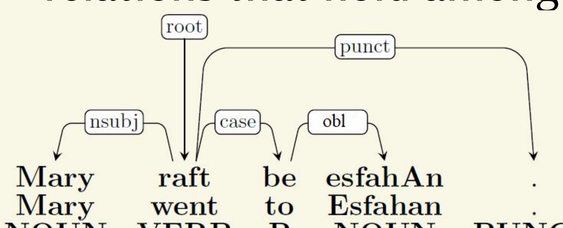
Homogeneous annotation of dependency relations using universal dependencies (UD): The case of P-drop in Persian

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

Background: dependency-based grammar

- In dependency-based grammar, syntactic structure is analyzed in terms of the words (or lemmas) in a sentence and an associated set of directed binary grammatical relations that hold among them [1].



Background: UD

- Dependency relations between the content words
- Function words are attached to the content words as their direct dependents.
- Useful for the analysis of typologically different languages
- Universal taxonomy with language specific elaboration
- UD maximizes parallelism across different languages
- UD facilitates comparative cross-linguistic studies [2], [3].

Case study: Preposition-drop in Persian across formal and informal registers

Purpose: Highlighting language-internal implications of UD

Language-internal implications: headedness rules in UD facilitate the study of language-internal variations, e.g. structural variation in any language where function words could be dropped optionally.

- P-drop - the omission of preposition in prepositional phrases (PP) - mostly occurs in informal Persian
- Mono-morphemic spatial prepositions (*be* "to", *dær* "in", *ru* "on", *tu* "at") can be optionally dropped in colloquial speech in [V_{mov} (to) PLACE] constructions [4].

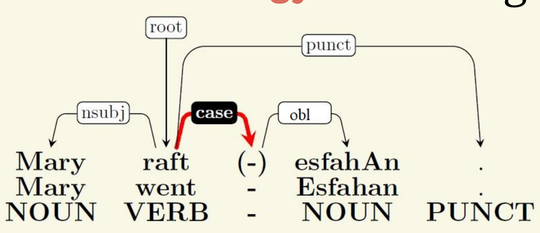
Examples: PPs with/without overt prepositions

(1) Mary raft **be** esfahAn.
Mary went to Esfahan.

(2) Mary raft esfahAn.
Mary went Esfahan.

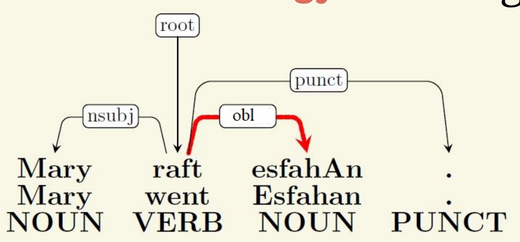
Non-UD-based analysis of P-drop

First strategy: Inserting an empty head node



Disadvantages: Not efficient: extra work for manual annotation or correction | Inconsistency in the analysis

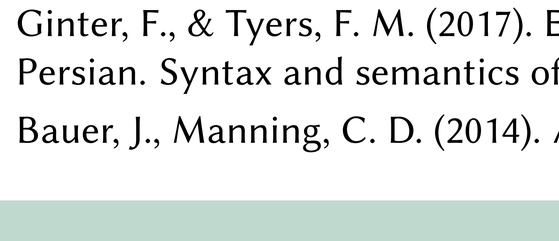
Second strategy: Linking the root to the content word



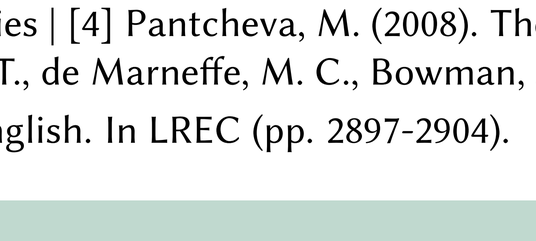
Disadvantages: Incoherence in the analysis of the PPs

UD-based analysis of P-drop

Content words function as heads
→ *similar solution for handling variation.*



Advantage: A homogeneous analysis of headed and headless PPs.



UD and language-internal variations

General implications:

- Minimizing the annotation effort
- Mximizing the homogeneity of dependency relations
- As NLP relies heavily on linguistic annotations [2], less variation leads to more accurate results.

Language-specific implications:

- In Persian, gold standard syntactic annotations are available mainly for formal register.
  - This favors performance on formal genre [5].
  - The performance of parsers on other genres suffers due to this bias.
  - UD analysis allows for using the same parsers for texts in both registers without a dramatic drop in accuracy.

References:

[1] Jurafsky, D., & Martin, J. H. (2017). Speech and language processing. | [2] 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.| [3] Nivre, J., Zeman, D., Ginter, F., & Tyers, F. M. (2017). EACL 2017 Tutorial on Universal Dependencies | [4] Pantcheva, M. (2008). The place of PLACE in Persian. Syntax and semantics of spatial P, 120, 305. | [5] 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).

Universität zu Köln
SFB 1252 PROMINENCE IN LANGUAGE