Scientific Report on NetWordS Exchange Visit

Harald Hammarström

November 8, 2012

1 Referential Information

This report concerns NetWordS Short Visit Grant 4779. The place, time and duration of the visit was CNRS, Villejuif, Paris, 17-22 September, 2012.

2 Scientific Report

2.1 Purpose of Visit

The purpose of the visit was to bring together one computational linguist interested in the languages of the world and one linguist specializing on African languages (plus associates) interested in computational methods.

In particular, we wanted to discuss a) the role of morphological tone and computational theories of learning tone patterns from examples, and b) cross-linguistic distributional properties of segments.

In addition to discussing ideas, we also wanted to pool resources, do demonstrations and agree on representations and formats.

2.2 Description of the work carried out during the visit and the main results obtained

One of the aims of the original plan was machine learning of morphology on languages with non-concatenative morphology, specifically African tone languages. This problem had been solved in the meantime, in the sense that tonal suprasegmentally inflecting words can be rewritten automatically as sequences of segments, and can therefore be addressed with traditional

methods developed for segmentally inflecting languages. This is because in all or nearly all tone languages, all and only tone marking and tone-bearing segments can be identified orthographically. While this seems obvious now, it was not clear from the outset that all tonal languages under consideration actually behave this way and that there would not be complications below the surface. Another researcher at CNRS met with during the visit, Nicolas Quint, is describing Koalib, a Heiban language with tonal inflection for case and has a concrete problem in accounting for tonal paradigms in the most economical and complete manner. He needs to finish curating the data (nearly 6000 items) and then we will apply machine learning methods to the (segmental translation of) this dataset.

Given this, the discussion on tone and the cross-linguistic distribution of segments quickly shifted to the more specific topic of segments used cross-linguistically for pronouns, and its role in comparative-historical linguistics. It was particularly useful that other researchers interested in this topic were present, in conjunction with the International Conference on the Reconstruction of Proto-Niger-Congo.

Ségérer's very elaborate database interface was demonstrated and explained.

Lexical reources on African languages were pooled.

We discussed a number of computational approaches to comparative-historical reconstruction. At least two new approaches will be tested on Ségérer's very large curated data collection.

2.3 Future collaboration with host institution (if applicable)

It is clear, especially from the presentations in the co-located conference, how difficult and gigantic a task it is to do comparative-historical reconstruction of large sets of languages in the classical manual way, i.e., without the assistance of computers. The interest in a long-term collaboration on this topic runs deep from both sides. Our next meeting is already scheduled to be this year.

2.4 Projected publications / articles resulting or to result from the grant

We foresee two publications as a result of the meeting

- One concerning the distrubtion of consonantal segments in pronouns cross-linguistically.
- One concerning algorithms for subgrouping of related languages on the basis of lexical innovations.

2.5 Other comments (if any)

The terms of the exchange programme and what costs it covers have been changing from the first call posted on LINGUIST LIST. We had to send the same information more than once and answers did not come on the dates they were advertised. This is annoying, and the administrative overhead is too large in proportion to the size of grant programme.