Název práce: Automatické párování tektogramatických stromů z česko-anglického paralelního korpusu
Autor: David Mareček
Katedra (ústav): Ústav formální a aplikované lingvistiky
Vedoucí diplomové práce: Ing. Zdeněk Žabokrtský, Ph.D.

Abstrakt: Cílem této práce je implementovat a zhodnotit softwarový nástroj pro automatické zarovnávání (alignment) českých a anglických tektogramatických stromů. Úkolem je najít odpovídající si uzly stromů, které reprezentují anglickou větu a její český překlad. Velké množství zarovnaných stromů získaných z paralelního korpusu může být užitečné pro trénování modelu pro transfer strojového překladu. Zároveň může posloužit lingvistům při studování překladových ekvivalentů mezi dvěma jazyky. Výsledky našich experimentů ukazují, že přesunutím problému alignmentu ze slovní roviny na tektogramatickou (a) zvýšíme mezianotátorskou shodu (b) můžeme vytvořit alignovací algoritmus, který využívá i stromovou strukturu věty a překoná nástroj pro alignment GIZA++ spuštěný na uzly tektogramatických stromů. To je pravděpodobně zapříčiněno tím, že tektogramatické reprezentace českých a anglických vět si jsou mnohem podobnější než samotné věty na povrchu.

Klíčová slova: tektogramatická rovina, word alignment, strojový překlad