MUNI
FAKULTA
INFORMATIKY

POSUDEK OPONENTA
DIPLOMOVÉ PRÁCE

Název práce:	*RapCor Support*
Autor práce:	*Bc. Michal Hala*
Vedoucí práce:	*doc. Mgr. Pavel Rychlý, Ph.D.*
Konzultant:	*doc. PhDr. Alena Polická, Ph.D.*
Oponent:	*RNDr. Vojtěch Kovář, Ph.D.*

Předložená diplomová práce se zabývá automatizací procesu budování korpusu francouzských rapových textů RapCor, resp. zefektivněním tohoto procesu. Po dvou krátkých teoretických kapitolách věnujících se morfologickému značkování a korpusům obecně je představen korpus RapCor a stávající proces jeho rozšiřování. Následuje těžiště práce, porovnání tří dostupných morfologických značkovačů pro francouzštinu a popis vylepšení procesu budování korpusu vedoucí k efektivnější a správnější manuální anotaci.

Text práce je psán relativně dobrou angličtinou, což hodnotím pozitivně. Rušivé jsou poměrně časté chyby ve slovosledu a nadužívání pasivních tvarů sloves, kvůli čemuž na několika místech není jasné, co je výsledkem práce studenta a co výsledkem externích okolností či práce někoho jiného. Také by bylo možné zapracovat na koherenci a čtivosti textu, na několika místech jsem musel dohledávat informace uvedené na jiném místě textu, jindy se naopak informace nadbytečně opakovaly. Ojediněle důležité informace chybí (např. kapitola 2 vzbuzuje dojem, že existují pouze pravidlové a neuronové značkovače, což není pravda). Typografická úroveň textu a práce s literaturou jsou na velmi dobré úrovni.

Praktická část práce je nepochybně užitečným, možná až stěžejním příspěvkem k budování korpusu RapCor. Z pohledu informatiky, resp. počítačové lingvistiky, jako výzkumného oboru však na mne práce působí spíše jako soubor víceméně přímočarých implementačních vylepšení již existujícího procesu, tedy spíše vývojová práce.

Přitom se nedá říci, že se při řešení žádné zajímavé otázky nevyskytly, např. kvantitativní srovnání morfologických značkovačů je jistě možné provést a řešení souvisejících problémů (rozdílná tokenizace, nekompatibilní značkové sady) jsou výzvy, které je třeba řešit, aby kvantitativní srovnání mohlo mít vypovídající hodnotu. Autor se však omezil na konstatování, že kvantitativní srovnání není možné, a nejlepší tagger byl pak vybrán de facto na základě pocitu z označkování pěti arbitrárně zvolených textů. To je podle mého názoru škoda a snižuje to hodnotu práce. Představoval bych si, že nejlepší tagger bude vybrán právě na základě kvalitně provedeného kvantitativního vyhodnocení.

Otázky k obhajobě:

- V sekci 6.1 popisujete vytvoření nové značkové sady zkrácením standardních „features” v Universal Dependencies s motivací, aby se značka vešla do jedné buňky tabulky. Je to ale správná motivace? Nejedná se o krok špatným směrem, kterým se odchýlíte od standardu Universal Dependencies, a značky budou navíc méně čitelné, a tedy i náchylnější k chybám v anotaci?

- Je opravdu nevyhnutelné (jak dvakrát v práci zmiňujete) i při korpusu této velikosti se časem uchýlit k použití databáze místo Google Spreadsheets? Kterou databázi byste použil a co přesně by to podle vás přineslo za výhody oproti současnému řešení?

Závěrem: Pokud by se jednalo o práci bakalářskou, hodnotil bych výborně nebo velmi dobře. Od výborné diplomové práce ale kromě většího rozsahu očekávám větší přesah do zajímavých problémů zpracovávaného oboru a ten jsem v této práci bohužel nenašel. Přesto si ale myslím, že práce **splňuje požadavky kladené na FI MU na diplomovou práci** a s ohledem na zmíněné připomínky navrhuji hodnotit ji známkou **C nebo D**, dle obhajoby.

Brno 13. června 2022 RNDr. Vojtěch Kovář, Ph.D.