FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Mgr. František Orangutan, Ph.D.
research associate
ÚFAL MFF UK
Malostranské náměstí 25
118 00 Praha
Czechia

November 21, 2021

Kancelář prezidenta republiky
Hrad I. nádvoří č. p. 1
Hradčany
119 08 Praha 1
Česká republika

An update on the Prague Dependency Treebank

I am writing to present our most recent findings within the project of the Prague Dependency Treebank (PDT).

The Prague Dependency Treebank 2.0 (PDT 2.0) contains a large amount of Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation (0.8 MW); in addition, certain properties of sentence information structure and coreference relations are annotated at the semantic level.

PDT 2.0 is based on the long-standing Praguian linguistic tradition, adapted for the current Computational Linguistics research needs. The corpus itself uses the latest annotation technology. Software tools for corpus search, annotation and language analysis are included. Extensive documentation (in English) is provided as well.

This version differs from the CD-ROM version in minor text corrections of the guide.

Please note that new versions of this corpus have been published: PDT 3.0 (2013), PDiT 1.0 (2012), PDT 2.5 (2012).

The Prague Dependency Treebank (PDT) is an open-ended project for manual annotation of substantial amount of Czech-language data with linguistically rich information ranging from morphology through syntax and semantics/pragmatics and beyond.

PDT version 2.0 is a sequel to version 1.0; PDT version 1.0 contains manual annotation of morphology and (surface) syntax. Version 2.0 adds the underlying syntax and semantics, topic/focus, coreference and lexical semantics based on a valency dictionary to the surface syntax and morphology that have been at the core of version 1.0. The corrections of version 1.0 are also included in version 2.0, even with the old data format preserved for those

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

who have already invested into its use.

The annotation in PDT 2.0 covers a large amount of Czech texts with interlinked morphological (2 million words), syntactic (1.5 MW) and complex underlying syntactic and semantic annotation (0.8 MW). The corpus itself now uses the latest annotation technology (standoff annotation using XML, RelaxNG-see Section 3.4, "Data formats" and the whole Chapter 3, Data).

PDT 2.0 is based on the long-standing Praguian linguistic tradition and adapted for the current Computational Linguistics research needs (see also Section 1.2, "Historical background of the project"). Software tools for corpus search, annotation and language analysis are included. Extensive documentation (in Czech and English) is provided as well.

Yours faithfully,

Franta Vopička