Posudek na bakalářskou práci Vojtěcha Jandy *Frekvenční distribuce nominální flexe v češtině*

Mgr. David Lukeš

5. září 2017

Posudek

Předložená práce vychází z velmi ambiciózního a chytře koncipovaného projektu, za nímž očividně leží velký kus vykonané práce a nemalé intelektuální úsilí. Nicméně příležitostně trpí nedostatkem důslednosti, pečlivosti a rozmyslu ve fázi přetavení nabytých vědomostí a dosažených výsledků do podoby finálního textu. Ve snaze poskytnout užitečnou zpětnou vazbu se následující připomínky ponesou spíše v kritickém duchu, ale rád bych předeslal, že výzkum považuju v jádru za obdivuhodný projev schopnosti samostatné badatelské práce, k níž bylo zapotřebí osvojit si analytické metody, které jdou bezpochyby nad rámec běžné průpravy bakalářského studia Obecné lingvistiky. I z tohoto důvodu upřímně doufám, že bude autor v započaté vědecké dráze pokračovat.

Kolega Janda si vytyčil za cíl prozkoumat, jak se u českých substantiv liší frekvenční distribuce pádů v závislosti na tom, do jaké kategorie životnosti spadají. Materiálově vychází z korpusu SYN2015, z nějž přebírá i morfologické značkování, podle nějž určuje příslušné frekvenční distribuce. Ručně pak anotuje jednotlivé kategorie životnosti na základě hierarchie, která vychází z existující literatury, ale obsahuje i některá vlastní doplnění (str. 21–23). Nejjemnější dělení představují tzv. sémantické kategorie; ty jsou sdruženy do pěti makrokategorií a zastřešeny ještě nejobecnější binární proměnnou životnost. Patrně z tohoto důvodu vychází autor v analýze pouze ze vzorku substantivních lemmat vytěžených z korpusu, neboť oanotovat takto řádově desítky tisíc lemmat je zcela jasně nad rámec bakalářské práce. Způsob výběru lemmat není popsán příliš detailně, mluví se o snaze pokrýt různé frekvenční hladiny, jakým způsobem a do jaké míry je této stratifikace vzorku dosaženo ovšem není z textu úplně jasné.

Relativní frekvenční distribuce pádů je v práci nazývána gramatickým profilem (GP), v návaznosti na koncept behaviorálního profilu z oblasti *usage-based linguistics*. Autor pracuje se třemi různými operacionalizacemi, které různými způsoby kromě distribuce pádů v úzu zohledňují i distribuci čísel (sg. vs pl.).

Kvantitativní zpracování je dvoufázové. První, exploratorní fáze spočívá v hierarchickém klastrování lemmat na základě podobnosti jejich GP. Optimální počet shluků (tak, aby byly co nejkompaktnější) je určen pomocí metody *average silhouette width*, přičemž u všech tří operacionalizací GP autor bere v potaz několik prvních řešení s nejvyšším skóre a volí to, které je nejblíž počtu stanovených makrokategorií životnosti (5), neboť ho "zajímá, jestli bude oněch pět klastrů (alespoň částečně) odpovídat mé pětistupňové hierarchii životnosti" (str. 28).

Druhá fáze statistického zpracování je inferenční. Klastry lemmat vytvořené v předchozím bodu jsou považovány za závislé proměnné, umístění na škále životnosti (ve všech třech úrovních granularity, tj. kategorie, makrokategorie i zastřešující binární proměnná životnost) jsou použity jako prediktory, a metodou podmíněných inferenčních stromů, resp. náhodných lesů se autor snaží určit, zda různé úrovně životnosti statisticky významně predikují přináležitost k různým klastrům. Jinými slovy, zda strukturu v datech nalezenou pomocí frekvenčních distribucí (GP), tj. klastry, lze alespoň částečně vysvětlit variabilitou v oblasti životnosti.

Právě napojení obou fází představuje slabý bod metodologické koncepce práce, neboť výstup jedné analýzy (klastrování) možná až příliš zbrkle žene na vstup analýzy další (stromy / lesy). Jak ukazuje analýza kontingenčních tabulek sémantických kategorií vs klastrů v odd. 4.1.3, shluknutí lemmat do klastrů není vždy intuitivní a zasloužilo by si ještě o něco hlubší prozkoumání a interpretaci, než jsou mu v textu věnovány. Nedostane se nám např. souhrnného komentáře ohledně namapování klastrů na makrokategorie, což je sice deklarovaný důvod, proč byla zvolena řešení s počtem klastrů blízko pěti (viz výše), ale reálně už se autor v textu k této otázce nevrací a neodpovídá na ni. Klastry samy o sobě jsou tedy v některých ohledech interpretační oříšek, a použijeme-li je v následných inferenčních analýzách jako závislou proměnnou, interpretační složitost se znásobuje, neboť není úplně jednoduše uchopitelné, jaké že vlastně kategorie nezávislé proměnné predikují. Tím nechci tvrdit, že jsou výsledky na nich založené špatně – jen orientace v nich je pak zejména pro čtenáře poněkud složitá, neb jsou vystavěné na vratkých konceptuálních základech.

Autor také mohl věnovat více času analýze stability jednotlivých identifikovaných klastrů. Jistým indikátorem může být i popsaná vizuální inspekce

tanglegramů, ty ale porovnávají výsledky klastrování napříč třemi různými operacionalizacemi GP, ne stabilitu klastrů uvnitř jedné operacionalizace. Takového odhadu je možno dosáhnout pomocí bootstrapování: opakovaně klastrujeme náhodné vzorky vybrané z celkových dat a sledujeme, které klastry se spolehlivě opakují. Tímto způsobem by bylo možné stanovit, kterým klastrům má smysl věnovat zvýšené interpretační úsilí, a které jsou naopak spíše náhodnými artefakty odvislými od daného vzorku. Domnívám se, že pokud by byla analýza klastrů takto dotažena do důsledku, včetně důkladnějšího rozboru toho, jak se v klastrech sdružují jednotlivé (makro)kategorie životnosti, byly by závěry práce v detailech lépe srozumitelné a více informativní. Každopádně by to na bakalářskou práci bohatě stačilo, nebylo by nutné se pouštět do druhé, inferenční fáze.

Rozumím potřebě opatřit výsledky p-hodnotou, ale někdy to může být spíš na škodu. Mnohem zajímavější veličinou je síla efektu (zde (ne)vyhraněnost klastrů z hlediska (makro)kategorií životnosti), a jako odhad statistické významnosti (= spolehlivosti, opakovatelnosti) by koneckonců mohly posloužit výsledky bootstrapování.

Je také potřeba si být dobře vědom, k čemu přesně testování významnosti nulové hypotézy (*null hypothesis significance testing*, NHST) slouží a k čemu ne, resp. co nám vlastně přesně o datech říká. Např. ať už se p-hodnota dostane pod sebenižší hranici, nelze tvrdit, že tím "byla jednoznačně vyvrácena nulová hypotéza a současně potvrzena hypotéza alternativní" (str. 49). p-hodnota je pravděpodobnost, že nasbíráme stejná data, jako máme, nebo dokonce extrémnější, za předpokladu, že platí nulová hypotéza. Tato pravděpodobnost může být arbitrárně nízká, ale nikdy nebude nulová. (Co víc, říká nám pouze to, zda se nám podařilo nasbírat tolik dat, abychom si mohli dovolit přestat pozorované rozdíly přičítat náhodnému šumu – ne jestli má nalezený rozdíl nějaký citelný reálný dopad, od toho je tu právě síla efektu, a co hůř, ani nám nezaručí, že ten rozdíl nevzešel z nějaké systematické, tj. nenáhodné chyby.) Vyvrácení nulové hypotézy tedy není událost, kterou lze *ex post* stupňovat podle toho, jak malého p jsme dosáhli. Je třeba si *předem* zvolit, jak malé p požadujeme, abychom *přestali věřit* v nulovou hypotézu – je to taková sázka sám se sebou. Buď si vsadím konzervativně (hladina 0,05), mám větší šanci vyhrát, ale výhra není tak cenná, nebo zariskuju, protože si věřím (hladina 0,01 nebo dokonce 0,001), a vyhrát je pak těžší a o to cennější. Ale nemůžu si vsadit až ve chvíli, kdy znám výsledky slosování, a nemůžu vyhrát víc, než kolik stanovila sázka. Stran alternativní hypotézy by pak přiléhavější tvrzení znělo, že pokud data vykazují dostatečně nízkou podporu pro H0, dovolím se věřit v H1.

Ještě ke srozumitelnosti a orientaci ve výsledcích: notně ji ztěžuje oz-

načení různých kategorií životnosti pomocí čísel místo mnemotechnických štítků. Už sám fakt, že je potřeba číselné sloupce při importu do R ručně převést na faktory (str. 33), by měl být signálem, že reprezentace dat není ideální (textové sloupce s opakujícími se řetězci převede R při importu na faktory automaticky). Navíc se během interpretace přidají i klastry, kterým taktéž není přiřazen žádný konceptuální štítek, opět jsou jen očíslovány. Když k tomu přidáme ještě čísla uzlů v inferenčních stromech a vše vynásobíme třemi dílčími analýzami pro tři různé operacionalizace GP, vzniká zmatek, v němž se místy ztrácí i autor: např. tvrzení, že "Zájmena, abstrakta, toponyma, věci denní potřeby a slovesná substantiva (uzel 4) se významně ($p<0,05$) shlukují v klastru 4" (str. 44) neodpovídá popisovanému obr. 3, kde v uzlu 4 jasně dominuje klastr 1, ne 4.

Tím postupně přecházím k problémům s prezentací výzkumu zmíněným v prvním odstavci posudku. Asi jako každé kvalifikační práci by byla textu prospěla korektura s lehkým časovým odstupem, v případě abstraktu pak kontrola od rodilého mluvčího (substantiva jsou anglicky *nouns*, slovo *nominal* obvykle označuje kategorii sdružující substantiva a adjektiva). S tím ovšem čtenář víceméně počítá a nepřijde mi smysluplné na tom přehnaně lpět. Závažnějším nedostatkem ale je nerovnováha mezi analytickým a syntetickým přístupem: často se velmi detailně dozvídáme co a jak, méně už proč a co z toho lze konceptuálně vyvodit. Příkladem budiž kapitola 3, která obsahuje vyčerpávající přepis použitých příkazů v R, včetně nahrání dat, to vše třikrát pro tři různé operacionalizace GP, ačkoli jsou postupy identické. Přitom by stačil obecný konceptuální popis těchto metod, zdůvodnění jejich použití a jak do sebe v rámci plnění stanovených cílů práce zapadají. Implementační detaily jsou z hlediska replikovatelnosti výzkumu pochopitelně také důležité, jejich místo je ale v samostatné příloze, která bude kromě okomentovaného skriptu obsahovat i podkladová data k analýze, bez nichž je přínos zveřejnění přesného postupu zlomkový.

Podobně působí i výsledková kapitola 4, která se převážně pod drobnohledem věnuje dílčím komentářům, nachází různé pravidelnosti, ale většinou už pro ona pozorování nenabídne soubornou konceptuální interpretaci, viz např.: "Z neživotných skupin zaujala abstrakta s rostlinami, tedy lemmata, jejichž gramatické profily v obou dosud prezentovaných modelech do značné míry narušují předpoklad tím, že se shlukují v klastrech s životnými substantivy" (str. 43). Proč se zrovna abstrakta a rostliny chovají podobně jako některá životná substantiva? V takové chvíli je potřeba se vrátit k definici kategorií (která mimochodem obsahuje konkrétní motivace k vydělení některých skupin), prozkoumat lemmata, která do nich spadají, projít GP charakteristické pro daný klastr, případně se podívat i do konko-

rdancí v korpusu (návrat ke zdrojovým datům) a pokusit se poskytnout vysvětlení, kde se tyto paralely v chování inkriminovaných lemmat v textu berou.

Finální textové zpracování tedy působí dojmem, že na něj po obrovském a úctyhodném kusu velmi dobré práce odvedené na datové analýze již nezbylo mnoho času, a to je škoda. Práce je místy psaná ještě z perspektivy před vykonáním výzkumu, toto by bylo záhodno sjednotit a vyhladit pro čtenáře různé slepé uličky, které již z hlediska perspektivy *post-hoc* nejsou relevantní. Formulace nejsou vždy úplně pregnantní, např. věta "V typologickém pojetí je životnost principem, že životnější jednotky mají tendenci se chovat jinak než jednotky méně životné." (str. 10) nám o konceptu životnosti sdělí jen to, že je jedním z mnoha kritérií pro klasifikaci jazykových jednotek , ale nic o jeho podstatě. Strukturace textu se postupem času čím dál víc opírá místo běžných konektorů o výčty ((i), (ii) atd.), převážně v situacích, kde jsou naprosto zbytečné (na přidělená čísla už se nikde dál neodkazuje) a působí naopak rušivě. Navzdory snaze o až příliš přesný popis postupu chybí zmínka o některých klíčových balíčcích pro R (např. `party`, z něhož pochází funkce `ctree`). Některé citace či parafráze jsou nepřesné, např. rozpětí hodnot *average silhouette width* je podle Levshiny (2015: 311) <0; 1> a minimální přijatelná hodnota 0,2; práce uvádí údaje <-1; 1> a 0,25 (str. 38). Podobně popis algoritmu pro podmíněné inferenční stromy se v některých bodech rozchází se zdrojem (Levshina 2015: 291):

> Má-li vybraná proměnná hodnot více, z datasetu je jedna oddělena a zbytek zůstane v celku. [Toto ze zdrojového textu nevyplývá, jen že se hodnoty, kterých proměnná nabývá, rozdělí do dvou skupin.] Má-li vybraná proměnná číselné hodnoty, algoritmus je rozdělí na dva stejně velké intervaly a data rozdělí podle nich (např. hodnoty 0–100 by byly rozděleny na intervaly 0–50 a 51–100). [O tom, jakým způsobem se číselná proměnná rozdělí na dva intervaly, zdrojový text také nic neříká, tj. nemusí to nutně být na dva stejně velké intervaly.]
>
> (str. 17; komentáře v hranatých závorkách: DL)

Přes výhrady uvedené výše nicméně chci zopakovat, že se jedná o úctyhodný kus práce, který je nepochybným dokladem autorovy schopnosti samostatně pracovat s literaturou, osvojit si složité metodologické postupy a kreativně je aplikovat. I proto je škoda, že interpretační stránka a výstavba textu tento potenciál plně nevytěžují. Navzdory tomu předložená stať bezpochyby splňuje požadavky kladené na bakalářskou práci; tímto

ji doporučuju k obhajobě a navrhuju hodnocení v rozpětí **velmi dobře** – **výborně**, v závislosti na průběhu obhajoby.

Témata k diskusi při obhajobě

V rámci obhajoby bych byl rád, kdyby se autor vyjádřil k následujícím otázkám / bodům:

- Prosil bych o podrobnější komentář k *usage-based linguistics* jako teoretickému východisku. Za pojítko autor označuje důraz na frekvenci, nicméně zatímco *usage-based linguistics* používá frekvenci k vysvětlení některých jevů v jazyce ("frekvence užití může podmiňovat změny v jazyce", str. 10), předkládaná práce používá koncept životnosti k vysvětlení právě frekvence ("životnost [se] projevuje v jazyce prostřednictvím frekvence", str. 10). Kauzalita je tu z hlediska role frekvence přesně opačná. Spadá tedy vůbec práce do *usage-based linguistics*? Pokud ano, tak v jakém smyslu?

- Proč byl z GP odebrán vokativ v práci okomentováno je (str. 21), ale rád bych se k tomu ještě vrátil. Není vokativ pád, na němž by se zrovna právě životnost měla markantně projevit?

- "Dále byla před výběrem slov k anotaci na sémantické kategorie odebrána ta lemmata, která měla víc než 9 tvarů s nulovým výskytem." (str. 21) Tento postup znamená, že existují-li frekventovaná lemmata, která se nicméně vyskytují jen v malém počtu forem, chceme je vynechat. Proč?

- Všechny tři prediktory u stromů / lesů (kategorie, makrokategorie, životnost) jsou založené na různě granulárním dělení téže škály životnosti, jsou tedy nutně korelované. Jaký to může mít dopad na použití metod stromů / lesů obecně, a konkrétně na analýzu důležitosti prediktorů v odd. 4.2.2?

- Autor se snaží odstínit vliv absolutní frekvence výskytu lemmatu na výsledky stratifikovaným vzorkováním. Pokoušel se absolutní frekvenci také zahrnout mezi prediktory a formálněji tak ověřit, zda má na GP nějaký vliv?

- Šlo by nějak lépe převést do češtiny termín *usage-based linguistics*? Jsou "lingvistické postupy založené na užívání" (a příbuzné varianty)

ustálený překlad? (Jde mi zejména o slovo "užívání", které se mi v tomto kontextu nezdá úplně jednoduše srozumitelné.)

Literatura

Levshina, Natalia. 2015. How to Do Linguistics with R. Data Exploration and Statistical Analysis. Amsterdam/Philadelphia: John Benjamins.