Report: Short Term Scientific Mission

COST-STSM-ECOST-STSM-IC1207-170214-038491

Sebastian Sulger

March 17, 2014

1 Purpose of the STSM

The STSM focused on a special type of multiwords: those that carry a special meaning with respect to discourse analysis. I have come across these items while working on the VisArgue project[1], focusing on the analysis of German political discourse. Initial examination of these items has revealed that they may realize some of the functions that are usually taken up by discourse particles, focus particles and connectors.

With respect to these multiwords, the mission aimed at three distinct issues. First, the multiwords at issue may be realized using various syntactic constructions (prepositional phrases, adverbial clauses, tag questions etc.), which makes it hard to identify them in running text. Second, in a careful linguistic analysis, we would like to not only identify the relevant multiwords, but also provide them with the correct treatment in terms of features or tags, lumping them together with single-word particles and connectors. The third issue that comes up concerns their disambiguation: since many of the multiwords in question have an alternate, purely compositional reading, we have to wonder how to separate apart the compositional from the non-compositional occurrences.

2 Description of the work accomplished during the STSM

During several meetings with the STSM host it was decided that a possible first step in identifying candidates for discourse multiwords consists in making use of available large-scale parallel corpora. Word alignments are especially useful in this task.

[1] http://www.visargue.uni-konstanz.de

On a theoretical note, the status of some of the phrases at issue is debatable. During the discussions with the host, it became clear that work on discourse multiwords must pay close attention to whether the phrases are in fact lexicalized/frozen idioms (in which case they would qualify as multiwords), or whether they allow a certain degree of modification (in which case they may not qualify as multiwords).

With respect to the disambiguation of ambiguous multiwords (discussed above), the host and I agree that machine learning experiments may provide clues as to how to tell apart true multiword usages from non-multiword/compositional usages. Features that may help disambiguating include context window strings, context window parts of speech, etc.

We also had a detailed look at the way multiwords are handled in the large-scale HPSG grammar developed at CSLI, the English Resource Grammar. Here, we payed special attention to mechanisms provided by that environment for handling various degrees of fixedness vs. variation in idiomatic expressions and multiwords.

During the second week of my STSM, I had meetings with members of the ParGram consortium[2], where I discussed varying ways of implementing multiword items in the LFG parser XLE.

3 Description of main results obtained

During the first week, the host and I have defined a corpus study that implements a lookup strategy for identifying discourse multiwords using the EuroParl corpus.[3] There are precomputed English-German word alignments available for EuroParl which can be exploited to bootstrap a discourse multiword candidate list. The corpus study is currently underway at the University of Konstanz.

Also underway is an experiment regarding ambiguous multiwords that feature multiword and non-multiword readings. First results indicate that for some of the multiwords, context window parts of speech does help in disambiguation.

Finally, as a direct outcome of the discussions with ParGram members and the host, the first German discourse multiwords have been successfully implemented in the large-scale German ParGram grammar using the XLE parsing environment.

4 Future collaboration with host institution

I plan to communicate the outcome of the current corpus studies with the STSM host. This is of high interest also to the host who has a growing interest in

[2]http://pargram.b.uib.no/

[3]http://www.statmt.org/europarl

information structure as seen from an HPSG perspective. The host is planning to visit the University of Konstanz this summer.

5 Foreseen publications/articles resulting or to result from the STSM

We are planning to publish the outcome of the studies currently underway in a conference paper soon.

6 Confirmation by the host institution of the successful execution of the STSM

Dan Flickinger, CSLI, Stanford University: I had a series of productive meetings with Sebastian Sulger during the week of his visit to CSLI in February, during which we exchanged summaries of our recent work on topics related to multi-word expressions, and discussed methods for addressing some of the known challenges for his project's focus on so-called discourse multi-words. We examined in some detail the implementation of such entries in the English Resource Grammar developed at CSLI, and looked into the various mechanisms the implementation provides for fixed vs. variable expressions, in particular for the range of uses of these entries in the political discourse text corpora relevant to their research work. The opportunity to work together on analysis of the issues and on candidate approaches, over the course of several days, was valuable, and opened several promising opportunities for more effective collaboration between our two groups.