Ressource-Free Corpus Construction and Partial Annotation of light verb constructions: Preliminary results on Persian journalistic corpora

Kim Gerdes & Pollet Samvelian
Université Paris 3 – LPP UMR 7018 – Signes (INRIA)
kim.gerdes AT univ-paris3.fr
Université Paris 3 – Mondes iranien et indien UMR 7528
pollet.samvelian AT univ-paris3.fr

Persian uses an abundance of light verb constructions and verbal collocations as the morphological formation of new simple verbs, though available, is no longer productive. Constructions where the light verb takes the predicative noun as a direct object (Oper1 in Mel'cukian terms) are particularly productive. Moreover, a quite large set of simple verbs can enter into these constructions. Some nouns have competing Oper1 verbs, with slight meaning variations. The usage seems also to be genre dependent.

(1) hame kaf zadand
everybody palm beat.pas.3.pl
'Everbody applauded.'

(2) motakkâ-yaš râ andâze zad (šenâs, 80)
mattress-poss.3.sgobj-marker measure beat.pas.3.sg
'She/he measured her/his mattress.'

(3) be u bohtân zadand (vaq-vaq sâhâb, 118)
to she/he slander beat.pas.3.pl
'They slandered her/him.'

It should be noted that although some of these sequences are lexicalized, many other sequences can be freely produced, with a more or less transparent meaning predictable from the context and the verbal construction. This fact constitutes the major problem for listing Persian verbal constructions in the lexicon and indeed Persian dictionaries display great variation with respect to the number of verbal collocations they contain. Another problem is the acquisition and the interpretation of such sequences by Persian native and non-native speakers. Our long term lexicological goal is to obtain reasonably complete lists of verbal lexical functions for a large set of nouns. An example of such a database containing lexical functions is the DiCo database for the DEC dictionary.

As for all corpus work on multi-word expressions, corpus based lexicon construction and quantitative measures on light verb constructions are hampered by the lack of sufficiently large, and sufficiently annotated corpora, and the situation is particularly severe for Persian as virtually no free resources of considerable size are available.

Our approach consisted of the following steps:

We developed a specialized web crawler for linguistic corpus research, named "GrosMoteur", which allows for

- Crawling a specific sight (or crawling freely with language testing)
- Cleaning and normalizing the text (different file encodings, html special characters etc. are standardized into a MySQL data base in unicode format)
- Cutting up the text into sentences.
- Easy Configuration within a graphical user interface.

We choose the Tehran based Kayhan newspaper for its quite well proofread texts. The current work is based on a corpus of 17 million words (>700000 different sentences), but the described procedures allow for effortless scaling.

The GrosMoteur also contains an automatic morphology learning algorithm that can compute complete morphological paradigms based on a list of affixes (see Goldsmith 2001 for comparable approaches). With a handwritten file containing the Persian conjugation[1] it can complete the paradigms for those verbs it finds a reasonable amount of attested forms for. The resulting morphological full form lexicon has 120,000 forms for about 2000 verbal lemmas.

With an additional list with the hand annotation of the 100 most frequent words of our corpus, we can obtain an (often erroneous) morphosyntactic annotation of about 70% of the words. The resulting partly annotated text data base makes it possible to extract sentences containing any form of a given verb (and, if desired, certain nouns at specific distances from the verb). It is for example possible to quantify the distribution of nouns preceding directly all forms of a verb like *zadan* "beat".

In spite of the error-prone crude approach, our preliminary results are of linguistic interest as no data on Persian light verb construction exist. We stumbled on constructions, not described in dictionaries and difficult to come up without a corpus, showing the astonishing freedom and inventiveness of light verb constructions in Persian.

Our short term goal is to add simple constituent and dependency structures to our annotation, again based on a semi-automatic learning system, to allow for more precise results and searches of long distance relationships. From a view point of Natural Language Processing, too, this work stands out as working on multi-word constructions of a non European language.

References:
Goldsmith, J. 2001. Unsupervised learning of. the morphology of a natural language. *Computational Linguistics* 27(2): 153-198.
Karimi-Doostan, M. 1997. *Light Verb Constructions in Persian*, PhD Thesis, University of Essex.
Mel'čuk I., Arbatchewsky-Jumarie N., Iordanskaja L., Mantha S., Polguère A. (1984-1999) Dictionnaire explicatif et combinatoire du français contemporain. *Recherches lexico-sémantiques*, Montréal, Les Presses de l'Université de Montréal, see http://olst.ling.umontreal.ca/dicouebe/index.php for the online version.

[1] With about 60 lines like "*3sgPreSubj: b|bi RAD d*" indicating that the 3rd person singular of the present subjective is formed with either b or bi as a prefix and d as the suffix.