The Prague Dependency Treebank of Spoken English is a collection of English spoken dialogs about personal photograph collections. In this work we enrich it with additional linguistic information targeted for information-extraction tasks. The original corpus consisted of three interlinked representations (audio, manual transcription and a morphological layer) annotated with formats based on the Prague Markup Language (PML) which is a backbone for the family of XML schemas for rich linguistic annotations of texts, such as morphological tagging and dependency trees.

We converted the morphological layer of the corpus into a treebank in the standard Penn Treebank bracketing style and enhanced it with part of speech tags, named entity labels, WordNet hypernyms and links to the lower layers of annotation. The pre-processed corpus data was given as an input into state-of-the-art NLP tools such as the Stanford parser and named entity recognizer, and the WordNet API to obtain the additional analyses. These annotations were added in such a way as to preserve the original PML format. We designed new XML schemas for the modified topmost layer of the corpus so that it could be appropriately displayed in editors for linguistic corpus processing, in particular, the powerful toolkit TrEd, which is a programmable graphical tree editor and browser for PML-compliant corpuses.

We can now formulate such queries as, for example, "find the tokens that are nouns and have a hypernym "anniversary".

The augmented corpus contains interesting strata of linguistic knowledge, is compatible with a specialized open-source query engine and is suitable for extensive information extraction.

Keywords: rich linguistic annotations, multi-layer corpus, PDTSE