

Building An Integrated Digital Tool for Language Resources

Anil Kumar Singh

Language Technologies Research Centre
IIIT, Gachibowli, Hyderabad, India-500032
anil@research.iiit.net

More and more linguists are now using computers and more and more computer scientists are becoming interested in language. The border between linguistics and natural language processing (if there was one) is now highly porous. In fact, it's not like a border at all — it's a discipline in itself. This is understandable since a lot of the work that linguists have to do can be done more effectively by using computers.

Many language resources are now available in electronic form and there are many tools for helping linguists perform their work. However, there are issues that need to be addressed before we can gain from using digital tools and electronic language resources as much as is possible. And there is a lot that is possible, even in the near future.

Perhaps the first range of problems that we should address are the small things. These are problems related to such 'minor' details like encodings, formats, convertors, input methods, and the like. For example, Indian languages use many standard and non-standard or proprietary encodings which make it very difficult to deal with, or even type, text in those languages. Even if it's a mere nuisance, not worth the attention of linguists and computers scientists, it's still a major bottleneck in working with Indian languages in any way. Similarly, the problem of varying formats for essentially the same kind of annotation also hinders proper use of language resources. This category of problems are technically not too difficult to address. We should get them out of the way, even if they are not very interesting to solve.

Interoperability of resources is also a major issue, not just in terms of formalism, but also the content. For example, many different kinds of lexical resources are available (WordNet, FrameNet, VerbNet, dictionaries, etc.). If they could all

be linked together, the benefits would be really worthwhile the effort. There has been some effort in this direction, but what we need is a seamless solution integrating the lexical resources.

To solve such problems, we will first need to study in details the needs of linguistic scholarship on the one hand, and to study the available resources on the other hand. Then we will have to find some common ground, which can form the basis for building integrated (or at least integrable) digital tools that allow resources to interoperate. To take a concrete example, if we are interested in the word 'book', we should be able to get all the information about it from various resources. This would include the synsets/hypernym/hyponyms etc. from the WordNet, verb frames from FrameNet, verb classes from the VerbNet, senses and example sentences from a dictionary, morphological information from somewhere else perhaps, and so on. This is not all. We should also be able to get the occurrence and co-occurrence statistics about this word from the corpora. The researcher should be able to do this easily and the results should be presented in human-friendly way.

What we need is basically a wrapper like (but with extra functionality) API for language resources, on top of which there is a user-friendly graphical user interface (GUI). This will allow a non-technologically-oriented language researcher as well as a software developer to get easy access to many (if not all) language resources in one place.