Univerzita Karlova

Filozofická fakulta

Ústav obecné lingvistiky

Bakalářská práce

Vojtěch Janda

Frekvenční distribuce nominální flexe v češtině

Frequency distribution of nominal inflection in Czech

2017

Vedoucí práce: Mgr. Jan Křivan, PhD.

Odborný konzultant: Mgr. Michal Láznička

Poděkování

Chtěl bych poděkovat vedoucímu své bakalářské práce, Mgr. Janu Křivanovi, PhD., za nekonečnou řadu připomínek a rad i trpělivost s mými dotazy. Rád bych také poděkoval odbornému konzultantovi této práce, Mgr. Michalu Lázničkovi, za to, že mi téma představil, přivedl mě k němu a pomohl mi se v něm orientovat.

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 14. 8. 2017 ……

Vojtěch Janda

Abstrakt

Tato práce využívající metod lingvistických přístupů na základě užívání ověřuje, zda lze vysvětlit rozdílnou frekvenční distribuci pádů u jednotlivých substantiv v češtině pomocí hierarchie životnosti. Z vyváženého korpusu současných psaných textů SYN2015 extrahuji gramatické profily substantiv, skládající se z informací o rodu a čísle; na těchto profilech provádím klastrovou analýzu, jež dělí gramatické profily do skupin substantiv s podobnou relativní frekvenční distribucí pádů. Na základě klastrové analýzy a modelování podmíněných inferenčních stromů potvrzuji, že životnost rozděluje sledovaný vzorek na dvě skupiny.

Klíčová slova:

hierarchie životnosti, klastrová analýza, korpusová metoda, usage-base linguistics

Abstract

Employing methods of usage-based linguistic approaches, this paper tests the claim that differerences in frequential distributions of cases of nominals in Czech can be explained with the animacy hierarchy. Grammatical profiles consisting of information about gender and number are extracted from SYN2015, a balanced corpus of contemporary written texts, and analysed by hierarchical clustering which groups the grammatical profiles according to similarities of relative frequential distribution of cases. The cluster analysis and subsequential conditional inference tree modelling that animacy divides the sample into two groups.

Key words:

frequency analysis, animacy hierarchy, cluster analysis, corpus method

Obsah

Seznam zkratek

1 – nominativ

2 – genitiv

3 – dativ

4 – akuzativ

6 – lokativ

7 – instrumentál

sg – singulár

pl – plurál

BP – behaviorální profil

GP – gramatický profil

1. Úvod

Tato práce je kvantitativním pohledem na českou nominální flexi. Za použití statistických explorativních nástrojů hledá kritéria pro predikci hodnoty sémantické kategorie životnosti na základě relativních frekvencí četnosti výskytu jednotlivých tvarů slova.

Hlavním východiskem práce je životnost jako koncept jazykové typologie, jelikož od něj se odvíjí návrh vlastního výzkumu. V typologickém pojetí je životnost principem, že životnější jednotky mají tendenci se chovat jinak než jednotky méně životné. Jako přirozená kategorie životnost odpovídá dělení okolního světa na živé a neživé, případně škálu, která je na každém konci uzavřená jedním z těchto pólů. V jazycích se životnost projevuje různými způsoby. Například v češtině se projevuje v morfologické kategorii životnosti u maskulin: u životných maskulin dochází k distinkci nominativu a akuzativu (tedy značení subjektu a objektu). Podobně z výzkumu (např. Silverstein 1976, Dixon 1979 aj.) jazyků s ergativním rozštěpením vyplynulo, že životnější jednotky mívají akuzativní značení, zatímco ty méně životné bývají značeny absolutivem. Na základě těchto pozorování se předpokládá, že sémantika může mít značný vliv na morfosyntax. Předpokládá se tedy, že se životnost v mnoha jazycích alespoň do jisté míry projevuje. Lingvistické postupy založené na užívání (tzv. usage-based linguistics; Bybee 1985, Hopper 1987, Diessel 2014) souběžně předopkládají, že frekvence užití může podmiňovat změny v jazyce. Pokud spojíme tyto dva principy dohromady, můžeme se ptát, zda se životnost projevuje v jazyce prostřednictvím frekvence (tj. zda životnější jednotky mají jinou frekvenční distribuci než jednotky méně životné).

Podobný přístup najdeme u Levshiny, jejíž případová studie (Levshina 2015: 301–321) dokládá, že některá anglická slovesa v některých konstrukcích daleko častěji vyžadují životné argumenty. V její práci jsou data analyzována metodou hierarchické klastrové analýzy, tedy heuristického statistického nástroje pro seskupování datových položek na základě podobnosti napříč prakticky libovolným množstvím kritérií. Tuto explorativní metodu ve své práci využiji také, protože mě zajímá, jestli se lemmata patřící do stejné sémantické skupiny podle životnosti užívají v jazyce podobně (mají podobnou relativní frekvenční distribuci), a tedy jestli lze stupeň životnosti substantiva či pronomina aproximovat právě na základě relativní frekvenční distribuce jeho tvarů. Výsledky explorativní metody, stejně jako jejich interpretaci, je pak třeba ověřit, což provedu jednak sekundárními nástroji klastrové analýzy, jednak metodou vytváření náhodných lesů, podtypem strojového učení.

Konkrétně se v této práci budu zabývat hledáním statisticky významného vztahu mezi frekvenční distribucí pádů a stupni životnosti. Pokud se mi takový vztah nepodaří prokázat,

položím si otázku, jaké jiné faktory mohou ovlivňovat rozložení lemmat napříč klastry. Výstupem práce bude odpověď na tyto otázky, vytvořená na základě těchto analýz.

Kapitola 2 je uvedením do teoretických a metodologických východisek a vysvětluje několik pojmů, s nimiž pracuji v empirické části: *hierarchie životnosti*, *gramatický profil*, *hierarchická klastrová analýza* a *náhodný les*.

Ve třetí kapitole popisuji design výzkumu od formulace výzkumných otázek přes vymezení předmětu výzkumu a způsobu práce s životností, kdy představím i výběr sémantických skupin pro statistické zpracování, až po detaily sběru a zpracování dat a popis postupu při aplikaci konkrétních statistických nástrojů, kterými jsou hierarchické klastrování gramatických profilů a měřítka vhodnosti zvolené metodiky. Zkonstruováním podmíněných inferenčních stromů a náhodných lesů poté ověřuji, jestli sestavené klastry gramatických profilů lze vysvětlit některou z připravených škál životnosti.

Nakonec z empirického zkoumání odvozuji závěry, které prezentuji v kapitole 5.

2. Teorie a metodologie

Tato první část práce popisuje teoretická a metodologická východiska, na nichž bude vystavěna část empirická, výzkumná. Postupně zde představím a vysvětlím pojmy *hierarchie životnosti*, *gramatický profil*, *hierarchická klastrová analýza* a *náhodný les*.

Základním předpokladem této práce je to, že jazyk jako systém je propojený s užíváním jazyka a tyto dva aspekty se navzájem ovlivňují. Tento postulát je základem přístupů založených na užívání (usage-based approach). Cílem lingvistického výzkumu založeného na užívání je vytvoření rámce pro výzkum jazyka s ohledem na jeho kořeny v základních kognitivních procesech, jako je kategorizace nebo analogizace (Diessel 2014). Mezi přínosy tohoto přístupu patří zjištění, že frekvence zřejmě hraje značnou roli v jazykové změně a vývoji jazykových struktur (viz např. Bybee 1985). Na tomto poznatku je vystavěna i metoda zkoumání pomocí *behaviorálních* a *gramatických profilů* (viz oddíl 2.1), kterou v této práci blíže představím a následně uplatním.

2.1 Gramatický profil

Koncept gramatického profilu (GP) navazuje na korpusově-sémantický výzkum Stefana Griese, Naokiho Otani, Dagmar Divjak a dalších, kteří v rámci svých prací představili a používali takzvané *behaviorální profily* (BP; např. Gries & Otani 2010, Divjak & Gries 2006 aj.). Tyto behaviorální profily představují sloučení syntaktického a sémantického korpusového výzkumu kolokability a souvýskytu, synonymie a antonymie. Jde v podstatě o informaci o tom, jaká slova se vyskytují v blízkosti zkoumaných lemmat, o odpovědi na následující otázky: jakého jsou slovního druhu, do jaké domény patří, jak často se s daným lemmatem pojí? Následná *hierarchická klastrová analýza* (viz samostatný oddíl 2.3) pak ukazuje vztahy a rozdíly mezi zkoumanými lemmaty na základě těchto informací. Metoda se skládá ze čtyř kroků. (i) Získání reprezentativního vzorku všech výskytů zvoleného lemmatu. (ii) Anotace konkordancí na řadu morfologických, syntaktických, sémantických a dalších charakteristik, následně označovaným jako ID tagy. (iii) Sestavení tabulky (BP), která ukazuje relativní frekvence souvýskytů každého lemmatu/smyslu s každým z ID tagů. Tyto relativní frekvence tvoří celek (100 %) pro každý souvýskyt u jednotlivých lemmat/smyslů. (iv) Vyhodnocení souvýskytových dat pomocí statistiky, tj. korelování, hierarchické klastrování aj. (Gries 2010). Tak například Gries & Otani (2010:141) za pomoci metody BP zjišťují, že co se ID tagů týče, anglické *tiny* ‚drobný' znamená to samé co *smallest* ‚nejmenší'

nebo že jako antonymum k *big* ‚velký' se chová *little* ‚malý', zatímco výraz *large* ‚velký' má podle behaviorálních profilů za antonymum výraz *small* ‚malý'.

Gramatický profil je v některých aspektech jednodušší. Při vytváření gramatického profilu nedochází k anotaci konkordancí, na místo toho je sledována předem zvolená proměnná či malá sada proměnných, které bývají již oanotovány v lemmatizovaných korpusech. GP je, stručně řečeno, informací o frekvenci užívání tvarů jednotlivých paradigmat, definovaná Jandou & Lyashevskayou (2011: 719) takto:

> *„A grammatical profile is the relative frequency distribution of the inflected forms of a word in a corpus."*

V jejich práci sloužily gramatické profily k ověření tzv. tradičního pohledu na problematiku tvorby vidových dvojic v ruštině. Podle tradičního pohledu lze vidový protějšek slovesa vytvořit jak sufixací, tak prefixací. Oproti tomu tzv. Isačenkova hypotéza říká, že vidový protějšek slovesa lze v ruštině vytvořit pouze sufixací (Janda & Lyashevskaya 2011: 719). Porovnáním gramatických profilů prefixem rozlišených a sufixem rozlišených vidových dvojic pak vyšlo najevo, že mezi těmito dvěma způsoby slovotvorby zřejmě není významný rozdíl (Janda & Lyashevskaya 2011: 752–754).

Další studii zabývající se ruskými slovesy a využívající gramatické profily přinesla J. Kuznetsova (2015), která pomocí této kvantifikace ukazuje společné sémantické vlastnosti (genderové stereotypy) sloves a jejich preferenci pojit se s argumenty příslušných jmenných rodů. Sledovanou charakteristikou v její práci byly relativní frekvence rodových koncovek sloves, příkladem jsou slovesa čarodějnictví, která se významně častěji vyskytují s koncovkami ženského rodu.

Podobně jako Kuznetsova (2015) zde sestavuji gramatické profily, abych na základě frekvencí různých tvarů substantiv odhalil možná sémantická pojítka. Ukázka gramatického profilu z mé práce, s frekvencemi již převedenými na relativní, je vidět v Tabulce 1.

Lemma, rod, skupina	**1SG**	**2SG**	**3SG**	**4SG**	**6SG**	**7SG**	**1PL**	**2PL**	**3PL**	**4PL**	**6PL**	**7PL**
Kmotra F 3	0,24	0,13	0,02	0,08	0,03	0,21	0,07	0,01	0,00	0,20	0,00	0,01

Tabulka 1: Gramatický profil ukazující zaokrouhlené relativní frekvence výskytů lemmata „kmotra" podle smíšené kategorie čísla a pádu.

Popis gramatického profilu lze shrnout tak, že jde o charakteristiku vytvořenou na základě jazykového užívání, která může zprostředkovat kognitivně-funkční vysvětlení některých jinak neodhalitelných jevů. Výhodou konceptu gramatického profilu je jeho původ v korpusových datech, respektive to, že gramatický profil je snadné použít v kvantitativním výzkumu, jelikož jde již z podstaty nejen o vyčíslitelný, ale přímo vyčíslený element.

2.2 Hierarchie životnosti

Druhým klíčovým pojmem této práce je hierarchie životnosti. Jde o koncept lingvistické typologie, který vychází z pozorování, že v mnoha jazycích světa jazykové jednotky, které označují více životné referenty, mají tendenci se formálně či svým užitím lišit od jazykových jednotek označujících méně životné referenty. Princip poprvé popsal Michael Silverstein (1976) ve své práci o ergativním rozštěpení v australských jazycích. Jím objevené rozdíly v chování jazykových jednotek ukázaly vhodnost rozlišování hierarchie (Příklad 1), která zahrnuje rozdíly v životnosti, osobě, referenčnosti (obvykle souhrnně nazývaná hierarchií životnosti či nominální hierarchií).

Příklad 1: *zájmena 1. osoby > zájmena 2. osoby > zájmena 3. osoby > vlastní jména > substantiva označující lidi > životná substantiva > neživotná substantiva*

– Silverstein 1976, cit. podle Dixon (1994: 85)

Jev byl od té doby rozpracován a aplikován na řadu jazykových struktur (např. Dixon 1979, 1994, Langacker 1991), viz shrnutí Comrie (1989). V současnosti je koncept dále propracováván, ale i zpochybňován. Například Elena Filimonova (2005) se zabývá konkrétními protiargumenty ke konceptu hierarchie životnosti a shledává, že jazykové jevy vymykající se hierarchické pravidelnosti lze obvykle vysvětlit probíhajícími změnami v pádovém systému daného jazyka, přičemž výjimky jsou nejčastěji způsobovány tradičně stabilnějšími zájmeny, která se pomaleji přizpůsobují novému pádovému systému.
Comrie (1989) považuje za základní rozsah této škály rozlišování (Příklad 2) životnosti v užším smyslu.

Příklad 2: *lidské > živé > neživé* *– Comrie (1989: 185)*

Mentální předěl mezi živým a neživým se projevuje v řadě jazyků – ty se ale obvykle liší v tom, kam kladou formální hranici na této škále nebo jaké jednotky do těchto tříd patří. Vztah mezi životnými a neživotnými jazykovými jednotkami lze pozorovat i v češtině, kde gramaticky na životnost pozitivně kódovaná maskulina mají v akuzativu jinou příponu než

v nominativu, zatímco u maskulin gramaticky neživotných se tyto shodují. Podobný jev pak lze pozorovat z indoevropských jazyků také v latině, v níž mají shodné značení nominativu a akuzativu neutra, tedy jmenný rod zahrnující především ne-lidské a neživotné skutečnosti. A právě zmíněné kritérium pádového značení vytváří v mnohých jazycích další stupně, které souvisejí s dalšími kritérii jako je referenčnost či osoba, i. e. osobní zájmena první a druhé osoby singuláru (což platí i o synchronně supletivních složkách paradigmat těchto zájmen v češtině) a lidská jména vlastní, označující typické agenty (Comrie 1989: 186). Podle Comrieho je tedy důležité si uvědomit, že životnost není jediným faktorem, který může chování nominálních jednotek ovlivňovat (Comrie 1989: 197-198). Vzhledem k těmto pozorováním se životnost sleduje odděleně ve vztahu s dalšími tzv. *škálami prominence* (vedle životnosti mezi ně patří *tématičnost*, *referenčnost*, *anaforičnost*, *pronominální osoby*; viz Heine & König 2010: 94).

Pro potřeby této budu pracovat s životností v širším smyslu, která zahrnuje následující stupně (Příklad 3):

Příklad 3: *zájmena 1. /2. osoby, zájmena 3. osoby > lidská vlastní jména > příbuzenské termíny > substantiva označující lidi > substantiva životná > substantiva neživotná*

Tato hierarchie bude dále rozpracována s ohledem na povahu dat, a to zejména v oblasti neživotných substantiv, která jsou oproti podrobně rozpracovanému systému životnějších jednotek dlouhodobě opomíjena, na což upozorňuje Ji (2017). Očekávaná hierarchie zahrnuje také příbuzenské termíny, neboť na jejich relevanci v oblasti nominálních hierarchií upozorňuje již Comrie (1989: 195). Použitím statistických metod popsaných v následujících oddílech je pak možné zjistit, zda čeština přes neexistenci explicitního značení vyšší životnosti nemá tendenci používat slova pro rodinné příslušníky v jiných gramatických kontextech než jiná lidská substantiva.

2.3 Klastrová analýza

V tomto oddílu představím techniku klastrové analýzy dat jako třídícího nástroje, který seskupuje jednotky na základě různých kritérií pomocí vyhodnocování vzdáleností mezi konkrétními jednotkami, vycházeje z práce N. Levshiny (2015).

Klastrová analýza dat se používá v rámci přístupu behaviorálních profilů, kdy je každá k analýze určená jednotka označkována a definována co nejpřesněji s ohledem na výzkumné otázky daného projektu, tedy konkrétně v této práci půjde v první řadě o sestavení

gramatických profilů se jmenným pádem jako proměnnou nabývající 12 různých hodnot (diskuse GP v této práci viz oddíl 3.3.4).

Samotný proces klastrování závisí na zvoleném typu analýzy (viz zde), všechny typy však sdílejí stejný základní postup. Každé dvojici jednotek je na základě jejích vlastností a zadaných kritérií přiřazena hodnota vzájemné vzdálenosti (disimilarity), podle níž algoritmus určuje, jestli patří do stejné skupiny (klastru), nebo ne. Vytváření klastrů je, vzhledem k obvykle velkému objemu dat (a tím pádem i obrovskému množství dílčích rozhodnutí), proces heuristický, tj. takový, který pouze aproximuje optimální řešení každého jednoho dílčího uzlu.

Popsaná metoda hierarchického klastrování se užívá v jedné ze dvou základních podob, podle odlišných přístupů označovaných jako (i) *bottom-up* a (ii) *top-down*. (i) Přístup bottom-up (také *aglomerativní*) probíhá tak, že každou dvojici klastrů (zpočátku jednočlenných) spojuje nebo nechává oddělené na základě míry jejich vzájemné podobnosti (nejmenší vzdálenosti). (ii) Přístup top-down (také *divizivní*) naopak bere data jako celek, v němž jednotky vyděluje na základě vzájemné odlišnosti (vzdálenosti; viz předchozí odstavec).

Aglomerativní i *divizivní* klastrování může být řízeno různými algoritmy. Levshina (2015: 310-311) zmiňuje několik hlavních: *complete, single, average, ward*. (i) Algoritmus *complete* slučuje klastry na základě nejmenších vzdáleností mezi vzájemně nejvzdálenějšími jednotkami z každé srovnávané dvojice klastrů. (ii) Algoritmus *single* slučuje klastry na základě nejmenších vzdáleností mezi vzájemně nejbližšími jednotkami z každé srovnávané dvojice klastrů. (iii) Algoritmus *average* slučuje klastry na základě nejmenší průměrné vzdálenosti mezi jednotkami jednoho a druhého z každé srovnávané dvojice klastrů. (iv) Algoritmus *ward* minimalizuje nárůst odlišností ve vzdálenostech mezi jednotkami klastrů. Výhodou tohoto algoritmu je vznik velice kompaktních klastrů.

Výstupem analýzy jsou jednotky rozdělené do klastrů s podobnými vlastnostmi na základě předem zadaných kritérií; výsledky budou prezentovány pomocí *klastrového dendrogramu*, což je grafické znázornění ukazující hierarchickou strukturu mezi vytvořenými skupinami jednotek dat. Hierarchie vytvořených klastrů je v *dendrogramu* znázorněná tak, že méně podobné klastry jsou spojeny výše, zatímco podobnější klastry jsou spojeny níže. Nejvýše položená jednotka (klastr) tak představuje celá data, zatímco spodní část dendrogramu je členěná na klastry jednočlenné, takže celé znázornění připomíná kořeny stromu.

Je však nanejvýš důležité si uvědomit, že klastrová analýza je metoda explorativní, a tak je nutno jakékoliv závěry ověřit jednak testováním hypotéz, jednak například nástroji na vyhodnocování vhodnosti daného klastrového řešení, což lze pro počet klastrů provést pomocí

změření průměrné šíře siluety (*silhouette width*) pro různé počty klastrů od 2 do n-1, kde *n* je počet klastrů. Pro zjištění vhodnosti konkrétní hierarchie lze využít tzv. *multiscale bootstrapping*, který pomocí velkého počtu iterací klastrové analýzy s parametry odpovídajícími původnímu provedení vypočítá *přibližně nezaujatou p-hodnotu* každého klastru a zároveň vyhodnotí, jakou pravděpodobnost na objevení se v novém řešení analýzy (*bootstrap probability*) každý klastr má, čímž určí jeho stabilitu (Levshina 2015: 315-316 a 320-321).

V případě pochybností o vhodnosti daného řešení na základě výše uvedených způsobů jejího ověření provedu klastrovou analýzu s jinými parametry či algoritmy.

2.4 Náhodný les (random forest)

Náhodné lesy jsou nástrojem pro neparametrické testování hypotéz. To znamená, že pomocí nich lze otestovat pravdivost hypotézy bez ohledu na rozdělení dat. Používají se jako alternativa k vícenásobným regresím, a to zejména na data menšího objemu, ovšem taková, do nichž zasahuje více predikujících faktorů (Levshina 2015: 291).

Náhodný les je ovšem až druhým stupněm analýzy: tak jako jsou skutečné lesy tvořené stromy, jsou i náhodné rozhodovací lesy tvořeny podmíněnými inferenčními stromy. Tyto stromy představují metodu regrese a klasifikace pomocí rekurzivního binárního rozdělování, které se provádí v následujících třech krocích: (i) rozhodnutí, zda jsou některé z nezávislých proměnných propojené se zvolenou závislou proměnnou. Nezávislá proměnná s nejsilnějším propojením je algoritmem vybrána a (ii) dataset je podle ní rozdělen na dvě části. Má-li vybraná proměnná dvě možné hodnoty, dataset je rozdělen podle nich. Má-li vybraná proměnná hodnot více, z datasetu je jedna oddělena a zbytek zůstane v celku. 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). (iii) První dva kroky se na každé „větvi“ opakují tak dlouho, až nezbude žádná proměnná ovlivňující výsledek na předem zadané hladině významnosti.

Mezi výhody této metody patří fakt, že algoritmus užívá permutaci k několikerému uspořádání dat, s tím že pokaždé provede statistiku znovu, čímž umožní získat p-hodnoty, které ukazují stabilitu každého dělícího uzlu, přičemž algoritmus testuje nulovou hypotézu a předpokládá nulový vliv nezávislých proměnných na data (Levshina 2015: 292).

Náhodný les je tvořen jednoduše velkým množství podmíněných inferenčních stromů. Z velkého množství stromů (lesa) pak lze vypočítat podmíněnou důležitost proměnných, která říká, které proměnné jsou pro daný soubor dat nejvíce relevantní a naopak. Nakonec lze jak

pro jednotlivý strom, tak pro les spočítat, nakolik se liší od náhody, čímž ověříme, jestli se vypočítaný model na ověřovaný dataset hodí.

3. Design výzkumu

V této kapitole detailněji popíši (3.1) jaké odpovědi by práce měla přinést (iii) stupně životnosti či sémantické domény, s nimiž budu pracovat (iv) samotný průzkum korpusových dat a jejich zpracování (v) analýzu vycházející z teorie popsané výše.

3.1 Výzkumná otázka

Dříve byly projevy tzv. hierarchie životnosti – animacy hierarchy (také nominální hierarchie – noun phrase hierarchy) zkoumány zejména v gramatice jazyků, které obsahují ergativní rozštěpení. V této práci hledám podobné projevy v užívání češtiny.

Mým primárním cílem je zjistit, zda se životnost denotátu projevuje v relativní frekvenci pádů. Je-li tomu tak, bylo by možné na základě frekvenční distribuce jednotlivých tvarů (tj. na základě gramatického profilu) přibližně předvídat, na jaké úrovni se v hierarchii životnosti daný výraz umístí. Podbodem tohoto záměru je otázka, zda se feminní a neutrální tvary budou shlukovat podle sémantické životnosti s maskuliny animaty a inanimaty, která jediná rozlišují životnost gramaticky. Zajímá mě tedy také to, jestli gramatický rod také může ovlivňovat frekvenční distribuci nominálních lemmat, nebo jestli na něm nezáleží.

Hlavní hypotézy a otázky lze tedy přesně formulovat takto:

> *H_0: Neexistuje významný vztah mezi frekvenční distribucí pádů substantiv a stupni hierarchie životnosti.*
>
> *H_1: Existuje významný vztah frekvenční distribuce pádů substantiv a stupně hierarchie životnosti.*
>
> *Alternativní výzkumná otázka: Jaké jiné faktory mohou ovlivňovat rozložení lemmat napříč klastry?*

3.2 Metoda

Cílem tohoto výzkumu je zjistit, do jaké míry spolu souvisejí frekvence dané formy a její význam, respektive – konkrétněji – jak spolu souvisejí relativní frekvence pádů jmenného lemmatu a jeho zařazení v sémantické kategorii životnosti. Chci-li takto vzájemně vztahovat dvě veličiny, je náležité nějak změřit validitu svých hypotéz. Existuje celá řada statistických nástrojů, které právě to umožňují, ale obecně lze o nich říci, že jejich výstupy jsou tím jistější, čím větší data jsou zpracovávána. I proto základní jazyková data získávám z korpusu a zpracovávám je kvantitativně.

3.2.1 Data

Základní soubor dat obsahuje všechny výskyty substantiv a osobních zájmen extrahovaných z vyváženého korpusu současné psané češtiny SYN2015 (Křen et al. 2015; o sběru dat viz níže). Každý z výskytů je morfologicky anotován, některá lemmata či jejich tvary jsou anotovány také stylisticky, tzn. jsou zařazeny do některého z nestandardních rejstříků (např. slovesný tvar „zavříno“ je označen jako „hovorový“). Pro tuto práci jsou ovšem důležité následující charakteristiky: (i) slovní druh (ii) jmenný rod (iii) číslo (iv) pád.

Za použití zmíněných značek vytvořím gramatické profily všech jednotek datasetu. Každému lemmatu přiřadím absolutní frekvence každého z pádových tvarů v obou číslech. (Později absolutní frekvence převedu na relativní při zpracování v klastrové analýze.) Mezitím vyberu na 350 lemmat různých stupňů životnosti a oanotuji na sémantické kategorie s ohledem na metodologické závěry z oddílu 2.2. Vzorek bude obsahovat lemmata s rozrůzněnou celkovou absolutní frekvencí, abych mohl vyloučit vliv frekvence užití lemmatu nad rámec jednoho tvaru.

3.2.2 Metody analýzy

Dále se budu věnovat hierarchické klastrové analýze. Vyzkouším různé metody vytváření klastrů s ohledem na metodologické závěry z oddílu 2.3 a vyberu tu, která nejlépe vystihne vybraný vzorek. Toho dosáhnu srovnáním vzniklých klastrových dendrogramů – stromových struktur shlukujících jednotky s podobnými profily.

Na základě vzniklých klastrů v případě potřeby upravím původní sémantické kategorie, aby lépe reflektovaly zjištěný stav. Dalším krokem potom bude náhodný les podmíněných inferenčních stromů, kterými ověřím existenci korelace mezi klastry a sémantickými kategoriemi.

3.3 Příprava dat

3.3.1 Sběr dat

Pro extrakci potřebných informací z korpusu jsem použil následujících dotazů v jazyce CQL (Corpus Query Language; morfologické značení viz Hajič 2004, Hajič et al. 2007, Jelínek 2008, Petkevič 2006):

1. „[tag="NN.*"]“ pro substantiva
2. „[tag="P.*"]“ pro zájmena.

Prostřednictvím rozhraní KonText jsem vypočítal frekvenční distribuci každého ze získaných slovních tvarů, které jsem společně s informacemi o lemmatu a morfologickými informacemi převedl do podoby tabulky.

3.3.2 Zpracování dat

Výchozí tabulka byla transformována tak, aby ukazovala podle absolutní frekvence sestupně seřazená lemmata s přiřazenými absolutními frekvencemi všech pádů v obou číslech.

Z tabulky byly odebrány frekvence výskytu pátého pádu (vokativu), neboť není zapojen do větné stavby, jeho užití je velice specifické a omezené jen na určitý typ lemmat. Užívání vokativu by navíc muselo být předem prozkoumáno v samostatné práci, aby mohlo být efektivně zohledněno zde.

Dále byla před výběrem slov k anotaci na sémantické kategorie odebrána ta lemmata, která měla víc než 9 tvarů s nulovým výskytem. Z celkového množství 153 307 lemmat vytěžených z korpusu jich z tohoto důvodu bylo odebráno 104 823. Vhodná slova k anotaci jsem tedy vybíral ze 48 484 možných.

Výběr pak probíhal s ohledem jednak na absolutní frekvenci (tzn. vybraná slova pocházejí z celé škály celkové absolutní frekvence), jednak na sémantické skupiny (viz oddíl 3.3.3).

3.3.3 Anotace na sémantické kategorie

3.3.3.1 Představení hierarchie

Jak bylo předznamenáno výše v oddílu 2.2, v oblasti životnosti budu vycházet z hierarchie životnosti v té podobě, v jaké ji představil Dixon (1994). Při výběru excerpt pro klastrovou analýzu jsem si však uvědomil, jak by bylo vhodné rozšířit tuto předpokládanou hierarchii směrem doprava, v oblasti méně životných jednotek, v souladu s kritikou například ze strany Ji (2017), která upozorňuje na výrazné upřednostňování životnější části hierarchie v dosavadních výzkumech v této oblasti.

Výsledný soubor kategorií ukazuje Tabulka 2 (číslo v závorce udává číslo přiřazené ve vstupní tabulce analýzy).

Sémantická kategorie [počet lemmat]	**Makro-kategorie**	**Životnost**
Osobní zájmena (včetně neživotných tvarů; 99) [3]	Zájmena	Ano
Lidská propria (vlastní jména; 1) [30]	Lidská propria	
Lidé (2, 3, 4, 5) [80]	Lidé	

Antropomorfní bytosti (6) [30]	Ne-lidé	
Domácí zvířata (7) [27]		
Zvířata, živí tvorové (8) [27]		
Rostliny (18) [23]	Neživé	Ne
Části těla (12) [24]		
Věci denní potřeby (13) [23]		
Oblečení (20) [17]		
Toponyma (14, 15) [24]		
Temporální výrazy (16) [17]		
Slovesná substantiva (9) [29]		
Zbylá abstrakta (17) [24]		

Tabulka 2: Sémantické kategorie podle životnosti v širším smyslu.

Číslování je víceméně arbitrární a mezery v něm jsou způsobeny úpravami seznamu v přípravné části práce. Vyřazeny jako irelevantní či nevhodně vymezené byly skupiny „city/nálady“ (10, rozděleno mezi „slovesná substantiva“ a „zbylá abstrakta“), „nezcizitelné“ (11, rozděleno mezi „části těla“, „věci denní potřeby“, „zbylá abstrakta“), jídlo (19, irelevantní a částečně konfliktní se skupinami „domácí zvířata“, „zvířata, živí tvorové“, „rostliny“), „vlastnosti“ (21, sloučeno se skupinou „zbylá abstrakta“).

3.3.3.2 Zdůvodnění rozdělení

Sémantické kategorie byly vytvořeny s ohledem na východiska popsaná v oddílu 2.2. „Makro-kategorie“ životnosti v Tabulce 2 tak odpovídají hierarchii navržené Dixonem (Dixon 1994) – výjimkou je kategorie zájmen, která jsou sloučena do jedné skupiny vzhledem k tomu, že pro jazyky světa nelze určit jednotné pořadí. Dalším důvodem pro sloučení osobních zájmen je nevhodnost zařazení tří samostatných kategorií po jedné položce.

Taktéž v souladu s tím, co bylo předesláno v oddílu 2.2, jsem vybraná lemmata roztřídil jemněji do skupin pracovně označených číslicemi. Makro-kategorie slov označujících lidi tak zahrnuje (i) specificky profesní pojmenování (skupina 2, např. „ředitel“), a (ii) slova označující lidi podle vztahu k ostatním (skupiny 3, 4, 5, např. „milenka“, „muž“, „bratr“).

Makro-kategorie ne-lidských živých apelativ je rozdělena na (i) antropomorfní bytosti (skupina 6, např. „duch“, „anděl“, „kentaur“), (ii) domácí zvířata (skupina 7, např. „pes“, „kráva“) a (iii) zvířata, živí tvorové (skupina 8, např. „zajíc“, „motýl“). K vytvoření těchto předělů mě vedla myšlenka, zda se výrazy pro domácí zvířata nebo mazlíčky užívají jinak než slova pro ostatní, divoká zvířata, hmyz apod. Antropomorfní bytosti jsou zařazeny, protože jde zcela jasně o tvory „živé“, zpravidla myslící, ale přeci jen ne doslova lidské.

Ve vybraných neživých slovech jsem vytvořil několik skupin, které by mohlo být zajímavé sledovat, protože by mohla být užívána trochu jinak než typické neživé skupiny výrazů, jakými jsou (i) abstrakta ve skupině 17 (např. „smutek“, „zlomek“) nebo (ii) slovesná substantiva skupiny 9 (např. „výroba“, „porovnání“). Vyděleny jsou tak např. (iii) rostliny (skupina 18, např. „réva“, „celer“), které jsou sice z biologického hlediska živé, ale chováme se k nim zpravidla jako k věcem – relevantním kritériem je fakt, že maskulina této skupiny se řadí mezi inanimata („celer“ aj.). (iv) Části těla (sk. 12, např. „hlava“, „penis“), (v) věci denní potřeby (sk. 13, např. „hřeben“, „propiska“) a (vi) oblečení (sk. 20, např. „košile“, „kabát“) jsou specifikovány jako skupiny relevantní pro posesivitu, která by také mohla ovlivnit podobu gramatických profilů. Konečně (vii) toponyma (sk. 14 a 15, např. „Brno“, „Temže“, „kopec“) byla zvýrazněna, protože obsahují především konkrétní místní názvy (a jsou tedy vysoko v hierarchii referenčnosti; viz Heine & König 2010), a na (viii) temporálních výrazech (sk. 16, např. „minuta“, „čas“) mě zajímá, zda se časové reference odlišují od zbytku abstrakt.

Sloupec „Životnost“ pak zařazuje makro-kategorie do jedné ze dvou skupin (i) živé (ii) neživé. Zájmena mohou označovat i neživotné referenty, ovšem v hierarchii životnosti (v širším smyslu; viz Dixon 1994) jsou podle svého užívání v řadě světových jazyků zařazena na vrchol. Důvodem, proč jsou v této hierarchii na vrcholu, by mohla být jejich vysoká diskurzivní prominence, resp. umístění na škále anaforičnosti (viz Heine & König 2010: 94).

Dále byla lemmatům přiřazena čísla těchto sémantických kategorií, načež jsem v programu R vytvořil jejich gramatické profily, abych mohl přistoupit k analýze statistickými postupy.

3.3.4 Příprava gramatických profilů

V tomto oddílu popíšu postupy vytváření gramatických profilů.

3.3.4.1 Načtení dat

```
data.final <- read.table (file.choose(), header = T, row.names = 1, sep="\t", encoding = "UTF-8")
```

V prvním kroku definuji proměnnou nazvanou `data.final`, do níž ukládám obsah vybraného souboru, kterým je v tomto případě tabulka ve formátu prostého textu. Tabulka obsahuje

frekvenční distribuci flexe excerpt (zde ještě s absolutními frekvencemi). Hodnoty jsou odděleny tabulátorem. První řádek tabulky, který obsahuje označení číselně-pádové skupiny („1sg“, „4pl“ apod.), je převeden na názvy sloupců, zatímco z prvního sloupce, který obsahuje seznam lemmat opatřených informací o jmenném rodu a sémantickém zařazení, vzniknou názvy řádků (viz Tabulka 1 v oddílu 2.1). Typ proměnné `data.final` je „data“.

3.3.4.2 Vytvoření gramatických profilů

```
data.table <- as.table(as.matrix(data.final))
```

Tato řádka skriptu definuje proměnnou `data.table` a ukládá do ní obsah proměnné `data.final` převedený nejprve do podoby (datatypu) matice a následně tabulky, což je nutné kvůli dalšímu kroku, kterým je výpočet relativních frekvencí pro každé lemma.

```
data.prop <- prop.table(data.table, 1)
```

Gramatický profil „číslopád“

Tímto základním postupem jsem vytvořil gramatické profily pro jedinou kategorii, kterou je morfosyntaktický pád rozlišující číslo, takže každému lemmatu je přiřazeno celkem 12 frekvencí (1sg vs. 1pl, 2sg vs. 2pl, atd.; viz Tabulka 1 v oddílu 2.1).

Gramatický profil „pád“

Vhodnost či nevhodnost zahrnutí gramatického jmenného čísla do výzkumu otestuji tak, že podobným procesem jako GP „číslopád“ nechám projít také GP „pád“. Ten jsem vytvořil (i) sečtením frekvenčních hodnot stejných pádů obou čísel pro každé lemma (hodnoty 1sg + hodnoty 1pl, hodnoty 2sg + 2pl apod.) a (ii) načtením do zvláštní proměnné v R, jejíž obsah jsem poté (iii) podobně jako u základní tabulky GP „číslopád“ převedl na relativní frekvence. Výsledný gramatický profil zobrazuje Tabulka 3).

Lemma, rod, skupina	**1**	**2**	**3**	**4**	**6**	**7**
Kmotra F 3	0,31	0,14	0,03	0,27	0,03	0,22

Tabulka 3: Gramatický profil „pád“ zobrazující zaokrouhlené relativní frekvence lemmatu kmotra podle samostatné kategorie pádu pro obě čísla

Použitý skript prozrazuje, že se od předchozího GP liší pouze pojmenováním proměnných. Zásadní změnou jsou však načítaná data.

```
data.final.jenom.pady <- read.table (file.choose(), header = T, row.names =
1, sep="\t", encoding = "UTF-8")
```

```
data.table.jenom.pady <- as.table(as.matrix(data.final.jenom.pady))

data.prop.jenom.pady <- prop.table(data.table.jenom.pady, 1)
```

Gramatický profil „2 kategorie“

Další možností bylo oddělit kategorie čísla a pádu, ovšem zahrnout je do gramatického profilu obě. Gramatický profil „2 kategorie“ jsem tedy vytvořil (i) sečtením frekvenčních hodnot stejných pádů obou čísel pro každé lemma (hodnoty 1sg + hodnoty 1pl, hodnoty 2sg + 2pl apod.), (ii) sečtením frekvenčních hodnot obou čísel pro všech šest zkoumaných pádů (1sg + 2sg + … + 7sg apod.), (iii) načtením do zvláštní proměnné v R, jejíž obsah jsem poté (iv) podobně jako u základní tabulky GP „číslopád“ převedl na relativní frekvence. Výsledný gramatický profil zobrazuje Tabulka 4).

Lemma, rod, skupina	**sg**	**pl**	**1**	**2**	**3**	**4**	**6**	**7**
Kmotra F 3	0,36	0,14	0,15	0,07	0,01	0,14	0,01	0,11

Tabulka 4: Gramatický profil „2 kategorie“ zobrazující zaokrouhlené relativní frekvence lemmatu kmotra podle samostatných kategorií a čísla pádu.

Použitý skript se opět liší pouze v názvech uložených proměnných a rozdíl oproti ostatním dvěma typům gramatických profilů spočívá v převedení dat do nové relevantní podoby.

```
data.final.2cat <- read.table (file.choose(), header = T, row.names = 1,
sep="\t", encoding = "UTF-8")

data.table.2cat <- as.table(as.matrix(data.final.2cat))

data.prop.2cat <- prop.table(data.table.2cat, 1)
```

3.4 Statistické testy

Dalším krokem bylo provedení statistických testů, které byly představeny v oddílech 2.3, respektive 2.4. V následujících dvou kapitolách představím celý postup včetně rozhodování o vhodnosti metrik, metod a algoritmů, které jsem měl k dispozici. V oddílu 3.4.1 popíši hierarchickou klastrovou analýzu, v oddílu 3.4.2 pak tvorbu náhodných lesů.

3.4.1 Hierarchická klastrová analýza

Spojení frekvencí a sémantických kategorií jsem nejprve hledal pomocí hierarchické klastrové analýzy. V této kapitole okomentuji postup v programu R.

3.4.1.1 Příprava a výběr metod

Pro hierarchickou klastrovou analýzu jsem využil balíčků (i) „Rling“ (Levshina 2014), (ii) „vcd“ (Meyer et al. 2016), (iii) „pvclust“ (Suzuki & Shimodaira 2015), (iv) „cluster“ (Maechler et al. 2017) a (v) „dendextend“ (Galili 2015) pro otevřený nástroj pro statistické analýzy R (R Core Team 2017).

(i) Balíček „Rling“ obsahuje funkce, které z hlediska hospodárnosti kódu zjednodušují některé procesy. Jako příklad slouží funkce `prop.table()`, jejíž použití jsem popsal v oddílu 3.3.4.2, a která umožňuje rychlé a jednoduché vytvoření proporční tabulky.

(ii) Balíček „vcd“ slouží ke zpracování a vizualizaci kategorických dat.

(iii) Balíček „pvclust“ slouží mj. k počítání nezaujatých p-hodnot, multi-scale bootstrappingu a výpočtu průměrné šířky siluety klastru.

(iv) Balíček „cluster“ slouží k výpočtu klastrů z matice disimilarity (disimilarity matrix).

(v) Balíček „dendextend“ přidává nové možnosti manipulace a srovnávání klastrů a klastrových dendrogramů.

Výpočet disimilarity

V první řadě bylo třeba vybrat vhodný způsob výpočtu vzdálenosti mezi hodnotami v gramatickém profilu, aby bylo možné vytvořit matici disimilarity. Levshina (2015: 306–307) doporučuje vyzkoušet více metod a výstupy porovnat, ovšem jako nejvhodnější pro výpočet disimilarity v behaviorálních profilech uvádí (i) euklidovskou metodu a (ii) metodu canberra. (i) Euklidovská metodu výpočtu vzdálenosti (disimilarity) je nejpřímočařejší a nejjednodušší na interpretaci, neboť jde o druhou odmocninu součtu druhých mocnin odlišností mezi všemi páry čísel v daném vektoru (tj. o přeponu pravoúhlého trojúhelníka, jehož odvěsny – každá v jednom rozměru – představují vzdálenost/disimilaritu každého páru čísel v daném vektoru). Příslušný kód pro gramatický profil „číslopád“ vypadá takto:

```
dist.eu<-dist(data.prop)
```

(ii) Metoda canberra je vážená manhattanská vzdálenost, tedy vážené provedení vzdálenosti získané součtem délky odvěsen pravoúhlého trojúhelníka (Lance & Williams 1967). Matice disimilarity se metodou canberra vypočítá pro gramatický profil „číslopád“ takto:

```
dist.can<-dist(data.prop, method = "canberra")
```

Klastry a příslušné dendrogramy jsem vygeneroval následovně:

```
(i)   hc.ward.eu<-hclust(dist.eu, method = "ward.D2")
      plot(hc.ward.eu)

(ii)  hc.ward.can<-hclust(dist.can, method = "ward.D2")
      plot(hc.ward.can)
```

Srovnal jsem klastry (viz Příloha 2, soubory srovnani.metod.eu.pdf a srovnani.metod.can.pdf) vytvořené na základě matic disimilarity obou metod a zjistil jsem, že dendrogramy jsou si velice podobné. S přihlédnutím k interpretačnímu hledisku (Levshina 2015: 307) jsem pro další postup vybral euklidovskou metodu.

Analogicky k matici disimilarity pro gramatické profily „číslopád" jsem vypočítal také vzdálenosti pro (i) gramatické profily „pád" a (ii) gramatické profily „2 kategorie".

```
(i)   dist.eu.jenom.pady<-dist(data.prop.jenom.pady)

(ii)  dist.eu.2cat<-dist(data.prop.2cat)
```

Klastrovací algoritmus

Klastrovací algoritmy byly popsány v oddílu 2.3. Za nejvhodnější řešení pro tuto práci považuji algoritmus ward, který díky výše popsaným mechanismům (viz 2.3, iv) vytváří kompaktní a snadno interpretovatelné klastry a je vhodný pro práci s behaviorálními profily (viz Levshina 2015).

3.4.1.2 Tvorba klastrů

Když jsem vybral příslušné metody, mohl jsem přikročit k samotné klastrové analýze.

Následující kód byl použit pro vytvoření klastrů a dendrogramů pro (i) gramatické profily „číslopád" (ii) gramatické profily „pád" (iii) gramatické profily „2 kategorie".

```
(i)   hc.ward.eu<-hclust(dist.eu, method = "ward.D2")
      plot(hc.ward.eu)
```

Do proměnné `hc.ward.eu` si ukládám hierarchii klastrů vytvořených metodou ward z matice vzdálenosti `dist.eu`. Druhý řádek pak na základě této klastrové hierarchie vytváří dendrogram (viz Příloha 2, soubor hc.ward.eu.pdf).

```
(ii)  hc.ward.eu.jenom.pady<-hclust(dist.eu.jenom.pady, method =
      "ward.D2")
      plot(hc.ward.eu.jenom.pady)
```

Zde si do proměnné `hc.ward.eu.jenom.pady` ukládám hierarchii klastrů vytvořených taktéž metodou ward z matice vzdálenosti `dist.eu.jenom.pady`. Třetí řádek pak na základě této

klastrové hierarchie opět vytváří příslušný dendrogram (viz Příloha 2, soubor hc.ward.eu.jenom.pady.pdf).

```
(iii) hc.ward.eu.2cat<-hclust(dist.eu.2cat, method = "ward.D2")
      plot(hc.ward.eu.2cat)
```

Zde si do proměnné `hc.ward.eu.2cat` ukládám hierarchii klastrů vytvořených taktéž metodou ward z matice vzdálenosti `dist.eu.2cat`. Druhý řádek pak na základě této klastrové hierarchie opět vytváří příslušný dendrogram (viz Příloha 2, soubor hc.ward.eu.2cat.pdf).

3.4.1.3 Vyhodnocení vhodnosti vzniklých klastrů

Jak jsem vysvětlil v oddílu 2.3, po vytvoření klastrů je vhodné nějakým způsobem vyhodnotit, jak vhodné rozdělení vzniklo a jak hrubě nebo naopak jemně výstupy klastrové analýzy interpretovat. Jedním takovým měřítkem je průměrná šířka siluety (více viz oddíl 2.3), dalším například vizuální porovnání dendrogramů.

Průměrná šířka siluety

Pro všechny tři typy gramatických profilů jsem si vypočítal průměrnou šířku siluety pro všechna řešení od 2 do *n*-1 klastru, abych zjistil vhodný počet klastrů, s kterým dále pracovat. Jako vodítko mi poslouží Kaufman & Rousseuw (1990), kde se jako hraniční hodnota, pod níž už obvykle není vhodné považovat klastrové řešení za dobře strukturované, uvádí 0,25.

```
(i)   asw<-sapply(2:377, function(x)
      summary(silhouette(cutree(hc.ward.eu, k=x), dist.eu))$avg.width)
      max(asw)
      #[1] 0.2986852, druhe nejlepsi sestiklastrove reseni [5]
      0.282443579
```

První příkaz definuje proměnnou `asw`. Do ní postupně ukládá výsledky výpočtů průměrné šířky siluety všech variant vypočtených klastrů pro profily „číslopád“. Druhý příkaz `max(asw)` z těchto průměrných šířek siluet vybere tu nejvhodnější. V tomto případě jde o dvouklastrové řešení. Jelikož ale chci detailnější pohled, zjistím si i několik nejbližších dalších řešení. Relativně silné je také šestiklastrové řešení, které má průměrnou šířku siluety jen o setinu a půl menší, ale pro mě stále ještě relevantní (průměrná šířka = `0.259889004`, přibližně o tři a půl setiny menší) je i pětiklastrové. S ním budu dále pracovat, protože mě zajímá, jestli bude oněch pět klastrů (alespoň částečně) odpovídat mé pětistupňové hierarchii životnosti (viz „Makro-kategorie“ v Tabulce 2)

```
(ii)  asw.jenom.pady<-sapply(2:377, function(x)
      summary(silhouette(cutree(hc.ward.eu.jenom.pady, k=x),
      dist.eu.jenom.pady))$avg.width)
      max(asw.jenom.pady)
      #[1] 0.3802994, druhe a treti nejlepsi peti- a sestiklastrove
      reseni [4] 0.302617422 [5] 0.292777841
```

I v tomto případě první příkaz definuje proměnnou `asw.jenom.pady`, do níž postupně ukládá výsledky výpočtů průměrné šířky siluety všech variant vypočtených klastrů pro profily „pád“. Druhý příkaz `max(asw.jenom.pady)` z těchto průměrných šířek siluet vybere tu nejvhodnější. I tentokrát je nejvýraznějším řešením to o dvou klastrech – průměrnou šířku siluety má o desetinu větší než dvouklastrové řešení v položce (i). Znovu jsem si ovšem nechal vypsat další silná řešení a zjistil jsem, že druhé nejsilnější má 5 klastrů a průměrnou šíři siluety stále ještě větší než dvouklastrové řešení v předhozí položce. Opět s ohledem na hierarchii životnosti stanovenou v oddílu 2.2 si pro další postup vyberu právě toto řešení.

```
(iii) asw.2cat<-sapply(2:377, function(x)
      summary(silhouette(cutree(hc.ward.eu.2cat, k=x),
      dist.eu.2cat))$avg.width)
      max(asw.2cat)
      #[1] 0.3293232, druhe a treti nejlepsi tri- a ctyrklastrove reseni
      [2] 0.323134996 [3] 0.313628280
```

Zde provádím obdobný výpočet průměrné šířky siluety pro hierarchii klastrů gramatických profilů „2 kategorie“. Když jsem porovnal nejlépe strukturovaná řešení pro tuto hierarchii klastrů, zjistil jsem, že dvou-, tří- a čtyřklastrové řešení mají téměř stejně vysoké hodnoty. Vybral jsem tedy pro další postup řešení čtyřklastrové, neboť je vhodnější pro porovnání s mou hierarchií životnosti.

Porovnání dendrogramů

Když jsem pomocí výpočtu průměrných šířek siluety klastrových řešení ověřil, že výstupy mé klastrové analýzy jsou relativně dobře strukturované (pro některé typy gramatických profilů lépe, pro některé hůře – viz oddíl Průměrná šířka siluety), přišlo na řadu přímé porovnání klastrových řešení pro různé gramatické profily prostřednictvím jejich grafických znázornění, tj. dendrogramů. Nástroje pro tento úkon mi poskytl balíček dendextend (Galili 2015). Díky tomuto balíčku pro R jsem mohl vytvořit graf ukazující vztahy mezi dvěma klastrovými řešeními.

```
(i)   dend.normal<-as.dendrogram(hc.ward.eu)
```

```
(ii)  dend.pady<-as.dendrogram(hc.ward.eu.jenom.pady)

(iii) dend.2cat<-as.dendrogram(hc.ward.eu.2cat)
```

Prvním krokem bylo převedení klastrů (které se dají zobrazit jako dendrogramy) přímo na objekty typu dendrogram. To jsem provedl pro klastry (i) „číslopád“ (ii) „pád“ i (iii) „2 kategorie“ pomocí k tomu zvlášť určené funkce `as.dendrogram()` z příslušných proměnných.

Dále jsem pomocí funkce `tanglegram()` vytvořil tři dvojice zarovnaných dendrogramů s naznačenými vztahy, tj. čarami spojujícími stejné lemma.

```
(i)   tanglegram(dend.normal, dend.pady, lab.cex = 0.15, lwd = 1.0,
      common_subtrees_color_branches = TRUE, main_left = "číslopády",
      main_right = "pády")

(ii)  tanglegram(dend.normal, dend.2cat, lab.cex = 0.15, lwd = 1.0,
      common_subtrees_color_branches = TRUE, main_left = "číslopády",
      main_right = "2 proměnné")

(iii) tanglegram(dend.pady, dend.2cat, lab.cex = 0.15, lwd = 1.0,
      common_subtrees_color_branches = TRUE, main_left = "pády",
      main_right = "2 proměnné")
```

Tanglegram v (i) porovnává klastrové dendrogramy gramatických profilů „číslopád“ a „pád“ (viz Příloha 2, soubor tanglegram-normal-pady.pdf); tanglegram ve (ii) porovnává klastrové dendrogramy gramatických profilů „číslopád“ a „2 kategorie“ (viz Příloha 2, soubor tanglegram-normal-2cat.pdf); tanglegram ve (iii) porovnává klastrové dendrogramy gramatických profilů „pád“ a „2 kategorie“ (viz Příloha 2, soubor tanglegram-pady-2cat.pdf). Kvůli velkému množství dat bylo nutné zmenšit popisky na 15 % pomocí `lab.cex = 0.15`, jinak by se při zobrazení na běžných obrazovkách překrývaly. Také tloušťka spojovacích čar byla upravena z běžných 3,5 na 1,0, a to parametrem `lwd = 1.0`.

Tyto tanglegramy poskytly základní přehled o tom, jaké jsou rozdíly v uspořádání klastrů napříč řešeními, ale vzhledem k různě vypočítaným dendrogramům se jejich struktura zdá odlišnější, než ve skutečnosti je. Překonat tuto překážku umožňuje funkce `untangle()`, respektive její verzi `untangle_random_search()`, která nejdříve spojí dané dendrogramy do seznamu dendrogramů (datatyp dendlist) a následně rotuje jejich větve a listy tak, aby (i) se klastry obsahující stejné jednotky nacházely v tanglegramu co nejblíže sobě a (ii) aby byla zachována hierarchická struktura klastrů.

```
(i)   untangle1<-untangle_random_search(dend.normal, dend.pady, R=40,
      leaves_matching_method = "labels")

(ii)  untangle2<-untangle_random_search(dend.normal, dend.2cat, R=40,
      leaves_matching_method = "labels")

(iii) untangle3<-untangle_random_search(dend.pady, dend.2cat, R=40,
      leaves_matching_method = "labels")
```

Funkce prorotuje každou z dvojic dendrogramů tolikrát, kolik je hodnota parametru `R`, v mém případě tedy čtyřicetkrát. Algoritmus náhodně vyhledává stejné hodnoty v obou porovnávaných dendrogramech, aby pak mohly být spojeny čarami při vytváření dendrogramu. Vyhledávané hodnoty mohou být buďto (i) pořadí hodnot, nebo (ii) názvy listů. Jak lze vidět v kódu výše, ve všech případech jsem parametr `leaves_matching_metod` nastavil na možnost (ii), označenou jako „`labels`", neboť je pro mě relevantní zarovnat listy (koncové jednotky dendrogramů) podle jejich názvů, tedy lemmat.
Vytvořil jsem tedy nové tanglegramy z výše vytvořených dendlistů:

```
(i)   tanglegram(untangle1, lab.cex = 0.15, lwd = 1.0,
      common_subtrees_color_branches = TRUE, main_left = "číslopády",
      main_right = "pády", which = c(1L, 2L))

(ii)  tanglegram(untangle2, lab.cex = 0.15, lwd = 1.0,
      common_subtrees_color_branches = TRUE, main_left = "číslopády",
      main_right = "2 proměnné", which = c(1L, 2L))

(iii) tanglegram(untangle3, lab.cex = 0.15, lwd = 1.0,
      common_subtrees_color_branches = TRUE, main_left = "pády",
      main_right = "2 proměnné", which = c(1L, 2L))
```

Oproti předešlé tvorbě tanglegramů bylo ještě nutno pomocí nastavení parametru `which` specifikovat, které z dendrogramů v dendlistu mají být porovnány (v tomto případě oba, které každý dendlist obsahuje). Výsledné tanglegramy viz Příloha 2 soubory (i) untangled.cislopady-pady.pdf (ii) untangled.cislopady-2promenne.pdf (iii) untangled.pady-2promenne.pdf.

3.4.2 Náhodný les

V tomto oddílu se věnuji popisu tvorby podmíněných inferenčních stromů a náhodných lesů. Stejně jako v oddílu 3.4.1 vždy uvádím konkrétní kód použitý v programu R a vysvětluji, co daný úsek kódu provádí.

3.4.2.1 Příprava dat

Prvním krokem je příprava dat do podoby, ve které je lze zpracovat zvolenými metodami.

```
(i)   cutree.normal.list<-list(cutree(hc.ward.eu, k=5))

      cutree.normal.table<-do.call(cbind, cutree.normal.list)
```

Úryvek kódu (i) prvním příkazem ořezává dendrogram klastrové hierarchie gramatického profilu „číslopád“ ve výšce, ve které jsou data rozdělena právě na pět klastrů, což je množství zvolené s ohledem na průměrnou šířku siluety klastrů v daném klastrovém řešení (viz oddíl 3.4.1.3). Takto oříznutý dendrogram třídy hclust je převeden na seznam klastrů, které jsou označeny lemmaty, jež do nich patří. Druhý příkaz do proměnné `cutree.normal.table` třídy *table* ‚tabulka‘ ukládá právě vytvořený seznam jako druhý sloupec, zatímco první sloupec vyplňuje lemmaty, která předtím představovala pouze značky. Nyní tedy mám k dispozici tabulku obsahující seznam lemmat a informaci o tom, do kterého z pěti klastrů každé lemma patří.

```
(ii)  cutree.pady.list<-list(cutree(hc.ward.eu.jenom.pady, k=5))

      cutree.pady.table<-do.call(cbind, cutree.pady.list)
```

V úryvku kódu (ii) prvním příkazem ořezávám dendrogram klastrové hierarchie gramatického profilu „pád“ opět ve výšce, ve které jsou data rozdělena právě na pět klastrů (množství zohledňující průměrnou šířku siluety klastrů v daném klastrovém řešení; viz oddíl 3.4.1.3). Takto oříznutý dendrogram třídy hclust je stejným způsobem jako předešlý (i) převeden na seznam klastrů, které jsou označeny lemmaty, jež do nich patří. Druhý příkaz do proměnné `cutree.pady.table` ukládá právě vytvořený seznam jako druhý sloupec, zatímco první sloupec vyplňuje značkami klastrů ze seznamu – lemmaty. Výsledkem je znovu tabulka obsahující seznam lemmat a informaci o jejich zařazení do jednoho z pěti klastrů – od předchozí se liší typem gramatických profilů, které byly klastrovány.

```
(iii) cutree.2cat.list<-list(cutree(hc.ward.eu.2cat, k=4))

      cutree.2cat.table<-do.call(cbind, cutree.2cat.list)
```

Kódem (iii) replikuji výše popsaný postup s tím rozdílem, že klastry jsou složené z gramatických profilů typu „2 kategorie“. S ohledem na průměrnou šířku siluety (viz 3.4.1.3) je však dendrogram oříznut na čtyři klastry. Výsledkem je jako v kódech (i) a (ii) tabulka

lemmat s přiřazeným číslem klastru, do něhož jsou příslušné gramatické profily „2 kategorie“ zařazeny.

Nyní si vytvořené tabulky uložím do textových souborů, abych je mohl snadněji zpracovat.

```
(i)   write.table(cutree.normal.table, file="cutree.normal.txt",
      row.names=TRUE, col.names=TRUE, sep="\t")

(ii)  write.table(cutree.pady.table, file="cutree.pady.txt",
      row.names=TRUE, col.names=TRUE, sep="\t")

(iii) write.table(cutree.2cat.table, file="cutree.2cat.txt",
      row.names=TRUE, col.names=TRUE, sep="\t")
```

V předešlém kroku vytvořené tabulky jsem později sloučil tak, aby ke každému z lemmat byla přiřazena informace o klastru, do něhož je zařazeno v každém z řešení (tj. jeden sloupec pro každý z typů gramatických profilů popsaných v oddíllu 3.3.4.2). Do dalších sloupců jsem vložil informace o zařazení každého z lemmat podle (iv) sémantické kategorie (v) makro-kategorie a (vi) životnosti ve smyslu Tabulky 2 (viz oddíl 3.3.3.1).

Poté jsem mohl tabulku načíst zpět do R a začít ji zpracovávat.

```
pralesy.data <- read.table (file.choose(), header = T, row.names = 1,
sep="\t", encoding = "UTF-8")

pralesy.data$kategorie<-as.factor(pralesy.data$kategorie)

pralesy.data$X2cat.4clust<-as.factor(pralesy.data$X2cat.4clust)

pralesy.data$cislopady.5clust<-as.factor(pralesy.data$cislopady.5clust)

pralesy.data$pady.5clust<-as.factor(pralesy.data$pady.5clust)
```

První ze zde uvedených příkazů slouží k načtení tabulky a kromě proměnné, do níž tabulku ukládá, se neliší od předchozích načtení jiných tabulek.

Další příkazy převádějí obsah numerických sloupců tabulky na faktorové (kategorické). Důvodem je způsob, jakým klíčové funkce `ctree()` a `cforest()` zacházejí s numerickými proměnnými (viz oddíl 2.4). Obsah sloupců `kategorie`, `X2cat.4clust`, `cislopady.5clust` a `pady.5clust` je sice číselný, ale tato čísla jsou jen arbitrárními pojmenováními, která usnadňují orientaci v datech, takže by bylo chybné je zpracovávat jako číselné kontinuum, jež by bylo použitými funkcemi rozděleno podle aritmetického průměru. Ostatní sloupce tabulky jsou jako faktorové nastaveny automaticky, takže není nutné je pomocí funkce `as.factor()` převádět.

3.4.2.2 Tvorba podmíněných inferenčních stromů

S připravenou univerzální tabulkou, tj. takovou, která obsahuje relevantní charakteristiky pro všechny tři typy gramatických profilů, mohu přikročit k ověření klastrových řešení popsaných v oddílu 3.4.1 pomocí metody podmíněných inferenčních stromů, které jsem představil v sekci 2.4.

```
set.seed(6)
```

Tento jednoduchý příkaz zadává takzvaný *random seed* ‚náhodné semínko', což je číslo, kterým se inicializuje generátor pseudo-náhodných čísel, jenž v případě funkcí `ctree()` a `cforest()` zajišťuje náhodnost dělení dat na větve. Zadání seedu není pro spuštění funkce vyžadováno, neboť program R dokáže vybrat seed automaticky. Každý algoritmus pro generování pseudo-náhodných čísel ovšem dojde ke stejným výsledkům pokaždé, když použije stejný seed. Tímto způsobem je zajištěno, že výsledky zde uvedených výpočtů bude možno kdykoliv replikovat. Výběr seedu ovšem neovlivňuje náhodnost vygenerovaných čísel (výběr seedu je arbitrární), a tak je možné pro hospodárnost kódu použít jeden seed pro vygenerování libovolného množství stromů či lesů.
Nyní tedy můžu přikročit k samotné tvorbě stromů.

(i)
```
strom.pady.5clust<-ctree(pady.5clust ~ kategorie + makrokategorie
+ zivotnost, data=pralesy.data)
```

Tímto příkazem definuji proměnnou `strom.pady.5clust`, do níž ukládám binární strom. Parametrem `data` určuji, že strom bude vytvořen na základě datasetu `pralesy.data`, tedy na základě tabulky vytvořené v oddílu 3.4.2.1 Závislou proměnnou v procesu tvorby stromu je `pady.5clust`, obsahující čísla klastrů gramatických profilů typu „pád". Naopak nezávislými proměnnými (také prediktory) jsou tři různě hrubé stupnice životnosti představené v sekci 3.3.3.1, označené jako `kategorie`, `makrokategorie` a `zivotnost`. Snažím se tedy v souladu s výzkumnými otázkami (viz 3.1) otestovat, zda se sledovaná lemmata shlukují do klastrů podle životnosti, což by znamenalo, že životnost opravdu má tendenci projevovat se v češtině rozdílnou relativní frekvenční distribucí pádů. Hladina významnosti je pro funkce ctree a cforest() automaticky stanovena na 0,05. Všechny vzniklé větve stromu jsou tedy alespoň na 95 % statisticky významné.

(ii)
```
strom.cislopady.5clust<-ctree(cislopady.5clust ~ kategorie +
makrokategorie + zivotnost, data=pralesy.data)
```

Dalším příkazem definuji proměnnou `strom.cislopady.5clust`, do níž ukládám druhý binární strom. Jeho závislou proměnou je `cislopady.5clust`, obsahující čísla klastrů gramatických profilů typu „číslopád“. Prediktory a podmínky tvorby stromu se od (i) neliší.

```
(iii) strom.X2cat.4clust<-ctree(X2cat.4clust ~ kategorie +
      makrokategorie + zivotnost, data=pralesy.data)
```

Jako poslední vytvářím ze stejného datasetu strom `strom.X2cat.4clust`, který ověřuje vliv tří zmíněných škál životnosti na frekvenční distribuci sledovaných lemmat, respektive to, jak se shlukují jejich gramatické profily typu „2 kategorie“. Závislou proměnnou tohoto stromu je `X2cat.4clust`, zatímco nezávislé proměnné a podmínky tvorby stromu zůstávají stejné jako v (i) a (ii).

```
(i)   plot(strom.cislopady.5clust)

(ii)  plot(strom.pady.5clust)

(iii) plot(strom.X2cat.4clust)
```

Tyto tři příkazy slouží k vygenerování grafického znázornění tří právě popsaných stromů. (i) ukazuje možný vliv životnosti na rozřazení gramatických profilů „číslopád“ do pěti klastrů, (ii) ukazuje možný vliv životnosti na rozřazení gramatických profilů „pád“ taktéž do pěti klastrů a (iii) ukazuje vliv životnosti na rozřazení gramatických profilů "2 kategorie“ do čtyř klastrů. Výsledky prezentuji v podkapitole 4.2, grafy lze najít v Příloze 2 (soubory (i) strom.cislopady.5c.pdf (ii) strom.pady.5c.pdf (iii) strom.X2cat.4c.pdf).

3.4.2.3 Tvorba náhodných lesů

Nyní pomocí náhodných lesů, tedy sad mnohokrát přepočítaných podmíněných inferenčních stromů, změřím důležitost každého z prediktorů.

Les pro klastrové řešení GP typu „číslopád“

```
set.seed(6)
```

Stejně jako při tvorbě stromu je pro opakovatelnost získaných výsledků nutné zadat random seed. Jak, jsem vysvětlil v oddílu 3.4.2.2, volba seedu je arbitrární, a tak není důvod, abych nezvolil opět číslo 6.

```
les.cislopady.5clust<-cforest(cislopady.5clust ~ kategorie + makrokategorie
+ zivotnost, data=pralesy.data, controls = cforest_unbiased(ntree=400,
mtry=2))
```

Nejdříve vytvářím samotný les. Konstruktor lesa (v tomto případě pro gramatické profily typu „číslopády“ rozdělené do pěti klastrů) vypadá podobně jako konstruktor stromu uvedený v oddílu 3.4.2.2, navíc ovšem obsahuje parametr `controls`, kterým lze nastavit podmínky vytvoření lesa. Pro tuto práci jsem zvolil nezaujatý způsob tvorby lesa, který bude obsahovat 400 stromů. Parametr `mtry` určuje počet proměnných náhodně vzorkovaných v každém rozhodovacím uzlu. Obvykle se pro tento parametr používá hodnota odpovídající druhé odmocnině počtu nezávislých proměnných použitých pro tvorbu lesa (Levshina 2015: 297), což je v tomto případě po zaokrouhlení na celá čísla rovno dvěma.

```
les.cislopady.5clust.varimp<-varimp(les.cislopady.5clust, conditional =
TRUE)
```

V tomto kroku počítám indexy důležitosti proměnných pro `les.cislopady.5clust`. Jelikož jde o les podmíněných inferenčních stromů, parametr `conditional` ‚podmíněný‘ je nastaven na `TRUE` ‚pravda‘.

```
round(les.cislopady.5clust.varimp, 3)
```

Tento jednoduchý příkaz slouží k zaokrouhlení indexů důležitosti proměnných pro `les.cislopady.5clust` na tři desetinná místa. Provedeno za účelem zlepšení přehlednosti grafu, jenž vytvářím v následujícím kroku.

```
dotchart(sort(les.cislopady.5clust.varimp), main="Podminena dulezitost
promennych pro GP cislopady v peti klastrech")
```

Tento úryvek kódu vygeneruje `dotchart` ‚bodový graf‘ znázorňující důležitost proměnných pro daný les.

Les pro klastrové řešení GP typu „pád“

```
set.seed(6)
```

Random seed pro přehlednost nastavuji opět na 6.

```
les.pady.5clust<-cforest(pady.5clust ~ kategorie + makrokategorie +
zivotnost, data=pralesy.data, controls = cforest_unbiased(ntree=400,
mtry=2))
```

Tímto příkazem vytvářím samotný les, v tomto případě pro gramatické profily typu „pády“ rozdělené do pěti klastrů. Kvůli důslednosti ponechávám počet stromů (400), a protože se nemění počet nezávislých proměnných, hodnotu parametru `mtry` nastavuji opět na 2.

```
les.pady.5clust.varimp<-varimp(les.pady.5clust, conditional = TRUE)
```

Zde počítám indexy důležitosti proměnných pro `les.pady.5clust`. Hodnota parametru `conditional` ‚podmíněný' je opět nastavena na `TRUE` ‚pravda'.

```
round(les.pady.5clust.varimp, 3)
```

Tento jednoduchý příkaz slouží k zaokrouhlení indexů důležitosti proměnných pro `les.pady.5clust` na tři desetinná místa.

```
dotchart(sort(les.pady.5clust.varimp), main="Podminena dulezitost
promennych pro GP pad v peti klastrech")
```

Tento úryvek kódu vygeneruje `dotchart` ‚bodový graf' znázorňující důležitost proměnných pro daný les.

Les pro klastrové řešení GP typu „2 kategorie"

```
set.seed(6)
```

Tento příkaz nastavuje hodnotu random seedu, který ani tentokrát neměním.

```
les.X2cat.4clust<-cforest(X2cat.4clust ~ kategorie + makrokategorie +
zivotnost, data=pralesy.data, controls = cforest_unbiased(ntree=400,
mtry=2))
```

Tímto příkazem vytvářím samotný les, v tomto případě pro gramatické profily typu „2 kategorie" rozdělené do čtyř klastrů. Kvůli důslednosti ponechávám počet stromů (400), a protože se nemění počet nezávislých proměnných, hodnotu parametru `mtry` nastavuji opět na 2.

```
les.X2cat.4clust.varimp<-varimp(les.X2cat.4clust, conditional = TRUE)
```

V tomto úseku kódu počítám indexy důležitosti proměnných pro `les.X2cat.4clust`. Hodnota parametru `conditional` ‚podmíněný' je opět nastavena na `TRUE` ‚pravda'.

```
round(les.X2cat.4clust.varimp, 3)
```

Indexy důležitosti proměnných opět pro přehlednost zaokrouhluji na tři desetinná místa.

```
dotchart(sort(les.X2cat.4clust.varimp), main="Podminena dulezitost
promennych pro GP 2 kategorie ve ctyrech klastrech")
```

Tento úryvek kódu vygeneruje `dotchart` ‚bodový graf' znázorňující důležitost proměnných pro daný les.

4. Výsledky

Tato kapitola je věnována prezentaci výsledků, které jsem získal postupem komentovaným v kapitole 3.

4.1 Výsledky klastrové analýzy

Primárním způsobem analýzy byla hierarchická klastrová analýza jejíž výsledky popisují oddíly 4.1.1 a 4.1.2.

4.1.1 Průměrná šířka siluety

Průměrná šířka siluety klastrů nabývá hodnot od 1 do -1 a říká, jak dobře je dané klastrové řešení strukturované (čím vyšší hodnota, tím lepší strukturovanost řešení).

Klastrové analýzy všech tří typů gramatických profilů ukazují, že vybraný dataset je strukturovaný. Nejkvalitnější strukturu lze nalézt u dvouklastrového řešení pro gramatické profily typu „pád" – průměrná šířka siluety v tomto řešení dosahuje hodnoty 0.3802994. Pětiklastrové řešení použité pro další srovnávání dosahuje hodnoty průměrné šířky siluety 0.302617422, což také ukazuje na jistou míru strukturovanosti. Zajímavým případem je hierarchie klastrů gramatických profilů typu „2 kategorie", která má téměř stejné hodnoty průměrné šířky siluety napříč řešeními dvouklastrovým (0.3293232), tříklastrovým (0.323134996) a čtyřklastrovým(0.313628280), které navíc ukazují poměrně značnou míru strukturovanosti. Nejhůř z tohoto testu strukturovanosti vyšla klastrová řešení gramatických profilů „číslopád". Ani jedno z řešení v tomto případě nepřesáhlo hranici 0,3, přičemž pouze prvních pět klastrových řešení dosáhlo hodnoty průměrné šířky siluety alespoň 0,25, tedy hranice, pod níž podle Kaufman & Rousseuw (1990) již data nebývají dobře strukturovaná.

4.1.2 Tanglegramy

Vizuálním porovnáním tanglegramů (viz Příloha 2 soubory (i) untangled.cislopady-pady.pdf (ii) untangled.cislopady-2promenne.pdf (iii) untangled.pady-2promenne.pdf) jsem zjistil, že napříč všemi třemi řešeními existují velmi stabilní klastry, jako příklad za všechny uvedu dobře patrné klastry toponym (skupina 14) či lidských proprií (skupina 1).

Statistické porovnání výsledků třech klastrových analýz z oddílu 3.4.1 je popsáno v následující podkapitole (4.2).

4.1.3 Kontingenční tabulky

V tomto úseku postupně prezentuji kontingenční tabulky, které znázorňují distribuci lemmat jednotlivých kategorií napříč klastry daného řešení. Za pomoci těchto kontingenčních tabulek

se pokusím jednotlivé klastry v každém řešení charakterizovat s ohledem na obsažené kategorie.

4.1.3.1 Distribuce sémantických kategorií v klastrech GP „číslopád“

První klastr v tomto řešení obsahuje zejména životná lemmata, řadí se do něj také zájmena. Proti očekávání se v tomto klastru životných lemmat objevuje také většina abstraktních lemmat (sk. 17).

Druhý klastr obsahuje jen málo lemmat napříč celou škálou životnosti, chybí v něm pouze zástupce zájmen, proprií, toponym a skupiny 5 (pracovně označena jako relační lidská pojmenování).

Třetímu klastru dominují propria, která jsou v něm shluknuta všechna. Klastr doplňuje několik lemmat lidských a životných ne-lidských.

Ve čtvrtém a pátém klastru se až na několik výjimek seskupila lemmata neživotná.

cislo.klastru	1	2	3	4	5
kat99	3	0	0	0	0
kat1	0	0	30	0	0
kat2	10	2	15	0	0
kat3	13	3	9	0	0
kat4	15	2	3	0	0
kat5	4	0	3	1	0
kat6	19	3	7	1	0
kat7	18	3	5	0	1
kat8	20	4	2	1	0
kat9	3	1	0	21	4
kat12	0	7	0	11	6
kat13	1	1	0	5	16
kat14	2	0	0	20	1
kat15	0	0	0	1	0
kat16	0	3	0	10	4
kat17	11	1	1	4	7
kat18	7	4	0	10	2
kat20	0	3	0	0	14

Tabulka 5: Kontingenční tabulka distribuce sémantických kategorií v klastrech GP „číslopád“

4.1.3.2 Distribuce sémantických kategorií v klastrech GP „pád“

V tomto řešení je první klastr až na zájmena výhradně neživotný, největší zastoupení má skupina 20 (oblečení).

Druhý klastr zahrnuje životná lemmata vyjma proprií, ačkoli například profese jsou také zastoupeny minimálně. Pozornost je ovšem třeba věnovat šesti rostlinám a 10 abstraktům, která mají v tomto GP frekvenční distribuci spíše jako životná.

Třetí klastr zaujme převahou proprií a profesních lemmat, ale také tím, že neobsahuje jediné neživotné lemma.

Čtvrtý klastr seskupuje některá zbylá lidská i ne-lidská životná substantiva, ovšem daleko víc jich v něm je neživotných.

Klastr pátý je složen téměř pouze z toponym.

cislo.klastru	1	2	3	4	5
kat99	3	0	0	0	0
kat1	0	0	30	0	0
kat2	0	4	23	0	0
kat3	0	13	9	3	0
kat4	0	18	2	0	0
kat5	0	3	4	1	0
kat6	0	14	13	2	1
kat7	0	19	4	4	0
kat8	0	17	9	1	0
kat9	2	0	0	26	1
kat12	9	0	0	14	1
kat13	8	0	0	13	2
kat14	1	2	0	4	16
kat15	0	0	0	1	0
kat16	4	0	0	10	3
kat17	3	10	0	11	0
kat18	1	6	0	16	0
kat20	16	0	0	1	0

Tabulka 6: Kontingenční tabulka distribuce sémantických kategorií v klastrech GP „pád“

4.1.3.3 Distribuce sémantických kategorií v klastrech GP „2 kategorie“

První klastr tohoto řešení obsahuje výrazně více výrazů neživotných než životných, pozoruhodným jevem je opět výskyt zájmen mezi těmito neživotnými lemmaty.

Druhý klastr kromě životných obsahuje také několik abstrakt a rostlin. Na tento jev jsem narazil (a upozornil) již u řešení pro GP „pád“ (viz sekce 4.1.3.2). Pozoruhodně velké zastoupení mají profese (sk. 2), příbuzenské termíny (sk. 4) a všechna životná lemmata ne-lidská.

Ve třetím klastru výrazně převládají lidská lemmata, největší zastoupení mají ovšem propria.

Čtvrtý klastr je chudý na lemmata kterékoliv skupiny, nejvíce zástupců mají části těla (sk. 12). Z hlediska životnosti jde jinak o nevýrazný klastr obsahující několik zbylých lemmat z blízkosti obou pólů škály.

cislo.klastru	1	2	3	4
kat99	3	0	0	0
kat1	0	0	30	0
kat2	0	15	12	0
kat3	7	6	9	3
kat4	6	11	3	0
kat5	2	5	1	0
kat6	2	21	5	2
kat7	3	18	3	3
kat8	1	22	1	3
kat9	28	0	0	1
kat12	16	0	0	8
kat13	22	0	0	1
kat14	23	0	0	0
kat15	1	0	0	0
kat16	14	0	0	3
kat17	19	4	1	0
kat18	17	2	0	4
kat20	14	0	0	3

Tabulka 7: Kontingenční tabulka distribuce sémantických kategorií v klastrech GP „2 kategorie“

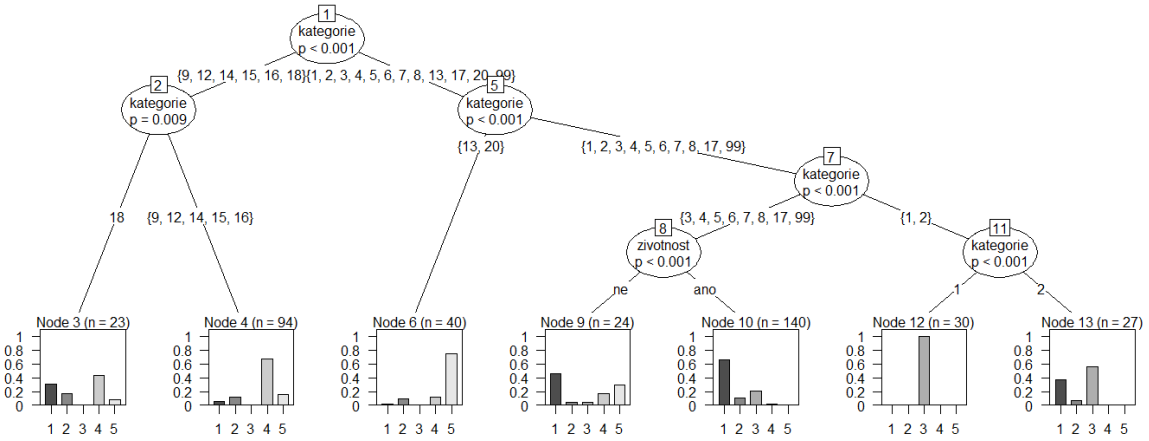
4.2 Výsledky modelů náhodných lesů

Náhodné lesy binárních stromů (podmíněných inferenčních stromů) posloužily jako statistické ověření hypotézy, že životnost ovlivnila složení klastrů tří různých typů gramatických profilů (viz oddíl 2.1), které představují relativní frekvenční distribuci flexe 378 vybraných českých nominálních lemmat.

4.2.1 Podmíněné inferenční stromy

V sekci 3.4.2.2 jsem vytvořil po jednom podmíněném inferenčním stromu pro druhé či třetí nejlépe strukturované řešení klastrové hierarchie pro každý z představených typů gramatických profilů. Cílem bylo zjistit, jak statisticky významná ona rozdělení jsou, a tedy jestli lze s 95% jistotou říci, zda se životnost projevuje v užívání češtiny prostřednictvím frekvenční distribuce nominální flexe.

4.2.1.1 Podmíněný inferenční strom klastrů GP „číslopád“

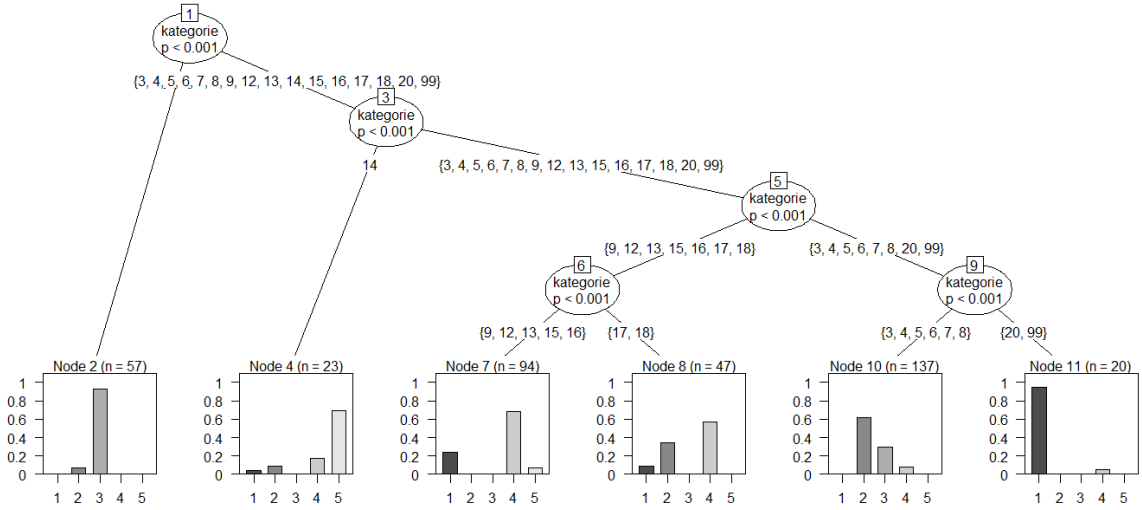


Obrázek 1: Graf podmíněného inferenčního stromu klastrů GP „číslopád“

Graf tohoto podmínečného inferenčního stromu (Obrázek 1; Příloha 2, soubor strom.cislopady.5c.pdf) ukazuje vydělení několika neživotných skupin. Jsou jimi: slovesná substantiva, části těla, toponyma, temporální výrazy, rostliny (uzel 2). Rostliny jsou odděleny do uzlu 3, lemmata z této skupiny jsou ovšem rozprostřena po všech klastrech s výjimkou klastru 3. Zajímavější ovšem je, že zbylé skupiny (uzel 4) jsou ze 70 % koncentrovány v klastru 4. Přesuneme-li se k uzlu šest, vidíme, že téměř 80 % lemmat ze skupin „oblečení“ a „věci denní potřeby“ se nachází v klastru 5, což znamená, že jsou v češtině užívána podobně – alespoň pokud jde o distribuci ve 12 hodnotách sloučené kategorie čísla a pádu. Pro vydělení uzlu 6 je p-hodnota menší než 0,01. Na základě klastrové analýzy tedy lze tedy pozorovat podobné užívání kategorií 13 a 20. Uzel 8 s výjimkou skupiny zbylých abstrakt, která poté tvoří samostatný uzel 9, zastupuje životná lemmata, ovšem bez lidských proprií a profesí. Uzel 10 není pro interpretaci příliš vhodný, neboť přesto, že slučuje většinu životných lemmat, většina jich spadá do klastru 5 společně s rostlinami a zbylými abstrakty. Za povšimnutí ovšem stojí asi 20 % těchto lemmat, která se objevila v klastru 3, v němž – jak ukazují uzly 13 a zejména 12 – převládají lidé. Uzel 12 zahrnuje výhradně lidská propria, která se všechna objevují v klastru 3. Profese (uzel 13) jsou podle tohoto modelu užívání ve více než polovině případů podobná vlastním jménům, ta ostatní se nejvíce shlukují v klastru 1 s rostlinami, životnými substantivy, zájmeny a abstrakty. Podle složení klastru 1 se lze domnívat, že téměř polovina rostlinných a abstraktních substantiv má distribuci sloučeného pádu a čísla jako lemmata životná.

Nejjasnějším výstupem tohoto stromového modelu tedy je, že se profese ve většině případů podobají lidským propriím.

4.2.1.2 Podmíněný inferenční strom klastrů GP „pád“

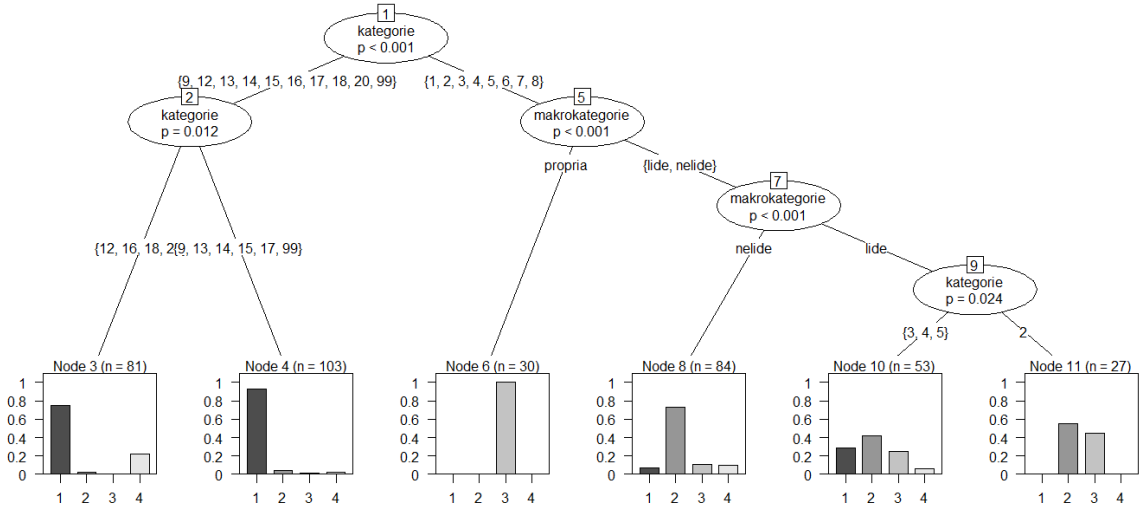


Obrázek 2: Graf podmíněného inferenčního stromu klastrů GP „pád“

Z grafu tohoto modelu (Obrázek 2; Příloha 2, soubor strom.pady.5c.pdf) odečítám vydělení proprií a profesí hned v prvním kroku. Všechny uzly se oddělují při p<0,001. Uzel 2 ukazuje, že propria a profese se vyskytují téměř výhradně v klastru 3, který sdílejí přibližně se třetinou všech ostatních životných substantiv, což znamená, že lemmata skupin 1 a 2 se od ostatních svým užíváním liší ať už započítáváme gramatické číslo (viz oddíl 4.2.1.1), nebo ho nerozlišujeme (Obrázek 2). Uzel 4 se skládá pouze z toponym, z nichž přibližně 70 % tvoří klastr 5, který je pro tuto skupinu téměř (viz uzel 10) exkluzivní. Uzel 6 zahrnuje zbylá neživotná lemmat s výjimkou oblečení. Přestože se z něj vydělují zbylá abstrakta a rostliny (uzel 8), většina uzlu 6 spadá do klastru 4, který jehož obsah je přibližně z 90 % neživotný. Signifikantní při p<0,001 je ovšem také vydělení zájmen a oblečení (uzel 11). Tento uzel spadá téměř výhradně do klastru 1, který sdílejí s částí neživotných substantiv. Uzel 10 ukazuje, že ostatní životná substantiva (tj. lidská mimo proprií a profesí, zvířecí a antropomorfní) se shlukují převážně v klastru 2, jejž sdílejí s částí uzlu 8 (abstrakta a rostliny). Asi třetina životného uzlu 10 spadá do klastru 3.

Stejně jako ve stromovém modelu pro GP „číslopád“ (viz oddíl 4.2.1.1) i zde vidíme jasnou tendenci profesních a propriálních lemmat shlukovat se dohromady, společně s částí ostatních životných. 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.

4.2.1.3 Podmíněný inferenční strom klastrů GP „2 kategorie“

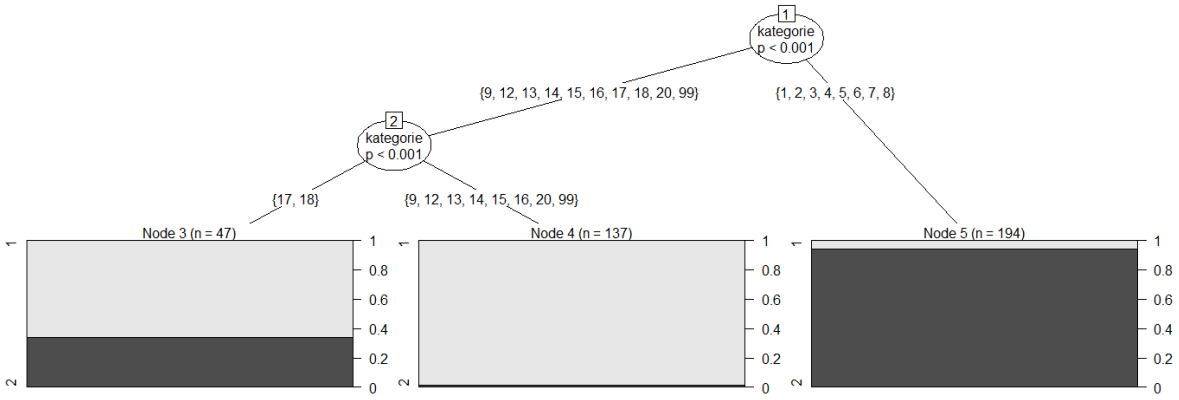


Obrázek 3: Graf podmíněného inferenčního stromu klastrů GP „2 kategorie“

Grafické znázornění tohoto modelu (Obrázek 3; Příloha 2, soubor strom.x2cat.4c.pdf) představuje jednoznačnější a interpretačně nejpřesvědčivější z představených možností. První krok rozděluje data na neživotná lemmata a zájmena (uzel 2) a lemmata životná (uzel 5) při $p<0{,}001$. Zájmena, abstrakta, toponyma, věci denní potřeby a slovesná substantiva (uzel 4) se významně ($p<0{,}05$) shlukují v klastru 4, zatímco části těla, temporální výrazy, rostliny a oblečení (uzel 3) mají značné zastoupení také v klastru 4, což je od uzlu 4 významně odlišuje. Při pohledu na dělení životného uzlu 5 lze pozorovat opět jednoznačné vydělení proprií (uzel 6) do klastru 3, který sdílí především s některými lidskými substantivy (uzel 9). V klastru 2 převládají ne-lidská životná substantiva doplněná o přibližně 50 % profesí a přibližně 40 % lidských substantiv skupin 3, 4 a 5. Tyto tři skupiny (uzel 10) jsou téměř rovnoměrně rozděleny do klastrů 1, 2 a 3, ovšem jisté zastoupení vidíme i v klastru 4, v němž se nacházejí převážně neživotná lemmata. Pro profese, v předchozích modelech řazené k propriím, je zde charakteristické rozdělení do klastrů 2 a 3 v poměru přibližně 55:45, většina se jich tedy shlukuje s ostatními životnými (mimo zájmena a propria).

Tento model přinesl takřka jednoznačné potvrzení pro rozdělení na životná a neživotná lemmata, zároveň ukazuje podobné tendence jako oba výše prezentované modely (např. odlišné užívání proprií a profesí, nepředpokládané zařazení zájmen atd.).

4.2.1.4 Podmíněný inferenční strom „Test životnosti“



Obrázek 4: Graf podmíněného inferenčního stromu dvou klastrů GP „pád“

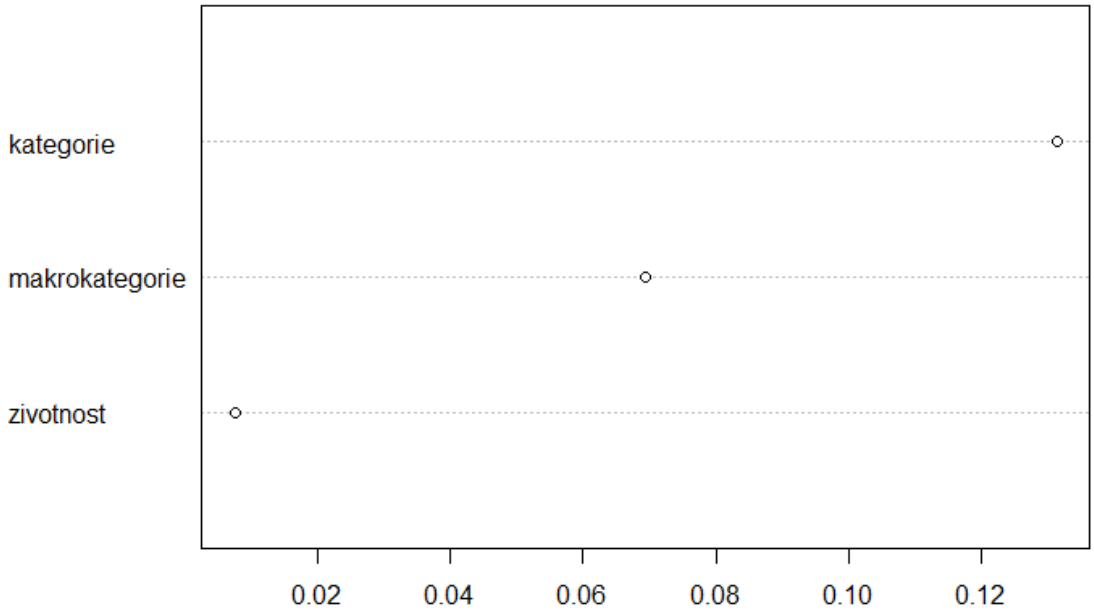
Strom v tomto grafu (Obrázek 4) zobrazuje test binární životnosti. Tmavá plocha představuje klastr 1, světlá plocha klastr 2. Typ GP „pád“ byl pro tento test vybrán, protože jeho dvouklastrové řešení vykazovalo jednoznačně nejvyšší průměrnou šířku siluet klastrů (0.3802994; porovnání viz 4.1.1). Uzel 5 obsahuje všechna životná lemmata s výjimkou zájmenných s tím, že téměř 100 % životných se nachází v klastru 1. Uzel 3 ukazuje zvláštní distribuci rostlinných a abstraktních lemmat, kterou jsem pozoroval u tohoto typu GP i v jemnějším rozdělení na 5 klastrů (viz 4.2.1.2). Uzel 4 je téměř úplným opakem uzlu 5 – obsahuje samá neživotná lemmata, zařazená téměř ze 100 % do klastru 2. Do uzlu 4 se řadí ovšem také zájmena, jejichž „neživotnému“ užívání se věnuji v oddílu 4.2.1.5. Při $p<0,001$ lze říci, že životnost skutečně rozděluje lemmata na dvě distinktivní skupiny prostřednictvím jejich relativní frekvenční distribuce.

4.2.1.5 Shrnutí

Poznatky ze stromových modelů pro všechny tři typy gramatických profilů naznačují, že je oprávněné se domnívat, že životnost ovlivňuje frekvenční distribuce nominálních lemmat v češtině. V oddílu 4.2.1.4 jsem pak ukázal, že neživotná lemmata se shlukují do jednoho klastru a životná do druhého. Dalším argumentem z jemnějšího klastrového řešení je v tomto ohledu rozdělení na uzly 2 a 5 ve stromu pro klastry GP „2 kategorie“. Proti přijetí alternativní hypotézy hovoří zařazení zájmen, která by se podle východisek prezentovaných v oddílu 2.2 měla řadit na životný konec škály životnosti, ovšem toto empiricky zjištěné zařazení, lze spekulovat, by se dalo vysvětlit tendencí češtiny elidovat zájmena ve funkci subjektu (čeština je tzv. pro-drop jazyk). Tím by bylo možné vysvětlit odlišnost v užívání zájmen a životných substantiv, jelikož zájmena tak mají mnohem méně výskytů v nominativu – pádu subjektu.

4.2.2 Náhodný les

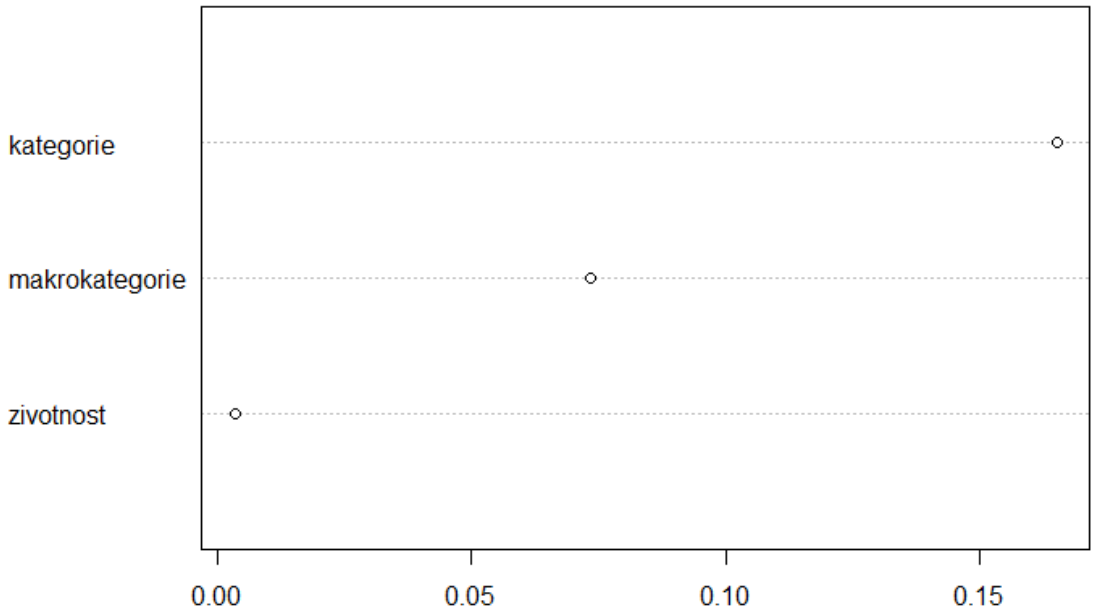
Pomocí náhodných lesů jsem spočítal, jak důležitý je každý prediktor pro rozdělení dat do klastrů.



Obrázek 5: Bodový graf důležitosti prediktorů pro rozdělení dat do klastrů GP „číslopád"

Konkrétní hodnoty indexu důležitosti prediktorů pro klastry GP „číslopád“ uvádím zde:

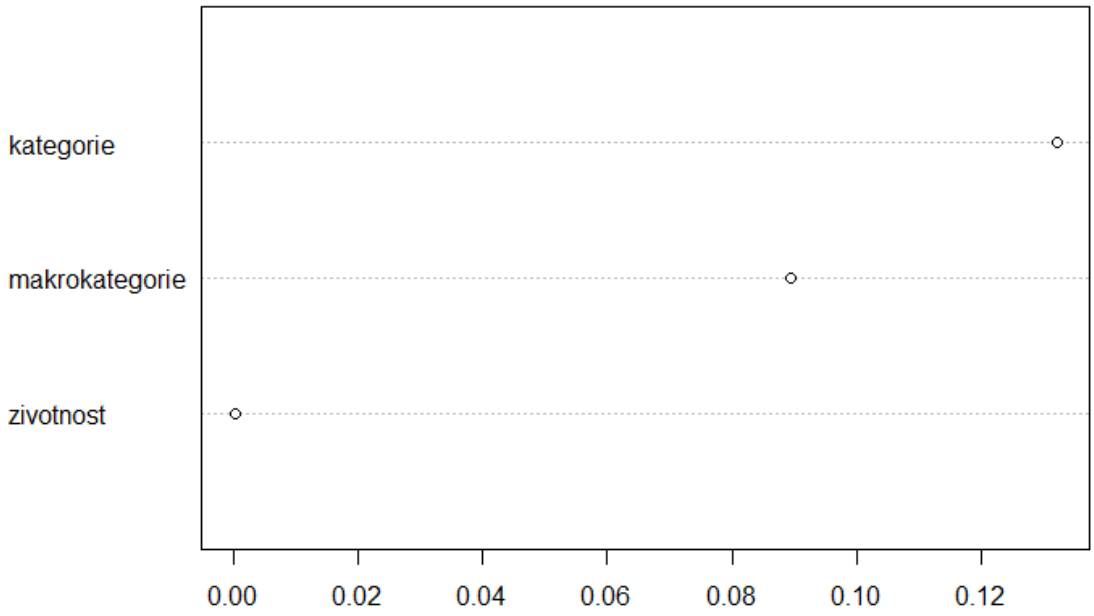
```
#   kategorie makrokategorie      zivotnost

 #       0.131          0.069          0.008
```



Obrázek 6: Bodový graf důležitosti prediktorů pro rozdělení dat do klastrů GP „pád"

Konkrétní hodnoty indexu důležitosti prediktorů pro klastry GP „pád“ uvádím zde:

```
#    kategorie makrokategorie      zivotnost
 #        0.165          0.073          0.004
```



Obrázek 7: Bodový graf důležitosti prediktorů pro rozdělení dat do klastrů GP „2 kategorie"

Konkrétní hodnoty indexu důležitosti prediktorů pro klastry GP „2 kategorie“ uvádím zde:

```
#    kategorie makrokategorie      zivotnost
 #        0.132          0.089          0.000
```

Z grafů (Obrázek 5, Obrázek 6, Obrázek 7) lze vyčíst, že jednoznačně nejvyšší důležitosti dosahuje u všech tří modelů prediktor `kategorie`. Tento fakt znamená, že anotace na více stupňů životnosti než jen na binární ano/ne byla dobrý krokem, neboť tato podrobnější anotace umožňuje zaznamenat jemnější rozdíly v užívání sledovaných lemmat.

Pro dělení dat je do jisté míry relevantní i proměnná `makrokategorie`, ale její význam není tak velký, s výjimkou gramatických pádů typu „2 kategorie“. Na grafu důležitosti prediktorů pro tento typ gramatických pádů (Obrázek 7) je vidět, že důležitost obou vedoucích prediktorů se liší jen málo, zatímco prediktor `zivotnost` zde vliv nemá.

5. Závěr

Tématem této práce bylo hledání pojítka mezi životností a frekvenční distribucí nominální flexe v češtině. Přístup jsem zvolil kvantitativní, vycházel jsem z poznatků lingvistického výzkumu založeného na užívání jazyka (usage-based linguistics). Využil jsem několika druhů explorativních statistických nástrojů, klíčovým byla hierarchická klastrová analýza, jejíž výstupy jsem následně vyhodnotil pomocí podmíněných inferenčních stromů.

Lemmata jsem sledoval prostřednictvím gramatických profilů, tedy souborů relativní frekvenční distribuce pro každé lemma. Bez ohledu na použitý typ gramatického profilu se lemmata shlukovala podle životnosti.

Nejlepších výsledků jsem však dosáhl použitím gramatického profilu, který nerozlišoval gramatické číslo. Klastrové řešení s těmito profily oříznuté na dva velké klastry, které vykazovaly vysokou míru strukturovanosti, rozdělilo lemmata na životná a neživotná na hladině významnosti $p<0,001$, čímž byla jednoznačně vyvrácena nulová hypotéza a současně potvrzena hypotéza alternativní, která říká, že existuje významný vztah mezi hierarchií životnosti a relativní frekvenční distribucí nominální flexe v češtině.

Zároveň se díky užití náhodných lesů (random forests) podařila ukázat užitečnost jemnějšího dělení škály životnosti na množství specifičtějších stupňů.

Hierarchie životnosti nebo také nominální hierarchie je však v českém jazyce téma otevřené dalšímu zkoumání. Výzkum by se nyní mohl ubírat například směrem k behaviorálnímu profilování sloves, jaké bylo provedeno například na ruštině či angličtině, pozornost je vhodné věnovat také interakci škál prominence, které nominální hierarchii ovlivňují.

Na závěr je pak vhodné vypíchnout důležitost usage-based přístupů v lingvistice, na nichž byla tato práce založena a které umožňují empiricky ověřovat mj. principy vývoje jazyka jako systému a odhalují tak nové způsoby propojení langue a parole.

6. Literatura

Bybee, Joan. 1985. Morphology: A Study of the Relation between Meaning and Form. Amsterdam: John Benjamins.

Comrie, Bernard. 1989. Language Universals and Linguistic Typology. 2. vydání. Oxford: Blackwell.

Diessel, Holger. 2014. Usage-based linguistics. In Mark Aronoff (eds.), *Oxford Bibliographies in "Linguistics"*. New York: Oxford University Press.

Dixon, Robert M. W. 1979. Ergativity. *Language* 55. 59-138.

Dixon, Robert M. W. 1994. Ergativity. Cambridge: Cambridge University Press.

Filimonova, E. 2005. The noun phrase hierarchy and relational marking: Problems and counterevidence. *Linguistic Typology 9*. 77–113

Galili, Tal. 2015. dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics.<doi:10.1093/bioinformatics/btv428>.

Gries, Stefan Th. & Otani, Naoki. 2010. Behavioral profiles: a corpus-based perspective on synonymy and antonymy. *ICAME Journal* 34. 121-150.

Gries, Stefan Th. & Divjak, D. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1). 23-60.

Hajič, Jan. 2004. Disambiguation of Rich Inflection (Computational Morphology of Czech). Vol. 1. Karolinum Charles University Press, Praha.

Hajič, Jan; Spoustová, Drahomíra; Votrubec, Jan; Krbec, Pavel; Květoň, Pavel. 2007. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*. ACL 2007, Praha. pp. 67–74.

Heine, B. & König Ch. 2010. On the linear order of ditransitive objects. *Language Sciences 32*. 87‑131

Hopper, Paul. 1987. Emergent grammar. In *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*. Edited by Jon Aske, Natasha Beery, Laura Michaelis, and Hana Filip, 139–157. Berkeley, CA: Berkeley Linguistics Society.

Janda, Laura A. & Lyashevskaya, Olga. 2011. Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian. *Cognitive Linguistics 22(4)*: 719–763.

Jelínek, Tomáš. 2008. Nové značkování v Českém národním korpusu. In: *Naše řeč*, 91, 1, pp. 13–20.

Ji, Jie. 2017-07. Animacy hierarchy within inanimate nouns: English corpus evidence from a prototypical perspective. Paper presented at Corpus Linguistics Conference 2017. 24 to 28 July 2017. University of Birmingham.

Kaufman, L. & Rousseuw, P.J. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken: John Wiley and Sons.

Křen, Michal et al. 2015. SYN2015: a representative corpus of written Czech. Institute of Czech National Corpus, Faculty of Arts, Charles University in Prague. <www.korpus.cz>.

Kuznetsova, Julia. 2015. Linguistic profiles: Going from form to meaning via statistics. Berlin, Boston: De GruyterMouton.

Lance, G. N. & Williams, W. T. 1967. Mixed-data classificatory programs I.) Agglomerative Systems. *Australian Computer Journal*: 15–20.

Langacker, R. W. 1991. Foundations of Cognitive Grammar, Vol. 2: Descriptive Application. Stanford: Stanford University Press.

Levshina, Natalia. 2014. Rling: A companion package for How to Do Linguistics with R. R p ackage version 1.0.

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

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. 2017. cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6.

Meyer, David, Zeileis, Achim, and Hornik, Kurt. 2016. vcd: Visualizing Categorical Data. R package version 1.4-3.

Petkevič, Vladimír. 2006. Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In: *Insight into the Slovak and Czech Corpus Linguistics (Šimková M. ed.)*. Veda, Bratislava, pp. 26–44.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. <https://www.R-project.org >.

Silverstein, M. 1976. Hierarchy of features and ergativity. *Grammatical Categories in Australian Languages*. 112–171.

Suzuki, Ryota & Shimodaira, Hidetoshi. 2015. pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling R package version 2.0-0.

7. Seznam příloh

1. Soupis excerpt
2. Grafy (Příloha 2.zip)

7.1 Soupis excerpt

LEMMA	KATEGORIE
Novák M	1
Karel M	1
Tomáš M	1
Hana F	1
Kateřina F	1
Josef M	1
Ondřej M	1
lenka F	1
Jiří M	1
David M	1
Monika F	1
Jan M	1
Petr M	1
Martin M	1
Pavel M	1
Richard M	1
Miroslav M	1
Petra F	1
Martina F	1
Vera F	1
Roman M	1
Vanessa F	1
Tamara F	1
Klíma M	1
Sokrates M	1
Evka F	1
Jíťa F	1
Vlasák M	1
Špotáková F	1
Třeštíková F	1
voják M	2
režisér M	2
poslanec M	2
ředitel M	2
prezident M	2
ministr M	2
policista M	2
předseda M	2
básník M	2
královna F	2
právník M	2
učitelka F	2
vrah M	2
správce M	2
fotograf M	2
číšník M	2
lékař M	2
starosta M	2
brankář M	2
zpěvák M	2
šéfredaktor M	2
výtvarník M	2
prostitutka F	2
horolezec M	2
mluvčí M	2
zřízenec M	2
poradkyně F	2
bratr M	3
strýc M	3
otec M	3
matka F	3
syn M	3
dcera F	3
máma F	3
babička F	3
dědeček M	3
teta F	3
bratranec M	3
vnuk M	3
vnučka F	3
kmotr M	3
synovec M	3
neteř F	3
praotec M	3
táta M	3
potomek M	3
sestřenice F	3
tchyně F	3
prarodič M	3
vnouče N	3
kmotra F	3
pramatka F	3
kamarádka F	4
soudruh M	4
kolegyně F	4
manželka F	4
kamarád M	4

kolega M	4
partner M	4
přítelkyně F	4
milenec M	4
milenka F	4
nevěsta F	4
parťák M	4
ženich M	4
spolužák M	4
partnerka F	4
snoubenka F	4
soudružka F	4
snoubenec M	4
spolužákyně F	4
expřítel M	4
žena F	5
muž M	5
slečna F	5
dívka F	5
holka F	5
panna F	5
ženská F	5
panic M	5
anděl M	6
čert M	6
démon M	6
upír M	6
troll M	6
kentaur M	6
duch I	6
robot I	6
přízrak I	6
monstrum N	6
strašidlo N	6
mumie F	6
vlkodlak M	6
golem M	6
zombie F	6
vodník M	6
skřet M	6
ježibaba F	6
mrazík I	6
rusalka F	6
harpyje F	6
divoženka F	6

hejkal M	6
jezinka F	6
upírka F	6
ufon M	6
rarach M	6
zlobr M	6
troglodyt M	6
Minotaurus M	6
pes M	7
kráva F	7
husa F	7
kůň M	7
kočka F	7
koza F	7
ovce F	7
kocour M	7
králík M	7
fena F	7
prase N	7
kuře N	7
slepice F	7
býk M	7
kachna F	7
štěně N	7
kotě N	7
kozel M	7
klisna F	7
morče N	7
vepř M	7
jezevčík M	7
kobyla F	7
andulka F	7
krůta F	7
prasnice F	7
beran I	7
medvěd M	8
motýl M	8
včela F	8
vlk M	8
zajíc M	8
slon M	8
orel M	8
tygr M	8
jelen M	8
červ M	8

krokodýl M	8
vrána F	8
sova F	8
myška F	8
cvrček M	8
potkan M	8
pavián M	8
opice F	8
ježek M	8
gorila F	8
vrabec M	8
čedič I	8
srna F	8
laň F	8
housenka F	8
Srnec M	8
koroptev F	8
činnost F	9
služba F	9
řízení N	9
rozhodnutí N	9
setkání N	9
hlasování N	9
transformace F	9
poranění N	9
výroba F	9
porovnání N	9
zkoumání N	9
vyjednávání N	9
napojení N	9
mrknutí N	9
chování N	9
jmenování N	9
zvyšování N	9
chápání N	9
stíhání N	9
zásobování N	9
zadržení N	9
pořízení N	9
vydávání N	9
trávení N	9
umírání N	9
vyvážení N	9
přispění N	9
donucení N	9

zalíbení N	9
hlava F	12
život I	12
ruka F	12
noha F	12
prst I	12
rameno N	12
mozek I	12
oko N	12
dlaň F	12
koleno N	12
kost F	12
břicho N	12
ucho N	12
žaludek I	12
plíce F	12
palec I	12
nehet I	12
loket I	12
prdel F	12
chodidlo N	12
penis I	12
hruď F	12
řiť F	12
hrtan I	12
kniha F	13
stůl I	13
židle F	13
zrcadlo N	13
počítač I	13
knížka F	13
mobil I	13
klíč I	13
nůž I	13
automobil I	13
talíř I	13
notebook I	13
záchod I	13
tužka F	13
peněženka F	13
hřeben I	13
sešit I	13
umyvadlo N	13
pyžamo N	13
obojek I	13

metro N	13
propiska F	13
internet I	13
ústí N	14
Lhota F	14
Praha F	14
Francie F	14
Ostrava F	14
Morava F	14
Šumava F	14
Evropa F	14
Londýn I	14
Brno N	14
Liberec I	14
Japonsko N	14
Asie F	14
Brusel I	14
Vltava F	14
Vyšehrad I	14
Sněžka F	14
Madagaskar I	14
Haná F	14
Everest I	14
Ještěd I	14
Dyje F	14
Temže F	14
kopec I	15
doba F	16
vteřina F	16
den I	16
čas I	16
hodina F	16
minuta F	16
týden I	16
období N	16
budoucnost F	16
ráno N	16
zítřek I	16
minulost F	16
odpoledne N	16
poledne N	16
úsvit I	16
včerejšek I	16
dnešek I	16
smutek I	17

hloupost F	17
zlomek I	17
zvláštnost F	17
takt I	17
relace F	17
rozmar I	17
doktrína F	17
scenerie F	17
etika F	17
mánie F	17
sorta F	17
tryzna F	17
přízeň F	17
autonomie F	17
morfologie F	17
sci-fi N	17
startup I	17
smůla F	17
nechuť F	17
naivita F	17
negativita F	17
rozkvět I	17
otročina F	17
okurka F	18
bříza F	18
meloun I	18
ořech I	18
dub I	18
banán I	18
bylinka F	18
plevel I	18
réva F	18
jetel I	18
obilí N	18
skořice F	18
hořčice F	18
celer I	18
petržel F	18
rybíz I	18
šafrán I	18
jitrocel I	18
konopí N	18
tymián I	18
hrách I	18
rákosí N	18

pórek I	18
čepice F	20
košile F	20
kabát I	20
klobouk I	20
tričko N	20
bunda F	20
svetr I	20
šátek I	20
šála F	20
blůza F	20

střevíc I	20
baret I	20
rukavice F	20
triko N	20
mikina F	20
burka F	20
kalhoty F	20
já/my NON	99
ty/vy NON	99
on/oni/ona/ony/ono/ona/on/ony NON	99