

## Leveraging the Existing Linguistics Cyberinfrastructure

The efforts to develop best-practice encoding standards and the tools that service these standards are noble and worthy endeavors. In particular, the field is in great need of data standards, not only to ensure interoperability between datasets—a move that in itself could usher in a new age of linguistic inquiry—but also to ensure that linguistic data endures. The crucial problem is convincing the greater linguistics audience to “make the leap” and adopt these new standards and use these new tools.

What is missing from most tool and standards development initiatives is a clear, relatively painless migration strategy from the current environment to the next. The focus of efforts has been on the endpoint, on engineering the environment of the future, with little or no emphasis on how to move linguists and their data from the present to that future. The consequence? Without achieving the necessary critical mass, both the tools, and the standards that underlie them, themselves risk obsolescence.

However, wide-scale adoption of new tools and technologies can achieve a degree of success if we recognize that a large scale, multi-lingual linguistics infrastructure already exists: the body of data and scholarly work on the Web. It is the rare linguist who does not record language data and analysis electronically, and increasingly, these data and analyses are being disseminated over the Web. This large database of multi-lingual data could conceivably be interoperated over *as it exists*, providing a testbed not only for (semi-)automated strategies for data migration and data reuse (even if only on-the-fly), but, if the output is properly encoded, could serve as a large-scale testbed for the very set of next generation tools that are being developed. Adapting existing and mature computational technologies, such as structured and unstructured text mining, clustering techniques for context-dependent term disambiguation, or automated methods for language identification, coupled with fairly robust machine learning strategies, makes it possible to do large scale mining and migration of the particular data types that exist on the current “linguistics Web”.

Potential candidates for large-scale data migration are semi-structured linguistic text objects, such as snippets of interlinear text. Capturing not only the objects themselves, but the relevant bits of the surrounding data and analysis, and including whatever citation information that can be found (both to help with term disambiguation and for appropriate source attribution), a database of such text objects could provide an infrastructure for testing large scale query and data interoperation methodologies, as well as the migration strategies themselves.

Thus, I propose that the move to a future cyberinfrastructure rests crucially on leveraging existing infrastructure and tools, focused first on particular, widely used structured data types, and adapting proven and tested computational methods to the greater goal.