30 March 2015

STSM – research report by Noam Ordan

The first phase of the project was, as planned, defining a set of grammar rules to identify multiword expressions in Hebrew. For this purpose a large corpus of Hebrew was crawled from the Web (10^9 tokens) and annotated for Hebrew morphology. The rules defined are in a form of regular expressions over part-of-speech tags, punctuation marks and words, such that the first word is the node and the second word encompasses its potential collocates. For example, to identify the nouns transitive verbs take the following rule is devised:

```
W1:[tag="verb"][tag="adverb"]{0,2}[word="at"][tag="numeral"]{0,2}
[tag="adverb"|word=","]{0,2} W2: [tag="noun"][tag!="noun"]
```

The first expression – indicated by W1 – searches for any verb in the corpus followed by up to two adverbs, followed by the Hebrew accusative particle “at”, followed by up to two numerals, and allowing again for up to two adverbs or a comma. The second word – indicated by W2 – simply defines a sequence of nouns. Other rules were devised to catch candidate synonyms for nouns, verbs and adjectives, subjects of passive verbs, construct- and absolute-state nouns, and predicative and attribute modifiers.

The second phase of the project involved a preliminary exploration of the results. The rules were rather loose in the sense that looking at the results showed that some of the collocations did not meet the requirements specified by the rules. However, we calculated a pointwise mutual information (PMI) for each word pair and this association measure filtered (1) pairs that do meet the rule specifications (with no evaluation as yet), and (2) oftentimes non-compositional collocations featuring up the list. The policy then was to prefer recall over precision under the assumption that the PMI association measure increases precision. In cases of light verbs, for example, high PMI scores indicate high level of idomaticty, such as “laqax” (“take”) whose high-scored nouns include the following expressions: “take the risk”, “take responsibility”, “take the time”, “take control” and then further down the list compositional phrases such as “take the medication”, “take the money” and “take the kids”.

The collaboration with the host is now further enabled. As a corollary of this research, I will now spend in Haifa another six months with further permission of my employers at Saarland University. Since the host’s research team focuses also on the study of translated language, we are going to apply these rules and methodology to study the distribution of multiword expressions across translated and non-translated language. In parallel, we are going to pursue this study in German; this will be done as part of a bachelor thesis in computational linguistics (by Tonio Süßdorf) supervised jointly by Prof. Josef van Genabith and myself at Saarland University.

I expect this study to yield publications in the next months; I will inform you of any progress as well acknowledge your support for relevant publications.