Závislostní syntaktická analýza věty je jedním za základních výzkumných témat v oblasti zpracování přirozeného jazyka. Výzkum algoritmů pro závislostní analýzu byl dosud zaměřen především na omezené množství jazyků, pro které existuje dostatek trénovacích dat (převážně jde o evropské jazyky). Z dosavadního výzkumu tedy není snadné vyvodit, který algoritmus bude nejvhodnější pro některý další jazyk.

Řešení, které navrhujeme, je pracovat místo jednotlivého analyzátoru se souborem několika analyzátorů. Konkrétně zkoumáme tři metody. Zaprvé, sloučíme výstupy jednotlivých analyzátorů do grafu a výsledný závislostní strom sestavíme prostřednictvím hlasování analyzátorů na jednotlivých uzlech. Zadruhé, váhy jednotlivých závislostních hran v grafu určíme na základě seskupení analyzátorů pomocí metody fuzzy clustering, jejímž vstupem je zde rozložení chyb jednotlivých analyzátorů napříč slovními druhy. Zatřetí, implementujeme meta-klasifikátor, který pro každé slovo v dané větě určí nejdůvěryhodnější analyzátor. Závislostní analyzátory, které používáme, jsou založeny na několika odlišných technikách -. jde o grafové analyzátory, přechodové analyzátory a analyzátory spočívající v konverzi ze složkových stromů. Právě tato škála různých analyzátorů nám umožňuje studovat chyby spojené s jednotlivými technikami a vybrat nejvhodnější kombinaci analyzátorů pro danou situaci.

Závislostní analyzátory jsou typicky vyhodnocovány pouze na základě počtu správně zavěšených uzlů, případně s přihlédnutím ke správnosti přiřazené syntaktické funkce. Pokud je pro některý z jazyků s dostatkem trénovacích dat a vysokým dosaženým skóre vyvinuta nová technika, její výsledky jsou ve srovnání se stávajícími technikami často stejné nebo horší. Posun v úspěšnosti se pak typicky omezuje jen na úzce vyměřenou jazykovou konstrukci nebo textovou doménu. Taková zlepšení pak ale nelze přenést na jazyky s malým množstvím dat.

Abychom se ujistili, že náš přístup přenositelný je, testujeme jej na velkých i malých datech pro několik jazyků z různých jazykových rodin. Zvláštní důraz klademe na jazyky s malým množstvím zdrojů, u kterých navíc experimentujeme s metodami částečně řízeného strojového učení (self-training). Pro jazyky s malým množstvím zdrojů může tento přístup výrazně přispět ke kvalitě analyzátorů i k efektivnější přípravě dalších trénovacích dat. V případě, že pracujeme se soubory analyzátorů, navíc vyvstává možnost přetrénovat jednotlivé analyzátory s pomocí výstupů zkombinovaných ze všech analyzátorů.

Ať už jde o jazyky s malým nebo velkým množstvím datových zdrojů, domníváme se, že počet správně zavěšených hran by neměl být jediným ukazatelem určujícím užitečnost jednotlivých analyzátorů. Pro změření reálného významu jednotlivých zlepšení jsme si jako konkrétní aplikaci vybrali strojový překlad.

Existuje řada přístupů ke strojovému překladu, založených na lingvistických pravidlech i na statistickém modelování. Pokud se daný systém strojového překladu skládá z posloupnosti jednotlivých komponent, vzniká otázka, jak vážně poškodí chyby vzniklé už na začátku posloupnosti výslednou kvalitu překladu. Například dvouprocentní chyba v určování slovních druhů může vést k podstatně větší chybě v následné syntaktické analýze a vliv na počet chyb v konečném překladu může být kvůli šíření chyby ještě větší. Snižování chybovosti komponent umístěných na začátku procesu je tedy podstatné, a to i proto, aby se badatelé mohli soustředit na samotné jádro překladu a ne na potlačování důsledků chyb vzniklých v analýze.

Abychom mohli experimentálně vyhodnotit vliv různých syntaktických analyzátorů na kvalitu překladu, používáme překladový systém TectoMT implementovaný v prostředí Treex. Na rozdíl od dnes dominujících systému frázového překladu je systém TectoMT založen na překladu přes závislostní strukturu nazývanou tektogramatický strom. V tomto experimentálním prostředí tedy můžeme zkoumat vztah mezi množstvím správně určených závislostních hran a správností překladu aproximovanou pomocí standardních měr BLEU a NIST.

Překážkou pro vyhodnocení těchto experimentů byla skutečnost, že pro angličtinu neexistovala žádná ručně anotovaná závislostní data. Standardním řešením bylo použít pro testování i trénování složkové stromy automaticky zkonvertované do závislostí, nicméně tento přístup vede k systematické chybě ve vyhodnocení. Z tohoto důvodu jsme ručně oanotovali závislostní stromy pro část anglických textů z překladové soutěže WMT 2012 a pro část složkového korpusu Penn Treebank.

V této disertaci ukazujeme možná zlepšení jak samotné závislostní syntaktické analýzy, tak strojového překladu, který z takové analýzy vět vychází. Vedle toho jsme se pokusili vyrobit ručně anotovaná data, která mohou v budoucnu posloužit jako zlatý standard pro další experimenty.