

Towards an extensible framework for the creation of linguistic tools

Jan-Torsten Milde
milde@fh-fulda.de

April 10, 2006

1. Statements

Good scientific tools are a prerequisite for scientific progress. Without good tools, a scientist is severely restricted with respect to the amount of processable data, the applicable methodology and the accuracy and correctness of the expected results.

Today, most scientific tools are realized as computer programs, processing digitized "raw" data created by some kind of sensoric device (e.g. a microphone, a video camera). These programs enable scientists to organize, analyze and visualize the data relevant to their specific domain. For the natural sciences a multitude of specialized programs have been developed. These tools are effectively used in every day scientific work.

This is completely different for the humanities. Here, the scientific community of a specific subfield is often very small in numbers. The scientists are excellent experts within their field, but usually are not trained to design and implement large and complex computer programs. As a result it becomes almost impossible to establish a tool development process based on the open source principles. In addition developing specialized scientific software is not attractive for commercial software companies, as the expected financial back flow is minimal. This dilemma results in a high demand for good tools.

1. What are the most pressing needs among possible cyberinfrastructure and/or digital tools for linguistics?

- *Stable tool platform.* Currently no generic platform for developing linguistic tools exist. When designing a new tool, developers often create their own system framework, which, over time becomes unmanageable and eventually collapses. Many programs are developed by single persons. As tool development is not regarded to be a "central" issue of linguistics, the development often ends once the scientist moves on to other fields.
- *Extensability and interoperability.* Linguistic tools are generally created and optimized to fulfill a single limited task. Even if the tools are available as open source (which is not very common), it usually is very hard to integrate or extend them due to missing documentation.
- *Adaptation to linguistic work flows.* Scientists in different areas of linguistics tend to use differing approaches for setting up corpora. Even within a single linguistic area there is much debate on

how to setting up a corpus "correctly". Most of the current tools are not capable to adapt towards these differences in work flow. Even worse, no formal description for these work flows exist.

2. What are some enduring challenges in creating cyberinfrastructure and/or digital tools for linguistics?

- *Corpus creation is a dynamic process.* For good reasons corpora are nowadays stored in XML based formats. The file formats are therefore structurally bound to the properties of XML, c.f. the data is organized in a tree structure.

The linguistic *interpretation* of the annotated data almost always takes multiple layers into account, ignoring the tree structure of the data format. The syntax of the data format therefore differs from perceived linguistic data structures. Even more important, the interpretation is not fixed. It changes from project to project and even within a project.

In order to establish an *agile corpus creation process*, this syntax to semantic mapping has to be made explicit. Unfortunately, how this can be achieved is still unclear.

A solution might be to adopt the idea of *early queries*, i.e. by integrating querying into the design phase and the data collection phase. By continuously formulating queries in a precise formal way, it becomes possible to adjust the corpus structure towards the intended usage. Errors in corpus design can be detected and corrected.

3. Which existing resources can be leveraged to create digital tools for linguistics?

- *Eclipse Workbench.* The robust extensible software infrastructure provided by Eclipse (see www.eclipse.org) makes it possible to efficiently design and implement linguistic applications. Eclipse provides excellent components (e.g. configurable editors, tree views) which are perfectly matching the structures of the scientific data under investigation. Existing tools are easily integrated into an Eclipse application. Furthermore, Eclipse-based tools are well portable to a number of operating systems.
- *UIMA.* UIMA is an open, industrial-strength, scaleable and extensible platform for creating, integrating and deploying unstructured information

management solutions from combinations of semantic analysis and search components.

- *Native XML databases.* Being able to efficiently store and query the XML-annotated corpora is a central demand. Native XML database systems (e.g. Tamino) use XQuery to query the stored XML structures. Instead of designing yet another linguistic query languages, linguists should translate the domain specific query languages into XQuery and rely on the high performance database systems.