Annotation of Multiword Expressions in the Farsi Section of the Universal Dependencies Project

Behrang QasemiZadeh, Heinrich-Heine-Universität Düsseldorf

Background

Annotations of multiword expressions (MWEs) in the Farsi section of the universal dependencies (UD) project are reviewed and compared with the English counterparts. For this language pair, comparable multiword structures are first identified. Their annotations in each of the corpora are then retrieved, their consistencies are checked, and some conflicting examples are gathered using the INESS corpus management system. We believe these inconsistencies are due to the syntax-oriented perspective taken for the annotation of MWEs in the UD project, i.e., MWEs are annotated to fill the void often caused by the lack of a clear syntactic relationship between their elements.

Farsi (Persian Language)

- Farsi, also known as Persian, is the official language of Iran and is spoken by about 100 million people.
- It belongs to the family of *Indo-European languages* and has a straightforward morphology and syntax:
  - Derivation and inflection are carried out by affixation;
  - Case markers are rarely used and the word order is not restricted (although the SOV pattern is dominant).
- Complications in the formal analysis of Farsi are often caused by the fact that it is written using the Arabic transliteration system:
  - This transliteration system hinders identifying the boundaries of words;
  - Put simply, white spaces may not represent the boundaries of words.
- Another peculiar characteristic of Farsi, which can be potentially interesting for the study of MWEs, is the way that actions and events are described:
  - The number of simple verbs in Farsi is extremely limited—there are less than 800 of which less than 300 are commonly used.
  - Actions and events are described using compound verbs. Most works (including the Farsi section of the UD project) traditionally limit the syntactic structure of verb compounds to the combination of a light verb and a non-verbial part (e.g., noun, adjective, prepositional phrase, etc.).

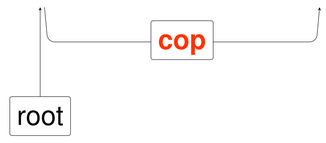
Farsi and English MWEs in UD

We closely examined three different syntactic relationships: `compound`, `name`, and `mwe`.

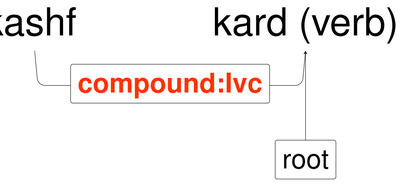
1. compound

- The language specific relation `compound:lvc` for annotating a light verb construct is introduced. In a number of cases, while the active voice of compound verbs is marked using `compound:lvc`, their passive voice is not marked in the same way:

Passive Voice: kashf shod (verb)

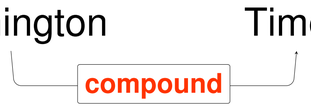


Active Voice: kashf kard (verb)



- In the English corpus, `compound` relation is asserted between elements of multi-word named entities and terms. However, in Farsi `nmod:poss` replaces `compound`:

English: Washington Times



Farsi: khabargozari (news agency) Fars

English: Higher Education

Farsi: amoozesh (education) aali (higher)

2. name

`name` is used for both Farsi and English in the same way. However, English honorifics are connected to their regents using `compound`; in Farsi, honorifics are connected to their regents using the `name` relation:

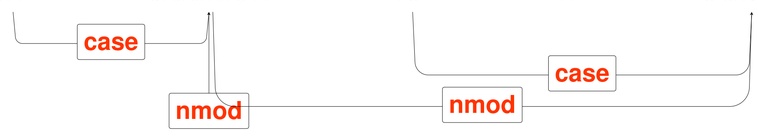
English: Mrs. Arafat

Farsi: khanom-e (Mrs.) Arafat

2. mwe

Similar to the English section, for Farsi `mwe` is mostly employed to mark prepositional compounds. Particularly those that end in the conjunction */ke/* (i.e., comparable role to 'that' in English). The use of `mwe`, however, is somehow confusing. Annotating every sequence of prepositions that ends with the conjunction /ke/, highly productivity structures, is an overuse of the relation `mwe` and inconsistent with the English annotations (in which 'that' is related using the relation case). Constructs other than the above-mentioned are also common, for instance:

English: In addition to ... outlets



Farsi: alaveh bar aan ...