Homogeneous annotation of dependency relations using universal dependencies (UD): The case of P-drop in Persian

Fahime Same
Universität zu Köln
f.same@uni-koeln.de

UD has been proven to be useful for the analysis of typologically different languages. In this abstract, the language-internal implications of UD is discussed. The case study presented here is limited to preposition-drop in Persian across formal and informal registers, but UD strategy has a much wider implication across different languages. UD headedness rules make it possible to provide a unified account for language-internal variations, for instance the structural variation in any language where function words could be dropped optionally.

Dependency grammar proves to be able to handle syntactic structures of languages with free word order better than phrase-structure grammar. In dependency grammar, the syntactic structure is shown "in terms of the words (or lemmas) in a sentence and an associated set of directed binary grammatical relations that hold among the words" (Jurafsky & Martin, 2017, p.1). Unlike most dependency-based parsing schemes, UD defines sets of dependency relations between the content words. Function words are attached to the content words as their direct dependents. Having lexical heads as the backbone of the syntactic analysis maximizes the parallelism across different languages and facilitates comparative cross-linguistic studies (Nivre et al., 2016).

P(reposition)-drop - the omission of preposition in prepositional phrases (PP) - is a linguistic phenomenon which often occurs in spoken Persian, where mono-morphemic spatial prepositions (be "to", dær "in", ru "on", tu "at") can be dropped (Pantcheva, 2008). A rough analysis of 400 post verbal PPs with the form [V_{mov} (to) PLACE] from the multilingual corpus sgs (Adli, 2011) shows that in over 80% of the cases, the preposition is dropped, while it is often present in similar sentences in the written register. Examples (1) and (2) show the presence and the absence of the preposition respectively.

(1). sgs, interview 38, line 437[1]
[…] raft-an **be** 'otAq-A-je xod-eSun.
went$_{3pl}$ **to** rooms themselves.
They went to their rooms

(2). sgs, interview 63, line 281
raft esfahAn.
went$_{3sg}$ Esfahan.
He went to Esfahan.

Following a dependency analysis in which function words are the head of a dependency relation, the result for the first example is shown in figure 1[2]. Applying the same analysis to the second example would be problematic, because the preposition which is the functional head is absent. The first strategy to adopt would be to insert an empty head node and link it to the root [figure 2]. The second strategy is to link the root to the content word [figure 3].

Following the first strategy is not efficient. It leads to less consistency and extra work for manual annotation or correction. The second approach results in incoherence in the analysis of the PPs. When the preposition exists, there is a dependency relation between the functional head and the root. In its absence, the relation is between the nominal element and the root.

[1] Persian examples are transliterated according to Adli (2011).
[2] In UD, the tag *nmod* shows the link between the noun phrase of a PP and what it modifies.

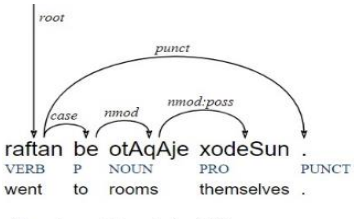

figure 1. non-UD analysis of (1)

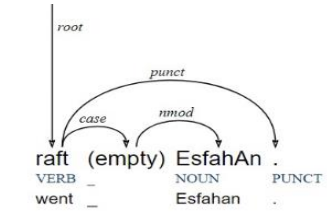

figure 2. First strategy for the analysis of (2)

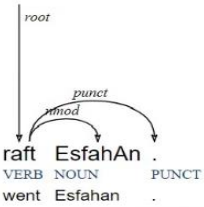

figure 3. Second strategy for the analysis of (2)

Since in UD content words are heads, the dependency relation analysis remains the same for both examples. Following UD leads to a homogeneous analysis of PPs headed by the preposition and the ones without a preposition.

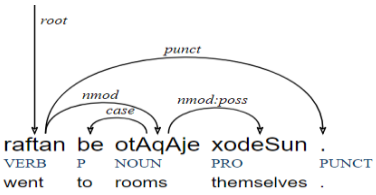

figure 4. UD analysis of (1)

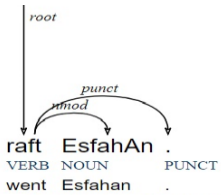

figure 5. UD analysis of (2)

To conclude, this approach minimizes the annotation effort and maximizes the homogeneity of dependency relations in case of language-internal form variations. As NLP relies heavily on linguistic annotations (Nivre, 2017), less variation leads to more accurate results. Additionally, the example presented here showed the variation across two registers. In Persian, gold standard syntactic annotations are available mainly for formal register (e.g. newspaper articles). This favors performance on formal texts, and the performance of parsers on other genres suffers due to this bias (Silveira et al., 2016). Having the same type of analysis for a phenomenon which shows register differences allows us to use the same parsers for both formal and informal registers without a dramatic drop in accuracy.

References:

Adli, A. (2011). Gradient acceptability and frequency effects in information structure: a quantitative study on Spanish, Catalan, and Persian. *Freiburg: Universität Freiburg dissertation*.

Jurafsky, D., & Martin, J. H. (2017). Speech and language processing. Draft of August 28, 2017.

Nivre, J. (2017). Perspectives on universal dependencies [PowerPoint slides]. Retrieved from https://stp.lingfil.uu.se/~nivre/docs/NivreRANLP17.pdf

Nivre, J., de Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... & Tsarfaty, R. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *LREC*.

Seraji, M., Ginter, F., & Nivre, J. (2016). Universal Dependencies for Persian. In *LREC*.

Pantcheva, M. (2008). The place of PLACE in Persian. *Syntax and semantics of spatial P*, *120*, 305.

Silveira, N., Dozat, T., de Marneffe, M. C., Bowman, S. R., Connor, M., Bauer, J., & Manning, C. D. (2014). A Gold Standard Dependency Corpus for English. In *LREC* (pp. 2897-2904).