# Information Structure with the Prague Arabic Dependency Treebank

Otakar Smrž, Petr Zemánek, Jakub Kráčmar, Viktor Bielický

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague
`padt@ufal.mff.cuni.cz`

## Abstract

The issue of information structure in language has been studied extensively both in the Prague School of Linguistics and in the Functional Generative Description [FGD, 11, 5]. This theory of representation of linguistic meaning is the framework for a family of multi-level annotation projects, in particular the Prague Dependency Treebank for Czech [PDT, 2, 3] and the Prague Arabic Dependency Treebank [PADT, 4, 12].

Information structure — the question of 'the given' and 'the new' in an utterance and how it is expressed — is considered to contribute to the linguistic meaning, and its annotation in PDT is part of the third, the most detailed and abstract level of linguistic description. Next to determining which elements in a sentence are context-bound and which are non-bound (the elementary distinctive feature from which the topic–focus dichotomy is derived), attention is also paid to the resolution of anaphoric relations and to capturing the communicative dynamism of a proposition (annotations of coreference and deep word order, respectively) [9].

In PADT, which now consists of the morphological and the analytical levels of description of Arabic, a similar annotation of information structure is being established. In our contribution, we would like to overview the theoretical concepts we work with, and present our formal treatment of a number of prototypical, yet corpus-based, instances of linguistic phenomena that have a principal impact on the structure of information in Arabic [6, 1].

The applicability of the general approach to written as well as spoken Arabic will be the main point of our account. In FGD, the description

of information structure incorporates also the notions of intonation center and stress, contrast, subjective word order, or potential ellipsis, which are directly connected to prosody. We will provide references to other related computational research, too [cf. e.g. 8, 7, 10].

# References

[1] Kristen E. Brustad. *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Georgetown University Press, 2000.

[2] Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká. Prague Dependency Treebank 1.0. LDC catalog number LDC2001T10, ISBN 1-58563-212-0, 2001.

[3] Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká. Prague Dependency Treebank 2.0. LDC catalog number LDC2006T??, ISBN 1-58563-???-?, 2006.

[4] Jan Hajič, Otakar Smrž, Petr Zemánek, Petr Pajas, Jan Šnaidauf, Emanuel Beška, Jakub Kráčmar, and Kamila Hassanová. Prague Arabic Dependency Treebank 1.0. LDC catalog number LDC2004T23, ISBN 1-58563-319-4, 2004.

[5] Eva Hajičová and Petr Sgall. Dependency Syntax in Functional Generative Description. In *Dependenz und Valenz – Dependency and Valency*, volume I, pages 570–592. Walter de Gruyter, 2003.

[6] Clive Holes. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press, 2004.

[7] Geert-Jan M. Kruijff. *A Categorial-Modal Architecture of Informativity: Dependency Grammar Logic & Information Structure*. PhD thesis, Charles University in Prague, 2001.

[8] Ivana Kruijff-Korbayová. *The Dynamic Potential of Topic and Focus: A Praguian Approach to Discourse Representation Theory*. PhD thesis, Charles University in Prague, 1998.

[9] Marie Mikulová et al. A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank. Technical report, Charles University in Prague, 2006.

[10] Oana Postolache. Learning Information Structure in the Prague Treebank. In *Proceedings of the ACL Student Research Workshop*, pages 115–120, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

[11] Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel & Academia, 1986.

[12] Otakar Smrž, Petr Pajas, Zdeněk Žabokrtský, Jan Hajič, Jiří Mírovský, and Petr Němec. Learning to Use the Prague Arabic Dependency Treebank. In *Perspectives on Arabic Linguistics*, volume XIX. John Benjamins, 2006.