Anotace mezivýpovědních textových vztahů na ÚFAL MFF UK

Zuzanna Bedřichová, Lucie Mladová

Pražský závislostní korpus (PDT 2.0), vytvořený a dále rozpracovávaný v Ústavu formální a matematické lingvistiky MFF UK, obsahuje česká jazyková data z oblasti žurnalistiky značkovaná na základě uceleného formálního popisu jazyka. Značkování korpusu PDT vychází z teorie funkčního generativního popisu, jehož základními hledisky jsou pojmy syntaktické závislosti, (zejména slovesné) valence a stratifikačního popisu jazyka – v PDT existují tři vrstvy lingvistického popisu: anotace morfologická, analytická (povrchově syntaktická) a tektogramatická (zachycující podkladovou syntax, sémantiku, aktuální členění a některé vztahy gramatické a textové koreference). Kromě těchto anotací vydaných ve verzi 2.0 probíhají některé další specializované anotační projekty, které se připravují pro další vydání PDT, např. anotace rozšířené textové koreference či zachycení víceslovných lexémů a pojmenovaných entit.

Dalším z projektů, které PDT 2.0 obohatí o novou vrstvu lingvistického popisu, je projekt anotace mezivýpovědních textových vztahů (dále MTV), na kterém se pod vedením prof. Evy Hajičové podílejí Šárka Zikánová, Lucie Mladová a Zuzanna Bedřichová. Cílem tohoto projektu je zachytit sémantickou výstavbu textu (ve světovém lingvistickém povědomí „discourse structure“), zejména pak sémantické vztahy mezi jednotlivými výpověďmi.

Anotace sama vychází z tektogramatického zápisu jazyka vyvinutého v předchozích fázích práce na korpusu, v němž každá jednotlivá věta (úsek textu mezi dvěma koncovými interpunkčními znaménky) zobrazena jako závislostní stromová struktura. Jedním závislostním stromem tedy v tektogramatickém zápise je i souřadné souvětí, které zahrnuje více výpovědí, nebo naopak syntaktický celek, který odpovídá pouze části výpovědi (například v případě parcelace). Proto nelze říci, že mezivýpovědní vztahy odpovídají vždy vztahům mezi jednotlivými stromy, ačkoli jednou ze základních idejí anotace MTV je systematicky spojit především jednotlivé tektogramatické stromy.

V aktuální první fázi se projekt zaměřuje na zachycení MTV vyjádřených tzv. textovým konektorem. Při vytváření klasifikačního systému MTV pak autoři koncepce anotace MTV do určité míry vycházeli z anotačních zásad filadelfského projektu Penn Discourse TreeBank a z jeho srovnání s klasifikačními možnostmi, které nabízí již existující systém sémantických vztahů tektogramatické roviny. Protože anotace pracuje s členěním textu na tektogramatické

stromy, jsou vztahy mezi výpověďmi nutně různého typu: zahrnují jak vztahy odpovídající řadě (nejen) koordinačních vztahů obdobných větné syntaxi (konjunkce, opozice, gradace, podmínka, přípustka apod.), tak vztahy reprezentující významové rozvinutí určitého tvrzení (specifikace, ekvivalence, rektifikace, restrikce apod.). Při koncipování anotačního schématu tak vyvstávaly různé otázky týkající se tradičního pojetí syntaxe, na které by práce s materiálem mohla přinést odpověď, stejně jako i řadu otázek dalších (např. zda existuje parataktický případ vyjádření účelu či jaký je vztah čistě kauzálních souvislostí mezi výpověďmi oproti kauzalitě tzv. nepravé, tj. mezi obsahy nevyjádřenými, inferovanými apod.).

Oproti filadelfskému projektu však přináší ten pražský několik nových hledisek, a to především přímé propojení anotace MTV s tektogramatickou rovinou a možnost využít již existující anotace aktuálního větného členění (např. některé kontrastní vztahy tak již není nutné znovu anotovat) či anotace rozšířené textové koreference. Výsledná komplexní anotace tak může přinést hlubší poznání struktury textu a být využita nejen jako materiál pro další počítačové využití v oblasti informační struktury (např. pro automatickou sumarizaci, strojový překlad apod.), ale také jako materiál pro řadu lingvistických bádání v oblasti nadvětné syntaxe, rétorické výstavby textu, textových strategií apod.

Tento výzkum je podporován následujícími grantovými projekty: GA ČR 405/09/0729 „Od struktury věty k textovým vztahům“, GA UK 102409 „Sémantická funkce tzv. synsémantik ve výstavbě jazykového projevu“ a GA UK 103609 „Textové (mezivětné) vztahy a jejich zachycení v jazykovém korpusu“.