The only option is open: Why should language technology and resources be free[*]?

Francis M. Tyers
Grup Tranducens
Dept. Lleng. i Sist. Inform.
Universitat d'Alacant

28th April 2011

Abstract

I would like to structure this paper in three parts. The first deals with how we use language technology resources, and impress that especially for marginalised and minority languages, these resources cannot exist in a vacuum. The second describes some of the principle problems faced by language technology and resources. Finally, I argue that the only viable option for the language technology sector in the Nordic countries is one of openness and free distribution.

First some definitions, when referring to *language technology*, it is taken to mean the software on which applications are based, for example a machine translation (MT) or spell-checking engine. When referring to *language resources*, it is taken to mean the data on which these application depend. For example, for a spellchecker, the dictionary, morphological rules, and error models. For a machine translation system, either the parallel corpora (if the engine is corpus based), or the dictionaries and rules (if it is rule based).

Both language technology and the resources on which it depends are interdependent. A spellchecking engine is no use without the data to run on it, likewise, a spelling dictionary is of limited use without the engine to run it.

There are three main problems facing language technology and resources. The first is *visibility*, or 'can the people who are looking for the resource find it?', the second is *availability* 'can it be used for what they want to use it for?' and finally *sustainability* 'will the resource still be available next year ... or in ten years?'

Imagine you have developed a spellchecker for a language, but it is not used because no-one knows about it, or worse still. Perhaps there is an existing spellchecker, which is no longer maintained but is more widely used because it is easier to find, or comes pre-installed. This is the problem of *visibility*.

[*] Free here refers to *freedom*, not *price*.

On the other hand, perhaps you are planning to work on machine translation systems between Swedish and the immigrant languages of Sweden. You find a source of bilingual lexica between Swedish and Kurdish, Swahili and Pashto, but they cannot be used because of prohibitive licensing terms. This is the problem of *availability*.

Finally, you develop a morphological disambiguator during a government-funded project. The project funding expires and work comes to a halt. There is no one left to make sure that the disambiguator is *visible* and *available* to other researchers and developers. This is the problem of *sustainability*.

For larger languages, these problems can be sidestepped by starting from scratch each time. As a result of the amount of funding available, and the larger number of speakers, the amount of effort expended in making a toolchain from scratch can be fairly minimal. One person year from a speaker population of 400 million is substantially more likely to be fundable than one person year from a speaker population of five hundred. Especially if the cost of specialist training is included – there are much more likely to be ready-trained linguists or programmers in a larger population.

This is still a tremendous duplication of effort. Furthermore, *availability* of resources for larger languages can have a direct effect on language technology for minority and marginalised languages. Consider for example the creation of multilingual applications, machine translation and bilingual dictionaries. If we want to create a dictionary of South Sámi and Finnish, then dictionaries of South Sámi and Norwegian and Norwegian and Finnish are likely to be useful – if they are available.

So, what are the solutions? The primary solution to all of these problems has been outlined very effectively by Scannell et al. (2006), the *pool*.

Bibliography

- Pedersen, T. (2008) ‘Empiricism Is Not a Matter of Faith’. *Computational Linguistics* 34(3), 465–470.

- Scannell, K., Streiter, O. and Stuflesser, M. (2006) ‘Implementing NLP Projects for Non-Central Languages: Instructions for Funding Bodies, Strategies for Developers’ *Machine Translation*