

Issue statement: Challenges in creating digital tools for linguistics

Susanna Imrie, SIL International
Susanna_Imrie@sil.org

The challenges for developers of software for linguistic research are wide-ranging and complex.

The top challenge is how to design software which is incredibly powerful in regard to the range of linguistic concepts and tasks it supports, yet has an intuitive, easy-to-use interface.

For example, how should a tool for interlinearization of texts interact with a corresponding digital lexicon to give a reasonable interlinear display at the phrase level? Each individual word in the phrase can be analyzed separately and linked to lexical entries. In addition, the combination of those words should be linked to a lexical entry for the phrase (e.g. idioms.) But how does one construct the interface for entering such analyses and then for displaying the results? Various options are possible. The fact that phrasal idioms can be discontinuous only serves to complicate the problem.

Another issue is providing tools for efficient development of large enough lexicons to adequately represent a language. The “dictionary development process” (see <http://www.sil.org/computing/DDP/>) is a new approach aimed at supporting rapid development of a sizable dictionary. A workshop is held among native speakers in which words are elicited by semantic domains. The next stage is to make iterative improvements to the very sketchy, draft lexicon; a number of bulk edit operations are carried out which gradually hone the data.

This approach (parts of which are being implemented in the SIL Fieldworks Language Explorer) raises the question of what operations (both in dictionary development and other aspects of linguistic research) would be useful to do with bulk edit as oppose to entering data one item at a time? For example, some morphological analysis could be done this way:

1. Filter for wordforms that end in “ment”.
2. Review words returned by the filter and remove any for which the “ment” is not a derivational suffix that derives a noun from a verb.
3. If necessary, create a new lexical entry for suffix “-ment”.
4. Specify that the “-ment” morpheme is the analysis of the “ment” part of the filtered words and that the words should be assigned the category of noun.
5. Make new lexical entries for the remaining “stem” part of the words, for each one that does not exist yet and assign them to the category of verb.

It would be great if an integrated computer tool could notice such patterns and bring them to the attention of the user, for example: “There are 300 wordforms that end in ‘ing’. Could ‘-ing’ be a suffix?”

Once work on dictionary development is completed in the editing environment, the problem of publishing presents itself. How does the linguist get from database to printed dictionary? What typesetting tools are recommended and what is the path for getting the data into the format needed by that tool? This is a particularly challenging problem when the dictionary involves an orthography not supported by current operating systems.

Another publishing format that needs to be in focus in today’s world is the internet. What are the best ways to publish lexicons on the internet? Given that the constraints of paper are not an issue in this environment, what are new ways of presenting data that should be made use of,?

Finally, there is the matter of the long term preservation of digital dictionaries. What should developers of a lexical database provide to facilitate users archiving their data?

These are just some of the challenges facing developers of software for linguistic research.