Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Lukáš Kyjánek
Název práce	Harmonisation of Language Resources for Word-Formation of Multiple Languages
Rok odevzdání	2020
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	Mgr. Magda Ševčíková, Ph.D. **Role** Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky, MFF UK

Text posudku:

Cílem diplomové práce bylo vytvořit sadu různojazyčných dat, v nichž budou slovotvorné vztahy (vztah mezi slovem odvozeným a slovem základovým/slovy základovými) a případně další slovotvorné rysy reprezentovány jednotným způsobem. Dostupné zdroje byly v diplomové práci zdokumentovány a některé z nich harmonizovány do jednotného formátu.

Po vymezení předmětu slovotvorného bádání v první kapitole, kde diplomant s pochopením kombinuje českou popisnou tradici s obecným přístupem typologických studií, jsou ve druhé kapitole popsány existující jazykové zdroje, z nichž lze čerpat informace o slovotvorné struktuře jednotlivých jazyků. Kromě dat, která se na slovotvorbu zaměřují výhradně nebo primárně, jsou uvedeny i zdroje, v nichž jsou slovotvorné rysy navršeny na anotaci jiného typu (wordnety, treebanky).

Z téměř padesáti katalogizovaných zdrojů je ve třetí kapitole, která je vlastním jádrem diplomové práce, k harmonizaci vybráno 27 zdrojů pokrývajících 20 jazyků. Jako cílová reprezentace jsou zvoleny stromové grafy používané v české derivační síti DeriNet a několika dalších zdrojích. Ostatní zdroje jsou na tuto reprezentaci převedeny poloautomatickou procedurou. Volba stromové struktury pro reprezentaci derivace a zvláště pak kompozice není neproblematická. Diplomant si je však těchto problémů vědom a např. k reprezentaci vztahů mezi kompozitem a základovými lexémy zavádí relace sekundární. V práci je také ověřena možnost převedení harmonizovaných dat zpět do původní reprezentace.

Výsledná harmonizovaná data byla zveřejněna pod názvem Universal Derivations 1.0 v repozitáři Lindat/Clariah-CZ. Zpřístupněny byly také harmonizační skripty a anotační rozhraní, které diplomant vytvořil pro ruční anotaci trénovacích dat; vše rovněž v příloze diplomové práce. I když harmonizační procedura dosáhla na jednotlivých zdrojích rozdílné úspěšnosti (důvody je třeba hledat ve strukturních rozdílech mezi jazyky, ale také v dostupnosti jazykových expertů pro daný jazyk a volbě harmonizační metody) a nechává tak prostor pro zlepšení v jednotlivých zdrojích, jedná se v mezinárodním kontextu o první kolekci harmonizovaných slovotvorných dat pro více jazyků. Kolekce je tak novým datovým zdrojem jak pro experimenty v počítačovém zpracování slovotvorby, tak i pro kontrastivní lingvistický výzkum této oblasti.

Diplomová práce je napsána srozumitelnou angličtinou. Po formální stránce má standardní strukturu, je zpracována pečlivě, a to včetně tabulek, obrázků a rozsáhlého seznamu citované literatury a jazykových zdrojů.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 16. června 2020 **Podpis**