Vojtěch Janda: ***Frekvenční distribuce nominální flexe v češtině***

Bakalářská práce. Praha: FF UK, 2017.

Posudek vedoucího.

Předkládaná práce se zabývá flexí českých substantiv. Metodologicky kombinuje postupy frekvenčního korpusového výzkumu, perspektivu funkčně-kognitivních přístupů založených na užívání i poznatky jazykové typologie. Cílem práce bylo pomocí kvantitativních metod ověřit, zda a do jaké míry lze frekvenční rozložení české pádové flexe (tj. rozdílné relativní frekvence pádových tvarů u jednotlivých substantiv) vysvětlit (aproximovat) pomocí typologické hierarchie životnosti. Práce dochází k závěru, že životnost má na frekvenční distribuci významný vliv.

Bakalářská práce je členěna celkem do pěti kapitol. První kapitola krátce představuje téma práce a základní východiska. Druhá kapitola probírá klíčové pojmy a metody práce: představuje jednak typologický koncept hierarchie životnosti, jednak koncept gramatického/behaviorálního profilu; následně je prezentována explorativní metoda klastrové analýzy, v níž se pracuje právě s gramatickými profily, nakonec je ukázána inferenční metoda náhodných lesů, která později slouží ke statistickému ověření, zda klastry nalezené pomocí předchozí metody reflektují životnost. Ve třetí kapitole je představen design výzkumu: (i) explicitně je představena výzkumná otázka, (ii) nastíněny jsou obě metody ve vztahu k datům; (iii) dále je v největší části kapitoly ukázán celý postup: jak byla extrahována data, jak byla sémanticky anotována na různé sémantické kategorie reflektující životnost (autor zvolil více stupňů rozdělení, aby mohl odhalit rozdíly na různých místech rozlišení životnosti) a jak byly připraveny gramatické profily (autor opět zvolil různé způsoby klasifikace, které kombinují kategorie pádu a čísla); následně je krok za krokem ukázáno, jak autor postupoval při tvoření statistických testů (klastrová analýza, náhodné lesy) a jejich vyhodnocování (průměrná šířka siluety, porovnání dendrogramů, bodové grafy) ve statistickém programu R. Čtvrtá kapitola se věnuje konkrétním výsledkům: (i) autor zde prezentuje výsledné dendrogramy (ty vzhledem ke komplikovanému grafickému řešení uvádí pouze v přílohách) a pomocí kontingenčních tabulek prezentuje rozložení sémantických kategorií do jednotlivých klastrů, (ii) modely podmíněných inferenčních stromů (součást metody náhodných lesů) slouží k ukázání, které sémantické kategorie životnosti měly nejvyšší vliv na rozložení substantiv do jednotlivých klastrů. Poslední pátá kapitola je věnována stručnému shrnutí závěrů.

Hodnocení

Autor v bakalářské práci jasně prokázal schopnost odborné lingvistické práce, zejména: (i) nastudoval si odbornou literaturu k teorii i metodologii, (ii) zorientoval se ve statistických analýzách, (iii) provedl a vyhodnotil vlastní empirický výzkum. Text je přehledně strukturován na kapitoly a oddíly, je psán srozumitelným jazykem, nezřídka však s formulačními nepřesnostmi. Postup práce je prezentován přehledně a schematicky. Pozitivní hodnocení převažuje, přesto je potřeba jeden důležitý aspekt zhodnotit více kriticky: práce není adekvátně dotažena z hlediska prezentace a interpretace výsledků.

V čtvrté kapitole jsou představeny výsledky statistických analýz. Jejich interpretace se bohužel zastavuje v okamžiku, kdy by kvantitativní výsledky měly významně doplnit úvahy týkající se jazyka a jeho chování. Práce tak nemůže dospět k adekvátnímu zpětnému vztažení vůči výchozím předpokladům, které se týkají samé podstaty tohoto výzkumu. V 4. kapitole autor hovoří o číslech klastrů a porovnává je z hlediska zastoupení jednotlivých sémantických tříd; není však zřejmé, jakými frekvenčními hodnotami pádů jsou substantiva v těchto klastrech naplněny. Klastrová analýza je metodou exploratorní – číslo klastru by nemělo být výsledkem, ale mělo by upozornit na zajímavé společné jazykové vlastnosti dat. Bylo by na

místě srovnat i klastry mezi sebou: ukázat například, v čem se liší/shodují u dendrogramu *pád* klastry 2 a 3, které zahrnují životná substantiva, co je charakteristické pro klastry s převahou neživotných substantiv apod. Zpětný pohled do dat by také vrhnul jiné světlo na zájmena. Přítomnost zájmen v klastru s neživotnými lze opravdu interpretovat tím, že čeština je jazyk pro-drop, ale taková úvaha musí vycházet z toho, že tento klastr je charakteristický mj. nízkou přítomností nominativu; zároveň by se ukázalo, že zájmena tvoří zcela zvláštní skupinu kvůli tomu, že mají velmi vysokou relativní frekvenci dativu.

Z hlediska metody náhodných lesů pak není tolik zajímavé závěrečné porovnání jednotlivých stupňů kategorizace (životnost, makrokategorie, kategorie – které jsou jen různě hrubým kategorizováním téhož), jako spíše ještě další detailní průzkum podmíněných inferenčních stromů, tj. prozkoumání skupin, které se nejjasněji vyčlenily (a snažit se o vysvětlení proč). Metodu náhodných lesů by také bylo na místě obohatit o další proměnné, které jsou v primárních datech k dispozici (zejména role absolutní frekvence, rodu).

Jednotlivé připomínky:

- Nekonzistence: na s. 15 autor uvádí, že bude dále pracovat s hierarchií z příkladu (3), tato hierarchie se však už dále neobjevuje, využívá se upravená hierarchie z Dixona (s. 21).
- Chybná interpretace: na s. 17 je uvedeno, že stromový algoritmus pracuje tak, že v případě kategoriálních proměnných odděluje vždy jednu hodnotu proměnné od ostatních hodnot. I v prezentovaných výsledcích je vidět, že hodnoty proměnných mohou být mezi dvě množiny distribuovány v různých kombinacích.

Otázka do diskuse: Práce uzavírá, že životnost má na pádovou distribuci značný vliv. Zároveň se ukazuje potřeba jemnější kategorizace. Dokázal byste navrhnout vlastní upravenou hierarchii životnosti pro tento jev v češtině, která by zohledňovala zjištěné výsledky? (Tj. která by kombinovala relevantní kategorie i makrokategorie.)

Vzhledem ke všemu výše uvedenému práci doporučuji k obhajobě a navrhuji hodnocení *velmi dobře*.

Jan Křivan, 5. 9. 2017