based database interfaces which mimic the power of regular expression searches of a corpus from the command line. Linguists have always been data-oriented people, and it is imperative for us to share our vast amounts of data.

