OPEN ACCESS

Diachronní korpusová analýza: slovosled českých posesivních adjektiv uvnitř nominální fráze

Jan Křivan — Michal Láznička

ABSTRACT:
Diachronic corpus analysis: the order of Czech possesive adjectives within nominal phrase. This paper is concerned with the diachronic development of the placement of Czech possessive adjectives relative to the head noun in Old and Middle Czech. At the same time, the aim of this study is also to introduce a possible way of approaching complex language data. We base our analysis on cross-linguistic synchronic generalizations regarding possessor placement which connect monolexemic possessors (which are high on the nominal animacy hierarchy) to the prenominal position. A sample of 1417 possessive adjectives obtained from available sources of Old and Middle Czech texts was annotated for an array of semantic and syntactic variables. The relationship between these variables and the possessor placement was analysed using classification trees and random forests. The results do not support the synchronic generalizations. We interpret this finding by positing two frequent, lexically partially filled constructions, *N Kristův* 'N of Christ' and *syn N-ův* 'son of N'. We conclude that the patterns observed in the data can be explained by the interaction of extralinguistic socio-cultural factors and the effects of frequency and similarity in these two constructions.

KLÍČOVÁ SLOVA / KEY WORDS:
adnominální posesivita, diachronní korpusová analýza, klasifikační stromy, lingvistika založená na užívání jazyka, náhodné lesy, postavení posesora uvnitř nominální fráze
adnominal possession, classification trees, diachronic corpus analysis, possessor placement within nominal phrase, random forests, usage-based linguistics

1. ÚVOD

V posledních několika desetiletích prochází lingvistika v mnoha svých subdisciplínách výraznou změnou, která je charakterizována důrazem na využití empirických metod zkoumání a práci s relativně velkými soubory dat. V tomto procesu dochází také k postupnému osvojování rigoróznějších metod práce a přebírání pokročilých statistických technik, které jsou vhodné pro analýzu dat s charakteristikami typicky vídanými v lingvistickém výzkumu. Průkopníky tohoto směru jsou studie psycholingvistické, které mají v určitém slova smyslu blíže k experimentální psychologii než k teoretické lingvistice. V posledních letech se však tento trend prosazuje i v korpusové lingvistice, lingvistické typologii či sociolingvistice (např. Poplack & Cacoullos, 2015; Widmer et al., 2017).

Po svém nástupu na počátku 80. let procházejí také podoblasti kognitivně a funkčně orientované lingvistiky v posledních desetiletích podobným empirickým obratem. S tím souvisí rostoucí zájem o lingvistiku založenou na užívání jazyka

(*usage-based linguistics*) a emergentistický pohled na gramatiku (Bybee, 1985, 2006; Hopper, 1987; Langacker, 1988). Tyto přístupy se dívají na jazyk jako na komplexní dynamický (někdy též *adaptivní*) systém (Beckner et al., 2009), tj. jako na „dynamický systém proměnlivých kategorií a flexibilních omezení, které jsou neustále přestrukturovávány a reorganizovány pod tlakem doménově nespecifických kognitivních procesů" (Diessel, 2017, vlastní překlad).

Základní principy jsou následující: (i) Jazyk slouží ke zprostředkování komunikace mezi jednotlivými členy daného jazykového společenství, kteří její pomocí naplňují své cíle. (ii) Mluvčí mezi sebou interagují s využitím inventáře jazykových prostředků různé úrovně uložených v paměti z předchozích komunikačních situací (*usage events*).[1] (iii) Na základě výsledků těchto interakcí jsou některé prvky tohoto mentálního inventáře zvýhodňovány a jiné naopak penalizovány. Jazyková struktura, systematičnost a idiosynkratičnost, které typicky pozoruje a popisuje lingvistika, se pak „vynořují" (*emerge*) jako vedlejší produkt těchto lokálních interakcí na úrovni jednotlivých mluvčích.

Tyto přístupy tak ze své podstaty umožňují vysvětlovat jazykové chování mluvčích jak pomocí sociálních faktorů (předpoklad bohatých informací o interakčním a obecně mimojazykovém kontextu jako součásti reprezentace), tak doménově nespecifických kognitivních mechanismů (efekty frekvence různé úrovně, podobnosti a analogie, recentnosti a primingu). Z výše uvedeného zároveň plynou dvě skutečnosti: (i) jednak se nepředpokládá, že by mentální gramatiky všech příslušníků daného společenství byly zcela totožné, (ii) jednak se tyto mentální gramatiky vyvíjejí během celého života mluvčího; v důsledku lze předpokládat, že po každé jednotlivé komunikační situaci dochází k aktualizaci těchto gramatik. To zároveň implikuje, že nositeli jazykové změny v tomto přístupu nejsou primárně děti osvojující si daný jazyk, ale spíše dospělí mluvčí.

Do popředí zájmu se tím dostávají variující, blízce synonymní struktury různého řádu, u nichž jsou zkoumány faktory, které vedou k výběru jedné ze „soupeřících" konstrukcí nad jinou. Stále častěji se pak u takových výzkumných problémů uplatňuje korpusový výzkum, při němž je velký vzorek anotován s ohledem na řadu teoreticky motivovaných proměnných a následně jsou konstruovány statistické modely, s jejichž pomocí se usuzuje na vliv daných faktorů na užití některé z variant konstrukce.[2] Takové modely potom jednak popisují a za pomoci sledovaných prediktorů vysvětlují pravidelnosti v daném datovém souboru, jednak predikují chování mluvčích. Tyto predikce je pak možno testovat nepřímo na dalších textech v následných korpusových analýzách či přímo v laboratorním prostředí manipulací zjištěných prediktorů v experimentálním kontextu.

1 Teoreticky vzato se jedná o jednotky sahající od jednotlivých artikulačních gest přes lexikální slova a zobecněné, lexikálně nespecifické konstrukce až po nadvětné celky.

2 Jak upozorňují Milin et al. (2016), tyto modely typicky nemají ambici být kognitivně realistické v tom smyslu, že by předpokládaly podobný typ analýzy (typicky lineární regrese) v myslích mluvčích při zpracování jazyka, volené techniky modelování vycházejí spíše z charakteristik dat a hledají kompromis mezi dostatečně výkonnou a přiměřeně komplexní analýzou.

Nutnou součástí uvažování o jazyce v rámci přístupů založených na užívání je dále logicky i diachronní výzkum. V případě existence konkurenčních prostředků je po každé úspěšné komunikační situaci posílena paměťová stopa užité varianty a její asociace s daným kontextem, zatímco reprezentace konkurenční varianty (či její asociace s daným kontextem) může být oslabena. Jedna z variant se pak postupně může stát pro daného mluvčího natolik úspěšnou či užitečnou, že druhou zcela přestane používat a její paměťová stopa postupně téměř zanikne. Tento proces se může opakovat u tak velkého množství mluvčích, že dojde k jazykové změně a jedna z variant se zcela ztratí či zůstane velmi silně kontextově omezena. Jazykový vývoj je tak přirozeně vnímán jako odraz a extenze stejných principů, které je možno pozorovat v každém okamžiku v mikroperspektivě jednotlivých komunikačních situací. Také výše popsaná metodologie je přirozeně aplikovatelná i v diachronním výzkumu. Studujeme-li období vývoje jazyka, pro které je dostupný relativně objemný textový materiál, a zaměřujeme-li se zároveň na dvě či více konstrukcí s podobnou funkcí, které si v určitém období konkurovaly, je možné s využitím zmiňovaného přístupu k jazyku lépe popsat a pochopit kontexty, které favorizují tu či onu konstrukci, a v důsledku lépe pochopit vývoj v dané fázi. Takové studie, které kombinují perspektivu mechanismů ovlivňujících reprezentace a zpracování jazyka na individuální úrovni s perspektivou vývoje v časových úsecích přesahujících jednotlivé generace, se začínají v poslední době objevovat stále častěji (např. Hilpert, 2013 či Wolk et al., 2013).

Tato studie vychází z výše uvedených principů a klade si dva cíle: na příkladu vývoje slovosledu v rámci adnominální posesivní konstrukce v češtině chceme za využití tohoto obecného metodologického rámce (i) představit možnosti využití moderních pokročilých statistických metod a (ii) zároveň popsat některé faktory, které mohly mít na vývoj konstrukce vliv, a tím rozšířit naše poznatky o této problematice v české historické mluvnici. Konkrétně se soustředíme na variaci v postavení posesivního adjektiva (posesora) v rámci adnominální posesivní konstrukce vůči řídícímu substantivu (posesu) v období od počátku 14. století do roku 1850.[3] Vycházíme přitom z několika mezijazykových pozorování, která spojují pozici posesora s některými sémantickými, syntaktickými a diskurzně-pragmatickými faktory. Vedle této výchozí hypotézy pak v rámci tohoto exploratorního výzkumu sledujeme další faktory, které mohly mít na postavení posesora vliv.

Struktura článku je následující: V druhém oddílu představíme vstupní hypotézu a výzkumné otázky a popíšeme dosavadní poznatky o slovosledu uvnitř jmenné fráze v češtině. Ve třetím oddílu popíšeme proces tvorby vzorku a povahu zdrojových souborů dat a následné anotace konkordancí ve vzorku včetně přehledu jednotlivých sledovaných proměnných. Ve čtvrtém oddílu představíme zvolenou metodu analýzy, klasifikační stromy a náhodné lesy. V pátém oddílu představíme výsledky analýzy spolu s jejich interpretací. V závěru článek krátce shrneme.

3 V celém článku používáme pojmy *posesor* pro vlastníka a *posesum* pro vlastněné.

2. VÝZKUMNÁ OTÁZKA A TEORETICKÉ PŘEDPOKLADY PRO VÝVOJ SLOVOSLEDU UVNITŘ NOMINÁLNÍ FRÁZE

2.1 VÝZKUMNÁ OTÁZKA

Způsob vyjádření posesivity uvnitř nominální fráze se v současné češtině vyznačuje obecně známou dichotomií: lexikálně vyjádřeného posesora lze za určitých morfologických podmínek vyjádřit buď posesivním adjektivem, nebo genitivní konstrukcí. Obojí způsob vyjádření má přitom (do značné míry) pravidelné koreláty z hlediska slovosledu: posesivní adjektiva se vyskytují uvnitř nominální fráze v antepozici (1a), zatímco genitivní konstrukce převážně (tj. s výjimkou zvláštních případů prenominálního genitivu) v postpozici (1b).

(1) a. *Janova kniha*
b. *Kniha krále Jana*

Tento stav je však relativně nový. Bylo už dříve zjištěno, že v nejstarších zdokladovaných obdobích, tj. od 14. století, mohla stát všechna adjektiva (tedy i relační a kvalitativní) jak v antepozici, tak v postpozici (Gebauer, 2007/1929). Variaci u posesivních adjektiv ilustruje příklad (2).

(2) a. *Jozef, jenž jest uprosil u Piláta [tělo Ježíšovo]*
(Vokabulář webový 2006–2016a, položka 6640, datace 1442)
b. *S [rychtářovým odpuštěním] prsty zdvihl.*
(Vokabulář webový 2006–2016a, položka 67078, datace 1490)

Tato slovosledná variace existovala až do 19. století. Dosud však nebylo podrobněji zkoumáno, jak se uvedená variace v průběhu let vyvíjela, ani nebylo podrobně popsáno, jak se měnil poměr obou sledů a které faktory měly vliv na vývoj k současnému stavu.[4]

Stanovili jsme si tedy následující výzkumnou otázku: jak se vyvíjelo slovosledné postavení adjektivního posesora uvnitř nominální fráze v češtině od počátku 14. století do roku 1850. Konkrétněji jsme potom chtěli především zachytit změny v poměru antepozice a postpozice v závislosti na čase a některých jazykových i mimojazykových faktorech.

4 Variaci a vývoji adnominální posesivní konstrukce v současné češtině se věnoval Vachek, který argumentoval ve prospěch pronikání genitivu na místo posesivních adjektiv tím, že v objektivním pořádku tématu a rématu je „žádoucí, aby označení vlastněného předmětu, jakožto členu určovaného, předcházelo před označením vlastníka, které platí za člen určující“ (Vachek, 1972, s. 147). Toto vysvětlení je založeno na zcela opačné argumentaci než vysvětlení, která předkládáme níže.

2.2 TEORETICKÉ PŘEDPOKLADY A VSTUPNÍ HYPOTÉZA

Pro náš výzkum jsme přijali teoretické předpoklady, které vycházejí z funkčně zaměřených přístupů (*usage-based linguistics*) nastíněných v úvodu tohoto článku. Využili jsme zejména poznatků jazykové typologie.

V germánských, románských a slovanských jazycích bylo na synchronních datech pozorováno (O'Connor, Maling & Skarabela, 2013), že nominální konstrukce s posesorem vyjádřeným jedním lexémem jsou kategoriálně omezeny na antepozici (viz výše (1) v současné češtině). Stejná tendence (tj. nikoliv kategoriální omezení, ale tendence v jazykovém užívání) se projevuje i v angličtině, pokud jde o preferenci prenominálního posesora v tzv. *'s* genitivu (3a) před posesorem prepozicionálním (3b).

(3) a. *John's book*
b. *A book of King John*

Nominální fráze s prenominálním posesorem jsou v anglickém úzu spojeny s vlastnostmi, které se na základě rozsáhlých zkoumání lingvistické typologie z různých jazyků světa přisuzují prototypickým agentivním argumentům a posesorům: zejména umístění vysoko na škále na tzv. nominální hierarchii životnosti, viz např. Aikhenvald (2013):

> Posesor má tendenci obsazovat pozici relativně vysoko v nominální hierarchii: prototypický posesor je životný nebo lidský a je vyjádřen osobním zájmenem nebo vlastním jménem.
> (Aikhenvald, 2013, s. 40, vlastní překlad)

S tím úzce souvisí, že se posesoři společně s agentivními argumenty často identifikují jako entity vysoce aktivované v informační struktuře věty (tj. jde typicky o známou informaci — téma) (pro češtinu viz Křivan, 2014). Zároveň platí, že v antepozici mají tendenci se vyskytovat entity syntakticky lehké, tj. jednoduché, nerozvité (Behaghel, 1909; nověji viz např. hypotézy Hawkinse, 2004).

Jedná se tedy v úhrnu o to, že prenominální posesivní konstrukce jsou podle těchto pozorování (diskuse viz O'Connor, Maling & Skarabela, 2013) typicky spjaty s nerozvitými (syntakticky lehkými) životnými posesory, kteří jsou z hlediska diskurzu známí (referenčně aktivovaní).

Naše vstupní hypotéza je proto následující: očekáváme, že právě takovéto entity se v historickém vývoji češtiny častěji vyskytovaly v antepozici a lze uvažovat o roli těchto konstrukcí v daných kontextech (tj. zejména o roli diskurzního statusu posesora v těchto konstrukcích) jako spouštěčů celé kategoriální změny. Rozhodli jsme se tedy především sledovat referenční, sémantickou a syntaktickou charakteristiku posesora a celého posesivního spojení, abychom se pokusili odhalit uvedené principy.

Jinými slovy, zatímco vznik konstrukce s posesorem v antepozici je na základě dosavadních výzkumů *usage-based linguistics* vysvětlen současným užíváním (rozložením diskurzně-pragmatických rysů v synchronních datech), v našem výzkumu se snažíme najít pro tato pozorování oporu v diachronních datech.

Na závěr této části musíme zdůraznit, že i přes řadu výše nastíněných teoretických předpokladů budeme provádět analýzu exploratorní. To znamená, že sledujeme nejen faktory podporující naše předpoklady, ale také další, zejména společenské a textové charakteristiky, které mohly mít na zkoumanou variaci vliv. V následujícím oddílu představíme skladbu vzorku a řadu charakteristik, které jsme anotovali, abychom dosáhli výše uvedených cílů.

3. SESTAVENÍ VZORKU DAT A ANOTACE JEDNOTLIVÝCH POLOŽEK

3.1 VZOREK DAT

3.1.1 ZDROJOVÁ DATA

K sestavení vzorku jsme se rozhodli primárně využít edičně zpracovaných dat infrastruktury RIDICS Ústavu pro jazyk český AV ČR. Data z období od 14. do 16. století jsme získali ze *Staročeské textové banky* (Vokabulář webový, 2006–2016a), data z období od 16. století do poloviny 19. století pocházejí ze *Středněčeské textové banky* (Vokabulář webový, 2006–2016b). Vzhledem k nižšímu objemu textů z období střední češtiny jsme data částečně doplnili o diachronní zdroje Českého národního korpusu (diakorp v6: Kučera et al., 2015) z let 1650 až 1850, s tím, že data z duplicitních zdrojů jsme zahrnuli vždy pouze ve verzi ze Středněčeské textové banky. Všechny uvedené korpusy jsou z povahy věci nereprezentativní, leč pro značnou část období se jedná o jediné dostupné (existující) texty dané doby.

3.1.2 EXTRAKCE A PŘÍPRAVA KONKORDANCÍ

Z uvedených zdrojů jsme pomocí regulárních výrazů získali konkordance všech potenciálních výskytů posesivních adjektiv, viz (4). Tímto postupem jsme extrahovali celkem 85 104 položek. (Za získání dat z textových bank ÚJČ děkujeme Borisovi Lehečkovi a Evě Lehečkové.)

(4) [word=".*(ov|in)(ých|ým|ými|ýma|ejch|ejm|ejma)"] | [word=".*ov[aoyuěi]"] | [word=".*in[aoyuěi]"] | [word = ".*(ó|uo|ů)v"]

Výsledky dotazu jsme uložili do tabulkové podoby. Každá konkordance obsahovala kromě samotného adjektiva a jeho levého a pravého kontextu také metadata z textových bank ÚJČ: jednoznačnou identifikaci položky, identifikaci zdroje, jeho přesné datování, uvedení padesátileté periody, literární druh a literární žánr. Položky diakorpu, kde je použito částečně odlišné členění, jsme převedli do této šablony.

Následně jsme ručně prošli a odfiltrovali všechny položky, které neodpovídají výskytu posesivních adjektiv. Tím jsme získali soubor 13 254 položek jako základ pro vytvoření samotného vzorku (9380 adjektiv odvozených z proprií, 3874 odvozených z apelativ).

3.1.3 VYTVOŘENÍ VZORKU

Data k analýze jsme se rozhodli připravit jako stratifikovaný vzorek, v němž bude (i) zajištěno dostatečné zastoupení jak adjektiv odvozených z proprií, tak adjektiv odvozených z apelativ, (ii) obsažen alespoň minimální počet dokladů ze všech padesátiletých period, který umožní statistickou analýzu, (iii) u period zastoupených více doklady bude využito většího množství položek.

K anotaci jsme připravili 1000 adjektiv z proprií a 500 adjektiv z apelativ (počet deapelativních adjektiv ve vzorku jsme tak oproti striktně poměrnému zastoupení mírně navýšili). V tabulkovém editoru jsme vytvořili systém vzorců, který spočítal doklady v každé z period. Následně jsme počty dokladů do vzorku depropriálních i deapelativních adjektiv zajistili ve dvou krocích: (i) v prvním skrutiniu jsme rozpočítali dvě třetiny dokladů rovnoměrně podle period (pokud doklady v dané periodě dosahovaly alespoň tohoto poměrného počtu), (ii) v druhém skrutiniu byla zbylá třetina rozdělena podle přebytků všech period (v jejich vzájemném poměru). Uvedený vzorek jsme uložili do tabulky, aby mohl být následně doplněn o anotační kategorie. Při práci na anotacích (viz oddíl 3.2.1) jsme pak museli ještě vyřadit další konkordance, které nebyly posesivními adjektivy nebo pro ně nebyly dostupné veškeré požadované informace. Finální vzorek tak obsahoval celkem 1417 výskytů, z nichž je 466 deapelativních a 953 depropriálních posesorů.

3.2 ANOTACE

3.2.1 ANOTACE ZÁKLADNÍCH CHARAKTERISTIK

V tabulkových souborech jsme s ohledem na výzkumnou otázku připravili základní anotační charakteristiky, které lze vysledovat z kontextu jednotlivých dokladů: (i) informace o posesorovi: (a) lemma posesora, (b) rozvití posesora (bez rozvinutí, koordinace, modifikace, apozice), (c) životnost posesora (lidský, ostatní životný, neživotný); (ii) informace o nominální frázi: (a) slovosled nominální fráze (celá nominální fráze členěná po slovech: tři pozice nalevo a napravo od řídícího členu NP), (b) syntakticko-sémantické určení modifikátorů (posesor, demonstrativum, kvantifikátor, relační adjektivum, kvalitativní adjektivum, genitivní konstrukce, předložková konstrukce, vztažná klauze), (c) pád řídícího členu NP, (d) řídící člen řídícího členu NP (jeho slovní druh), (e) (ne)projektivita posesora a řídícího členu v rámci klauze, (f) pozice řídícího členu NP vůči slovesu (před slovesem: téma, za slovesem: réma). Uvedené charakteristiky byly prakticky anotovány tak, aby bylo možné další zpracování kategorií (kombinování kategorií, vylučování hodnot z kategorií, jejich rozdělování i slučování). Základní anotaci dokladů připravily pod naším vedením studentky FF UK Lucie Salzmannová a Anna Stuchlá, následně jsme provedli kontrolu všech položek a vyřadili jsme další konkordance, které nebylo možné zahrnout do vzorku (viz oddíl 3.1.3), abychom mohli provést finální anotaci pro potřeby statistických výpočtů.

3.2.2 FINÁLNÍ ANOTACE KATEGORIÍ PRO STATISTICKÉ ANALÝZY

Připravené položky jsme sloučili do jedné tabulky a připravili tak datový soubor pro statistické zpracování. Všechny sledované proměnné a způsob jejich vytvoření představíme na následujících řádcích po jednotlivých tematických skupinách.

Závislá proměnná
Závislou proměnnou (A) jsme extrahovali pomocí jednoduchého vzorce z komplexních informací o slovosledu nominální fráze.

(A) pozice posesora vůči řídícímu substantivu (*pr.position*): antepozice (*initial*), postpozice (*final*)[5]

Metadata
Vzhledem k diachronní povaze výzkumu jsme museli pečlivě zvážit časové hledisko. Jednou z možností je pojmout čas jako spojitou proměnnou, kde jsou díla reprezentována pomocí vročení. Spojitý charakter této proměnné by však zachytil i řadu idiosynkratických rozdílů (např. mezi jednotlivými tituly), proto jsme využili přístup, který pracuje s intervaly a zároveň přímo na základě dat identifikuje periody, v nichž dochází k diachronní změně ve variaci sledované závislé proměnné (A).

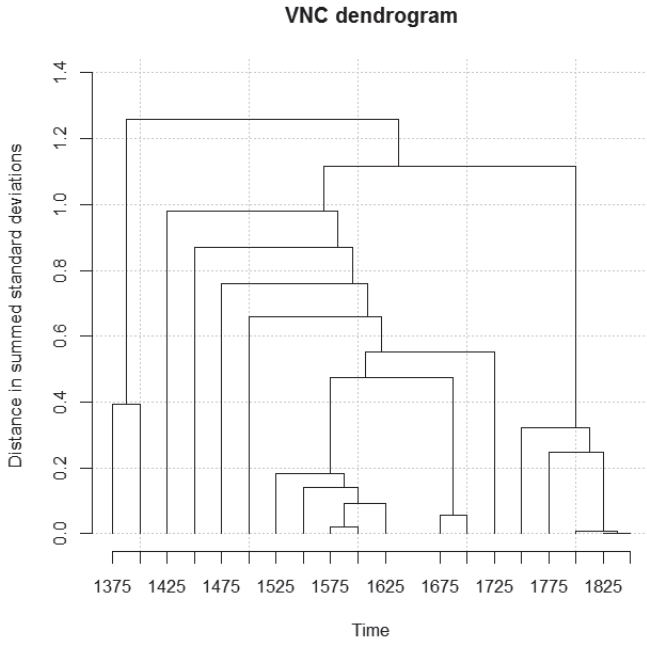
Časovou osu jsme nejprve rozdělili na periody po 25 letech (nejjemnější dělení, v němž bylo ještě smysluplné získaná data kvantitativně porovnávat, zároveň jsme kvůli nedostatku primárních dat sloučili nejstarší tři období 1300–1375 a z téhož důvodu zůstalo neobsazeno období 1625–1650). Následně jsme v těchto obdobích pomocí metody *variability-based neighbour clustering* (VNC, Gries & Hilpert, 2012a) a vypočítaného dendrogramu identifikovali celkem tři periody, v nichž se data z hlediska závislé proměnné chovají nejvíce kohezivně (tj. seskupují se do klastrů): 1300–1400, 1400–1725, 1725–1850, viz graf 1 a výsledná proměnná (B).[6] Výpočet jsme provedli v programu R (R Core Team, 2017) postupem zveřejněným na webu Gries & Hilpert (2012b) pomocí připravené funkce *vnc.individual*.

Zároveň jsme upravili anotaci žánrového členění (C).

(B) časové členění na periody metodou VNC (*period.id*): 1. rané období do konce 14. století (*early*), 2. střední období od začátku 15. století do první čtvrtiny 18. století (*high*), 3. pozdní období od druhé čtvrtiny 18. století do poloviny 19. století (*late*)
(C) žánry (*genre*): disputace (*arg*), bible a její části (*bib*), slovníky (*dict*), zábavná literatura (*fun*), kroniky (*chron*), naučná literatura (*learn*), naučná náboženská literatura (*learn.cr*), legendy (*leg*), noviny (*newsp*)

5 Kurzivou v závorce uvádíme názvy proměnných a jejich hodnot užité v anotačním schématu i při výpočtech.

6 Jedná se o využití klastrové analýzy, která je vzhledem k analyzované časové dimenzi omezena tím, že vytvoření uzlu je možné pouze u přímo sousedících datových bodů (odtud tedy *neighbour clustering*).



GRAF 1: Dendrogram vygenerovaný pomocí metody *variability-based neighbour clustering* se sledovanými časovými obdobími. Osa y = Vzdálenost jako součet směrodatných odchylek, osa x = časová období.

Sémantické vlastnosti posesivního spojení

Tři další charakteristiky se týkají sémantiky vztahu mezi posesorem a posesem. Následující kategorie jsme doanotovali na základě dalších úvah nad daty: status specifičnosti posesora založený na typu jména (D), ne/zcizitelnost posesa (E), sémantický typ posesa (F).

(D) status specifičnosti posesora založený na typu jména (*status*): proprium unikátní (*uni.prop*), proprium ostatní (*ot.prop*), apelativum unikátní (*uni.apel*), apelativum relační, tj. neunikátní (*neuni.apel*), apelativum ostatní (*ot.apel*)
(E) zcizitelnost posesa (*alienable*): zcizitelný (*a*), nezcizitelný, tj. příbuzenské vztahy a části těla (*i*)
(F) sémantika posesa, tj. řídícího členu fráze (*h.class*): příbuzenské a jiné osobní vztahy (*soc.rel*), část těla (*b.part*), událost (*event*), instituce (*instit*), místo (*place*), produkt (*product*), věc (*thing*)

Tyto kategorie si zaslouží stručný komentář. V případě charakteristiky posesora (D) se jedná o komplexní proměnnou motivovanou výchozí hypotézou. V první úrovni členění na typ jména posesora (propria, apelativa) vycházíme z nominální hierarchie životnosti (Silverstein, 1976), v níž stojí (lidská) propria výše než apelativa. Při dalších úvahách o statusu posesora nebylo vzhledem k povaze anotovaných konkordancí (nedostatečně dlouhý kontext) možné bezpečně určit míru kontextové aktivovanosti

daného referenta. Abychom zjemnili kategorizaci, včlenili jsme do kategorie v druhé úrovni rys specifičnosti, který vyjadřuje rozdíly v mimotextové aktivovanosti (jedinečná reference, relační entita, ostatní), tj. např. unikátní referenty typu *Mojžíš* je možno považovat za známé (aktivované) bez ohledu na kontext vzhledem k znalostem o světě. Do této kategorie jsme při anotaci řadili především postavy z biblických příběhů či legend, u nichž lze právě takový status předpokládat: je-li v textu zmíněn Abraham, není třeba jej blíže specifikovat jako referenta, neboť dané proprium bude automaticky přiřazeno k biblické postavě. Jednalo-li by se o jiného referenta, je nutno tento vztah v textu explicitně vytvořit. Logicky se tato kategorie týká především proprií, v případě apelativ se jedná o lemmata *bůh*, *hospodin*, *ďábel* a *satan*, která je možno považovat za svého druhu propria. U tzv. relačních entit je součástí jejich znalosti inherentní vztah k jinému referentu (Seiler, 1983), který přispívá ke zvýšené aktivaci; do této kategorie zahrnujeme apelativa označující rodinné členy či vládce. Zcizitelnost posesa (E) pak byla zvolena jako relativně objektivní ukazatel možné fixovanosti daného spojení, neboť lze předpokládat (viz např. Haspelmath, 2008), že nezcizitelná posesa se s vyšší frekvencí objevují v adnominální posesivní konstrukci, což může mít za následek určitý konzervační efekt a taková spojení mohou déle odolávat změně. Sémantický typ posesa (F) byl sledován jako ukazatel typu vztahu vyjádřeného užitím posesivního zájmena, od vlastnictví v užším slova smyslu po vyjádření sémantické role v nominalizacích.

Syntaktické vlastnosti nominální fráze a jejích částí
Proměnné pád řídícího členu (G) a řídící člen řídícího členu (H) jsme extrahovali přímo z primární anotace. Kombinace těchto proměnných umožňuje identifikovat pozici fráze v syntaktické struktuře.

(G) pád řídícího členu (*h.case*): přímý, tj. nominativ, akuzativ, genitiv (*dir*), nepřímý, tj. dativ, lokativ, instrumentál, vokativ (*indir*)
(H) pozice řídícího členu v syntaktické struktuře, tj. co je řídícím členem řídícího členu (*h.head*): sloveso (*v*), substantivum (*n*), předložka (*p*)

Syntaktická komplexnost nominální fráze a jejích částí:
Na základě detailních informací o řídícím členu fráze i na základě konkrétně vypsaných modifikátorů a posesorů jsme pomocí vzorců extrahovali proměnné týkající se komplexnosti (I)–(K). Komplexnost posesora přitom vychází z pozorování v literatuře týkajících se syntaktické váhy posesora. Ke sledovaným proměnným jsme zařadili i neprojektivitu (L).

(I) komplexnost řídícího členu fráze (*h.cmplx*): komplexní, tj. v koordinaci nebo apozici (*yes*), nekomplexní (*no*)
(J) komplexnost nominální fráze, tj. přítomnost dalšího modifikátoru společně s posesorem (vyjma demonstrativa) (*mod.other*): komplexní (*yes*), nekomplexní (*no*)
(K) komplexnost závislého členu, tj. posesora (*pr.cmplx*): rozvitý (*yes*), nerozvitý (*no*)
(L) neprojektivnost nominální fráze (*nonproj*): neprojektivní (*yes*), projektivní (*no*)

Informační struktura nominální fráze
Poslední sledovaná proměnná — rematičnost nominální fráze (M) — byla k dispozici přímo z předzpracovaných dat získaných z korpusu (viz 3.2.1).

(M) umístění v informační struktuře klauze, tj. rematičnost (*focus*): preverbální pozice a pozice v klauzi bez slovesného predikátu, tj. není réma (*no*), postverbální pozice, tj. je réma (*yes*)

Při členění této kategorie vycházíme z jednoduchých definic tématu a rématu (téma: o čem se mluví; réma: co se říká o tématu, např. Daneš, 1985). Pro účely výzkumu zároveň vycházíme z toho, že v textech převažuje objektivní pořádek slov, a tedy že slovosled lze považovat za hlavní indikátor informační struktury: u prvků v preverbální pozici očekáváme, že budou v tématu a že budou přístupné častěji než prvky v postverbální, rematické části klauze. Nominální fráze uvnitř neslovesných klauzí (obvykle nadpisy, slovníková hesla) považujeme za nerematické, anotujeme je tedy společně s frázemi v preverbálních pozicích.

4. ANALÝZA DAT

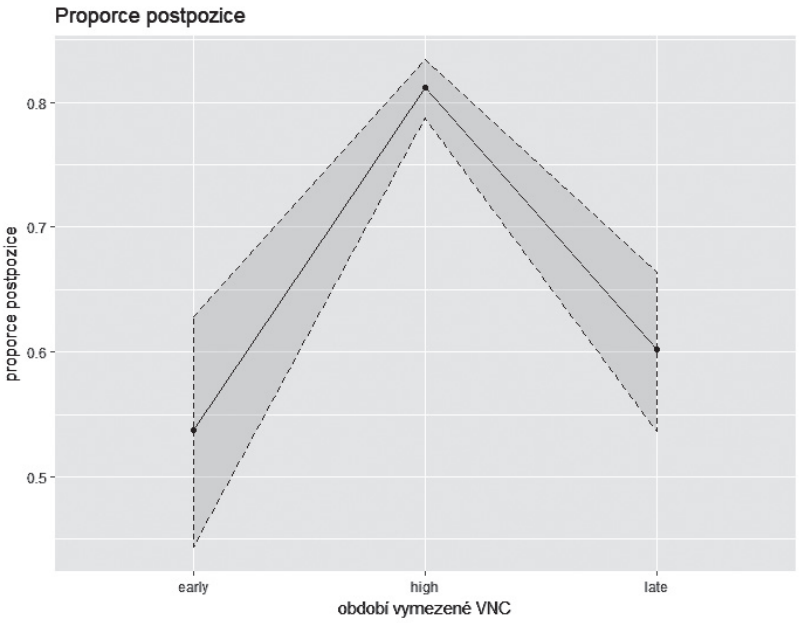
Po anotaci a vyloučení všech nežádoucích případů jsme v analýze pracovali celkem s 1417 výskyty, z nichž bylo 466 deapelativních a 953 depropriálních posesorů. Tři detekovaná období byla zastoupena následujícím počtem výskytů ve 151 různých textech: 1. rané období do konce 14. století: 108 výskytů ve 12 textech; 2. střední období od začátku 15. století do první čtvrtiny 18. století: 1083 výskytů v 96 textech; 3. pozdní období od druhé čtvrtiny 18. století do poloviny 19. století: 226 výskytů ve 43 textech.

4.1 VÝVOJ POZICE POSESORA VŮČI ŘÍDÍCÍMU SUBSTANTIVU

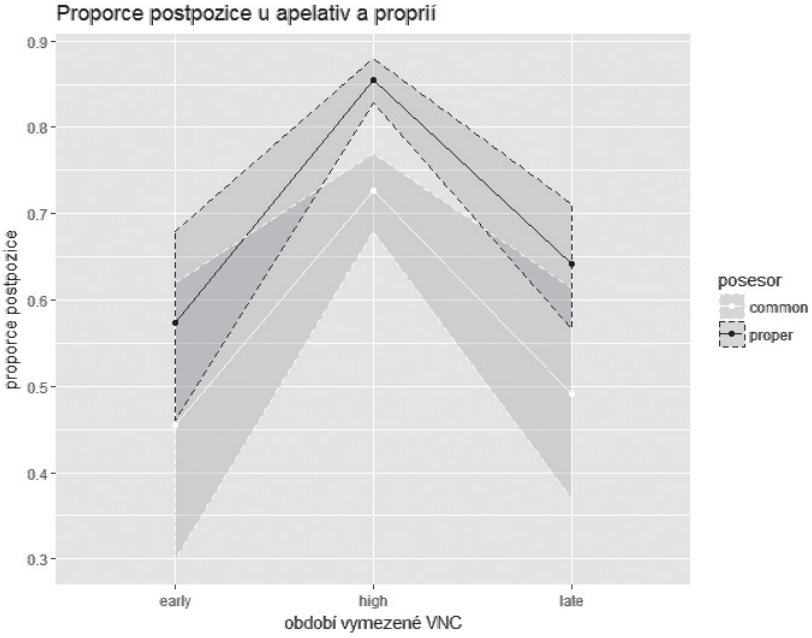
V prvním kroku jsme pro vizuální inspekci analyzovaných dat vynesli do grafu poměr antepozice a postpozice posesora v nominální frázi v průběhu času, a to jak (i) pro celý soubor bez ohledu na typ posesora (graf 2), (ii) tak s členěním na apelativa a propria (graf 3).[7]

Tyto grafy ukazují značnou míru variace po celé zkoumané období, v němž poměr mezi výskytem postpozic a antepozic neklesá pod 30 %, pouze depropriální adjektiva se ve středním období (*high*) blíží téměř kategorickému užívání postpozice. To naznačuje, že k definitivnímu ustálení pozice posesora došlo až v období na přelomu 19. a 20. století a později. Zároveň pozorujeme v datech určitý trend k nárůstu preference postpozice ve středním období a poté k opětovnému poklesu. Dále při pohledu na rozdíly v rozložení sledu pro typ posesora vidíme větší preferenci postpozice u proprií. Již na první pohled tak vidíme, že rozložení v datech příliš neodpovídá naší výchozí hypotéze. K tomuto zjištění se vrátíme v diskuzi po před-

7 Grafy byly vytvořeny v R (R Core Team, 2017) s použitím balíčku *ggplot2* (Wickham, 2009), verze 2.2.1.



GRAF 2: Poměr posesivních adjektiv v postpozici a antepozici ve třech periodách vymezených metodou VNC. Čárkované čáry značí 95% konfidenční intervaly.



GRAF 3: Srovnání poměru deapelativních a depropriálních posesivních adjektiv v postpozici a antepozici ve třech periodách vymezených metodou VNC (*common* dole, *proper* nahoře). Čárkované čáry značí 95% konfidenční intervaly.

stavení plné analýzy, jejímž cílem je zjistit, které ze sledovaných prediktorů k volbě slovosledu přispívají.

4.2 KLASIFIKAČNÍ STROMY

Anotovaný datový soubor jsme analyzovali pomocí metody klasifikačních stromů a jejího rozšíření pomocí náhodných lesů. Klasifikační a regresní stromy jsou neparametrickou metodou pro analýzu různých typů dat, při níž jsou pozorování rozdělena do podskupin na základě vztahu hodnot závislé proměnné a zvolených prediktorů (nezávislých proměnných).[8] Jedná se přitom o exploratorní metodu v tom smyslu, že strom je tvořen čistě na základě struktury analyzovaného datového souboru. Zatímco u jiných metod, jako je logistická regrese, se modeluje analyzovaný datový soubor s předpokladem určitého procesu generování daných dat, v případě klasifikačních stromů pouze dělíme datový soubor na základě kombinací hodnot prediktorů. Datový soubor je rozdělen na určité množství podskupin na základě binárního rekurzního dělení. V analýze jsme přitom využili přístup ke CART, který využívá podmíněné inference (*conditional inference*) (Hothorn, Hornik & Zeileis, 2006). Ten oproti klasickému přístupu (Breiman et al., 1984) řeší jednak problém přeučení stromu (*overfitting*) a z něj plynoucí nutnost jeho prořezávání (*pruning*), jednak problém toho, že původní algoritmus upřednostňuje při štěpení numerické prediktory a dále takové, které nabývají více hodnot a mají tak více potenciálních bodů pro štěpení. Algoritmus dělení u *conditional inference trees* přitom postupuje tak, že je otestováno, zda je některý z prediktorů asociován s danou závislou proměnnou, a na základě toho je zvolen prediktor, jehož míra asociace je nejvyšší. Zvolený prediktor slouží k rozdělení dat na dva podsoubory, v nichž prediktor nabývá pouze určitých hodnot, např. pro prediktor s hodnotami *a*, *b*, *c* může jedna skupina obsahovat pozorování s hodnotou *a*, druhá pozorování s hodnotami *b* a *c*. Tyto kroky jsou pak opakovány na jednotlivých podsouborech tak dlouho, dokud není možno odmítnout nulovou hypotézu o neexistenci asociace mezi prediktorem a závislou proměnnou. Vzhledem ke své podstatě je tato metoda vhodná v situacích, kdy jsou prediktory vzájemně korelované, což je v případě jazykových dat velmi častý případ; často můžeme pozorovat „koalice" faktorů, jejichž jednotlivý příspěvek je ve výsledku relativně malý, které však společně spolehlivě působí na hodnotu sledované proměnné určitým směrem. Metoda je též vhodná v situacích „malého *n* a velkého *p*" (*small n large p*), kdy sledujeme velké množství prediktorů při relativně malém počtu pozorování.[9] Další výhodou v porovnání s regresními koeficienty je intuitivní interpretace stromových struktur. Vzhledem k charakteru lingvistických dat je tak tato metoda vhodným doplněním standardní analytické výbavy (pro aplikaci v sociolingvistickém variačním kontextu viz Tagli-

8 Užívá se zkratka CART — *classification and regression trees*. V případě kategoriálních proměnných se pak hovoří o klasifikačních stromech, v případě numerických proměnných o regresních stromech. Pro uvedení do metody na příkladu lingvistických dat viz Baayen (2008) a Levshina (2015).

9 Všechny uvedené faktory mohou potenciálně způsobovat problémy při použití jiných regresních metod.

amonte & Baayen, 2012, pro aplikaci v korpusovém výzkumu např. Rezaee & Golparvar, 2016).

Analýzy jsme prováděli s použitím statistického softwaru *RStudio* (RStudio Team, 2016), což je IDE (*integrated development environment*) pro jazyk R (R Core Team, 2017). Klasifikační stromy jsme vytvořili pomocí balíčku *party* (Hothorn, Hornik & Zeileis, 2006).[10]

Nejprve jsme pomocí kódu (5) načetli datový soubor a za pomocí funkce `ctree()` z balíčku *party* jsme vytvořili klasifikační strom, abychom mohli analyzovat rozložení antepozice (*initial*) a postpozice (*final*) posesora v nominální frázi s ohledem na hodnoty všech sledovaných prediktorů.

(5)
```
adjectives <- read.csv(file = "adjectives.csv")
library(party)
set.seed(705)
strom.adj <- ctree(pr.position ~ period.id + status + pr.cmplx
+ genre + alienable + h.class + h.case + h.head + h.cmplx
+ focus + nonproj + mod.other, data = adjectives)
```

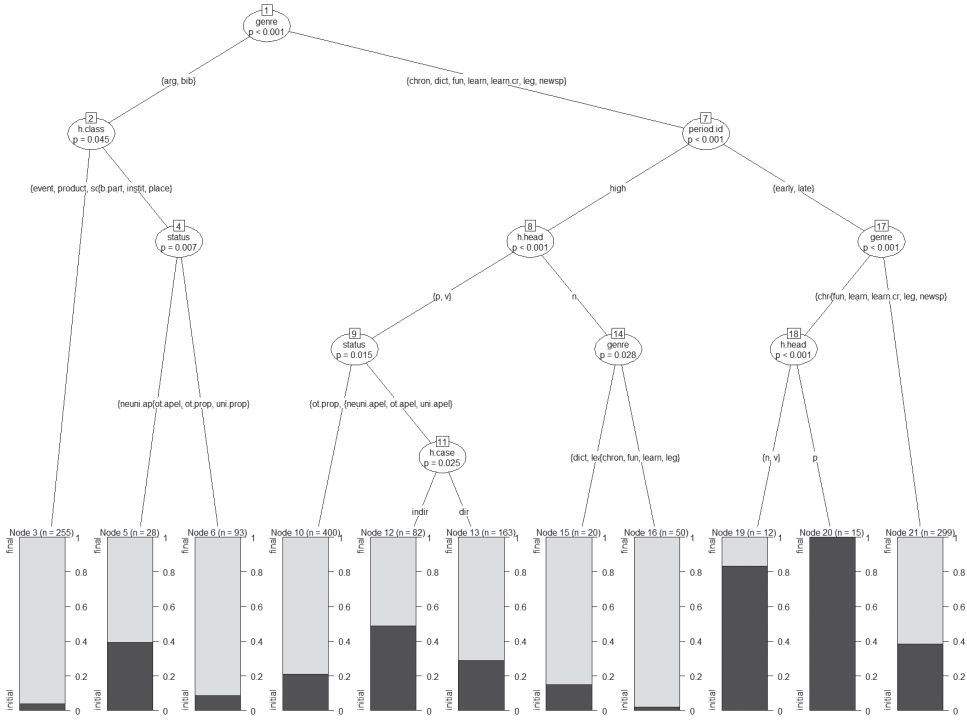
Vzniklý strom je možno vizualizovat pomocí funkce `plot()` (6) jako graf 4.

(6)
```
plot(strom.adj)
```

Nejvyšší uzel 1 ukazuje první štěpení na základě prediktoru s nejvýraznější asociací se závislou proměnnou, v každém uzlu pak vidíme jednak název prediktoru, dále hodnotu *p* z testu vztahu mezi prediktorem a závislou proměnnou, u každé z větví pak najdeme informace o dělení hodnot prediktoru do obou podsouborů. V koncových uzlech, kde žádné další štěpení nebylo na základě testu spolehlivé, najdeme „listy" stromu, tedy výsledné podskupiny pozorování vydělené na základě postupného větvení. Zde najdeme informaci o velikosti podsouboru a o poměru hodnot antepozice a postpozice v něm. Vidíme, že zatímco v některých listech je výrazné zastoupení jedné či druhé hodnoty, jinde najdeme poměrně vysoké zastoupení obou hodnot.

Ve výsledném stromu pozorujeme nejvýraznější vliv žánru. V grafu pod uzlem 2 jsou seskupeny biblické a (celkově silně podreprezentované) polemické texty, které vykazují všeobecně výraznou preferenci postpozice; výjimkou je malá podskupina konkordancí pod koncovým uzlem 5 vymezená sémantickou třídou řídícího členu a statusem posesora. U ostatních žánrů dále sledujeme vliv období: v uzlu 7 se vyděluje střední doba od obou okrajových, u nichž pozorujeme obecně vyšší výskyt antepozice; k možnému vysvětlení vlivu žánru při štěpení v uzlu 17 se vrátíme v diskuzi. Ve středním období v grafu pod uzlem 8 pozorujeme obecně vyšší frekvenci postpo-

10 Pro větší názornost uvádíme i kód použitý při analýze, anotovaný soubor je k dispozici na úložišti Open Science Framework <osf.io/bxa34>. Při analýze jsme pracovali s následujícími verzemi: RStudio, verze 1.1.383; R, verze 4.2.3; balíček *Hmisc*, verze 4.0.3; balíček *party*, verze 1.2.3.



GRAF 4: Klasifikační strom vygenerovaný na základě modelu strom.adj (viz (5)).

zice, a to s výjimkou deapelativních posesorů v přímých a ještě o něco více v nepřímých pádech (viz listy pod koncovými uzly 12 a 13).

Takto vytvořený klasifikační strom je možno dále hodnotit z hlediska jeho přesnosti. Postup spočívá v tom, že na základě vytvořeného modelu vygenerujeme predikce pro všechny datové body analyzovaného souboru a ty pak porovnáme s pozorovanými hodnotami. Dalším krokem pak může být využití stejného modelu pro nová, neanalyzovaná data anotovaná dle stejného schématu. Taková data jsme ovšem v době analýzy neměli k dispozici. Pomocí kódu v (7) zobrazíme v tabulkovém formátu predikované a pozorované hodnoty závislé proměnné pr.position z datového souboru *adjectives*, viz tabulka 1.

(7) `table(predict(strom.adj), adjectives$pr.position)`

pozice posesora	**postpozice (pozorováno)**	**antepozice (pozorováno)**
postpozice (predikováno)	1071	319
antepozice (predikováno)	2	25

TABULKA 1: Porovnání predikovaných a pozorovaných hodnot závislé proměnné z datového souboru *adjectives*.

V první buňce vidíme, že pro 1071 datových bodů byla predikována postpozice, přičemž tuto hodnotu mají datové body i v anotovaném souboru, naopak 319 datových bodů, pro něž byla predikována postpozice, má ve vzorku antepozici, jde tedy o nesprávně predikované případy. Z tabulky 1 získáme výpočtem poměru správně klasifikovaných případů klasifikační přesnost modelu, která činí 77 %, tedy o 27 % více než náhodný model, zároveň je patrné, že model je velmi přesný při předpovědi postpozice, naopak pro většinu antepozic predikuje taktéž postpozici.

V případě binární závislé proměnné je též možné pomocí funkce `somers2()` z balíčku *Hmisc* (Harrell, 2017) vypočíst index konkordance (*C*-index), který vyjadřuje poměr případů, kdy je predikovaná pravděpodobnost dané hodnoty závislé proměnné vyšší než pravděpodobnost druhé možné a ta je zároveň i hodnotou pozorovanou. Umožní nám to kód (8).

(8)
```
library(Hmisc)
strom.adj.pred <- unlist(treeresponse(strom.adj)) [c(FALSE, TRUE)]
somers2(strom.adj.pred, as.numeric(adjectives$pr.position) – 1)
```

Získáme tak *C*-index 0,74, který je v literatuře popisován jako indikátor přijatelné diskriminační síly modelu (viz Hosmer & Lemeshow, 2000, s. 162). Vidíme, že získaný model je relativně úspěšný, ponechává ovšem též značný prostor pro zlepšení. Toho je možno dosáhnout rozšířením metody o náhodné lesy.

4.3 NÁHODNÉ LESY

Analýzu pomocí klasifikačních stromů lze dále rozšířit a zpřesnit pomocí metody generování náhodných lesů (*random forests*) (Breiman, 2001), která byla vyvinuta pro zmírnění některých nedostatků CART. Metoda je založena na generování velkého množství stromů, které jsou konstruovány stejným způsobem jako jednotlivé stromy; každý strom v daném lese je přitom modelován na náhodně vybraném vzorku z analyzovaného datového souboru za použití náhodné kombinace prediktorů v modelu, jejichž počet je předem určen (standardně jde o druhou odmocninu celkového počtu prediktorů). Na základě jednotlivých stromů je možno predikovat hodnotu závislé proměnné pro pozorování, která do daného vzorku nebyla zařazena, a tak lze pak hodnotit prediktivní sílu celého modelu.

Náhodný les je možno s využitím balíčku *party* vytvořit pomocí kódu (9). Argument `ntree` přitom stanoví počet jednotlivých vytvořených stromů, `mtry` potom určuje, kolik prediktorů má být v každém modelu použito, v tomto případě generujeme 500 klasifikačních stromů, každý s využitím čtyř z celkových 12 sledovaných prediktorů.

(9)
```
set.seed(75)
les.adj <- cforest(pr.position ~ period.id + status + pr.cmplx
+ genre + alienable + h.class + h.case + h.head + h.cmplx
+ focus + nonproj + mod.other, data = adjectives, control =
cforest_unbiased(ntree = 500, mtry = 4))
```

Ačkoli takový model z logiky věci neumožňuje podobnou vizualizaci jako jednotlivé stromy, můžeme i pro vytvořený les určit jeho přesnost. To provedeme pomocí kódu (10). V prvním kroku opět získáme tabulky predikovaných a pozorovaných hodnot, poté index konkordance.

(10)
```
table(predict(les.adj), adjectives$pr.position)
les.adj.pred <- unlist(treeresponse(les.adj)) [c(FALSE, TRUE)]
somers2(les.adj.pred, as.numeric(adjectives$pr.position) – 1)
```

Použitím metody náhodného lesa získáváme model s výrazně vyšším indexem konkordance (0,85 — výborná diskriminační schopnost modelu), zároveň vzrostla i prediktivní síla modelu (správně predikuje v 81 % případů, úspěšnější je predikce antepozice, zůstává však stále relativně dosti chybová), viz tabulka 2.

	klasifikační strom		**náhodný les**	
pozice posesora	**postpozice (pozorováno)**	**antepozice (pozorováno)**	**postpozice (pozorováno)**	**antepozice (pozorováno)**
postpozice (predikováno)	1071	319	1062	261
antepozice (predikováno)	2	25	11	83
přesnost	77 %		81 %	
C-index	0,74		0,85	

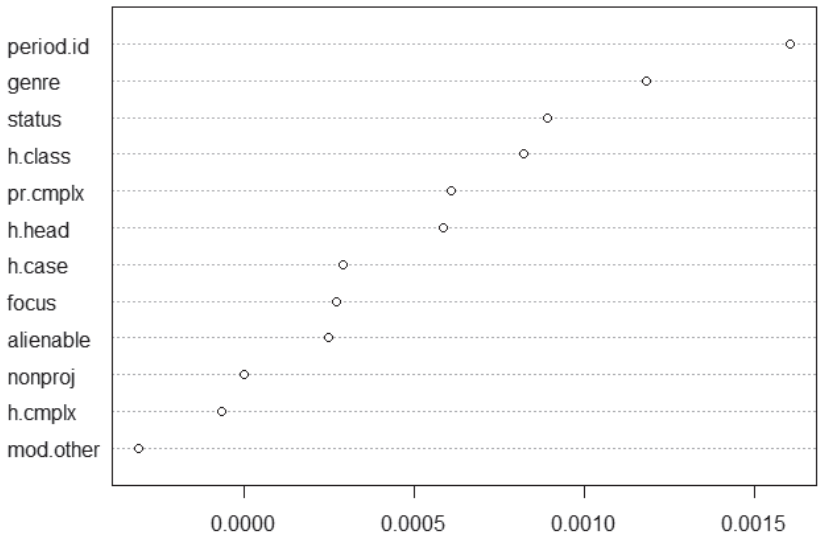
TABULKA 2: Srovnání prediktivní síly modelů klasifikačního stromu a náhodného lesa.

Vedle toho můžeme na základě metody lesa získat další důležitou informaci: model nám umožňuje posoudit význam jednotlivých prediktorů pro přesnost modelu. Metoda je založena na permutačním testu, kdy jsou hodnoty jednoho prediktoru v datovém souboru permutovány (Strobl et al., 2008). Permutovaný prediktor je spolu s ostatními nezměněnými prediktory použit pro predikci hodnot pozorování mimo vzorek (viz výše) daného lesa. Pokud se přesnost při použití permutovaného prediktoru sníží, dovodíme, že daný prediktor je důležitý, pokud naopak zůstane srovnatelná, předpokládáme, že prediktor (a jeho hodnoty) není pro daný datový soubor důležitý. Funkce `varimp()` implementovaná v balíčku *party* umožňuje tuto analýzu provést způsobem, který zároveň zohledňuje korelace mezi prediktory (argument `conditional = T` v následujícím kódu). Tyto hodnoty získáme způsobem uvedeným v (11).

(11)
```
les.varimp <- varimp(les.adj, conditional = T)
```

Nevýhodou této metody jsou vysoké komputační nároky (výpočet pro popsaný les trval na laptopu ML několik hodin). Výsledkem jsou hodnoty pro jednotlivé proměnné, které vyjadřují právě pokles prediktivní síly. Příkazem v (12) získáme graf s vypočtenými hodnotami seřazenými podle velikosti, viz graf 5.

(12)
```
dotchart(sort(les.varimp))
```



GRAF 5: Významnost jednotlivých prediktorů vyhodnocená permutačním testem.

Z grafu 5 můžeme jednak vyčíst, že jako výrazně důležitější prediktory se ukazují období, žánr, status posesora, sémantická třída řídícího členu, o trochu méně též komplexnost posesora, nadřazený řídící člen; naopak ostatní prediktory zřejmě nehrají z hlediska struktury dat příliš velkou roli. Zároveň je nutno poznamenat, že vypočtené hodnoty jsou velmi nízké. To ukazuje na to, že prediktory jsou výrazně korelované a žádný z nich nemá sám o sobě relativně větší sílu.

Celkově tedy můžeme konstatovat, že vytvořený model náhodného lesa je poměrně přesný, ač existuje prostor pro zlepšování. Dále i vzhledem k tomu, co ukázal jednotlivý klasifikační strom, vidíme, že jako nejvýznamnější se ukazují prediktory období, žánr a až následně typ posesora, sémantická třída řídícího členu, jeho řídící člen, okrajově jeho pád. Získané výsledky blíže diskutujeme v následujícím oddílu.

5. DISKUZE

5.1 SHRNUTÍ PRAVIDELNOSTÍ V DATECH

Představená analýza poukazuje na pravidelnosti, které jsou teoreticky smysluplné a lze je shrnout následujícím způsobem:

— V datech je značná míra variace, po všechna sledovaná období přitom sledujeme značný podíl postpozice.
— Významnější podíl na strukturování dat má žánr, období vzniku, dále také sémantické a referenční vlastnosti posesora a syntaktické vlastnosti řídícího členu.

— Bible jakožto samostatný žánr vykazuje výraznou preferenci postpozice.
— Depropriální posesor má obecně vyšší preferenci postpozice než posesor deapelativní.
— Ve středním, datově nejvíce reprezentovaném období sledujeme mírný nárůst preference postpozice.
— Pozice adnominální posesivní konstrukce v nepřímých pádech vykazuje nižší preferenci postpozice.
— Adnominální posesivní konstrukce závislé na substantivech vykazují výraznou preferenci postpozice.

Naše výchozí hypotéza předpokládala, že posesor, který je depropriální, s vyšší mírou aktivace (v anotačním schématu unikátní) a syntakticky lehký, bude vykazovat tendenci objevovat se v antepozici dříve a s větší pravděpodobností. Pokud izolujeme výskyty s takovou hodnotou statusu specifičnosti posesora (nerozvinuté unikátní proprium), nalezneme 513 konkordancí, z nichž je ve všech obdobích většina užita v postpozici. Při pohledu na deapelativní posesory pak zjistíme, že i zde má většina konkordancí posesora v postpozici, viz srovnání v tabulce 3. Ze tří sledovaných stupňů aktivovanosti mají potom největší podíl postpozice unikátní deapelativní posesoři.

	období		
posesor	**1. rané**	**2. střední**	**3. pozdní**
nerozvité unikátní proprium	18 / 32	60 / 330	17 / 56
unikátní apelativum	1 / 1	14 / 57	3 / 3
relační apelativum	16 / 11	71 / 161	12 / 19
ostatní apelativum	1 / 3	16 / 51	16 / 8

TABULKA 3: Srovnání nerozvinutých unikátních depropriálních posesorů se skupinami deapelativních posesorů bez ohledu na rozvitost (v jednotlivých buňkách je počet výskytů v antepozici/postpozici).

Z hlediska sémantických a diskurzně-pragmatických vlastností jsou tedy trendy v datech přesně opačné, než byl náš vstupní předpoklad. Domníváme se, že tento fakt je možno s využitím východisek popsaných v úvodu studie smysluplně interpretovat. Než tak ovšem učiníme, je nutno ještě konstatovat, že analýza pomocí náhodných lesů ukázala byť malý příspěvek komplexnosti posesora, která byla součástí výchozí hypotézy. V tomto případě situace odpovídá mezijazykovým pozorováním uvedeným v oddílu 2: z celkového počtu 150 rozvitých posesorů se většina (114) objevuje v postpozici. Abychom vysvětlili nejvýraznější pozorované trendy, které jsou zároveň rozporné s ohledem na výchozí hypotézu, blíže jsme prozkoumali některé další kategorie.

5.2 VYSVĚTLENÍ TRENDŮ ODPORUJÍCÍCH VÝCHOZÍ HYPOTÉZE: ANALÝZA LEMMAT POSESORA A LEMMAT POSESA

5.2.1 ZÁKLADNÍ ZJIŠTĚNÍ

Vzhledem k zobecnitelnosti i velkému množství hodnot jsme do modelů nezahrnuli dvě základní anotační charakteristiky lemma posesora a lemma posesa a zároveň jsme při předcházející konstrukci vzorku nezohledňovali konkrétní lexikální obsazení konkordancí. Jak ukazuje tabulka 4, při pohledu na frekvenci lemmat v obou kategoriích pozorujeme některá výrazně nadreprezentovaná lemmata. Průměrná frekvence pro posesorské lemma činí 3,7, pro řídící člen pak 3,1.

lemma posesora	**frekvence**	**lemma posesa**	**frekvence**
Kristus	197	syn	155
král	113	dům	43
hospodin	50	tělo	39
David	41	kniha	33
Mojžíš	41	pokolení	25
císař	30	slovo	23
Ježíš	27	víra	21
otec	24	jméno	20
Šalamoun	22	dcera	19
Jezukristus	21	smrt	18

TABULKA 4: 10 lemmat posesora a 10 lemmat posesa s nejvyšší frekvencí ve vzorku.

Tabulka 4 ukazuje, že posesorské lemma *Kristus* je ve srovnání se všemi ostatními výrazně frekventovanější. Pokud bychom chápali různá jména Krista jako jedno „konceptuální" lemma *Ježíš*, bylo by z celkového počtu 1417 pozorování 245 obsazeno tímto lexémem. Podobně je relativně velké množství případů obsazeno řídícím lexémem *syn*. Vzhledem k tomu, že obě tyto množiny jsou disjunktní, je celkový počet konkordancí obsazených jedním z těchto dvou lemmat 400. Při pohledu na pozici posesora pak mají obě tato lemmata výraznou preferenci postpozice (347 postpozic). Lemma *syn* se navíc často objevuje ve spojeních typu *A syn B-ův* či *synové B-ovi* (ve smyslu potomci, národ, pokolení) s depropriálním posesorem, často postavou z bible s jedinečnou referencí. Většina těchto kontextů se tak logicky vyskytuje v biblických textech. Podobně lemma *Ježíš/Kristus* je samozřejmě spjato s biblickými texty a šířeji s křesťanskou tematikou (učené náboženské texty, legendy, ale též kroniky).

Identifikovali jsme tak dvě spojení, která výrazně preferují postpozici po celý sledovaný časový úsek a jsou asociována s depropriálními posesory se značným podílem unikátních referentů, kteří jsou asociováni s biblickou či obecně náboženskou tematikou. Při pohledu na syntaktické rámce, v nichž se tato spojení objevují, je dále patrné, že především lemma *syn* se objevuje v přímých pádech s frekvencí o 10 % vyšší, než je tomu v rámci ostatních pozorování. Počet užití obou diskutovaných lemmat

v přímých pádech je zároveň v obou případech téměř dvakrát vyšší než celková frekvence druhého nejpočetnějšího lemmatu (188 konkordancí pro lemma *Ježíš*, 130 konkordancí pro lemma *syn*), obě sledovaná spojení se tedy prototypicky objevují v subjektové či objektové pozici.

5.2.2 VYSVĚTLENÍ V SOULADU S PŘIJATÝMI TEORETICKÝMI PŘEDPOKLADY

Spojíme-li zjištěná fakta s východiskovým pojetím jazyka popsaným v úvodu, lingvistikou založenou na užívání jazyka, která klade při vysvětlování jazykového chování důraz jak na sociální a obecně mimojazykový kontext, v němž se uživatelé jazyka pohybují, tak na doménově nespecifické kognitivní pochody, dostáváme poměrně zajímavý a realistický interpretační rámec pravidelností odhalených zvolenou analýzou.

Domníváme se, že je možno předpokládat existenci dvou lexikálně částečně obsazených konstrukcí s fixovaným sledem, sice *N Kristův/Ježíšův* a *syn/synové N-ův/N-ovi*, jejichž existence a užívání jsou v českých datech podmíněny sociokulturním kontextem, v němž byla literární tvorba spojena s kosmopolitními biblickými a křesťanskými tématy a v němž byla dále četná díla cele či částečně překládána, adaptována či sloužila jako vzor pro vznikající domácí, českojazyčnou literární tradici. Převažující slovosledná konfigurace u těchto konstrukcí tak může být ovlivněna jazykovým kontaktem.

Na základě uvedených zjištění a úvah pak lze předpokládat situaci, v níž tyto dvě zmiňované a jim podobné konstrukce slouží jako výrazné exemplář. Jejich relativně vysoká frekvence mohla mít za následek, že konstrukce přitahovaly jiná spojení, která jim byla podobná svými sémantickými ((unikátní) depropriální posesor) či syntaktickými vlastnostmi (subjektová či objektová pozice), případně též sdílenými textovými typy (tento potenciální efekt může do určité míry vysvětlovat vyšší zastoupení postpozice pod koncovým uzlem 21 prezentovaného stromu v grafu 5, který na rozdíl od uzlu 18 obsahuje žánry učené náboženské i nenáboženské literatury a legendy). Naopak spojení, která jsou svými sémantickými i syntaktickými vlastnostmi nejméně podobná vymezeným konstrukcím (relační a neunikátní deapelativní posesor v nepřímém pádu), vykazují vyšší míru variability, a tedy i relativně nižší míru preference postpozice. Naše interpretace tak předpokládá působení sociálních a kulturních faktorů, které vnesly do psaného jazyka určité struktury, a ty byly následně včleněny do repertoáru jazykových prostředků a mohly působit organizaci jazykového systému prostřednictvím popsaných mechanismů.

6. ZÁVĚR

V tomto textu jsme představili perspektivu, z níž je možné studovat vývoj jazyka se zohledněním představ tzv. *usage-based linguistics* o fungování jazyka na úrovni jednotlivých mluvčích i celých jejich společenství a o provázanosti jazykových i mimojazykových faktorů různé úrovně, na jejichž základě je možné interpretovat zdánlivě náhodnou variabilitu v jazykových datech. Zároveň jsme názorně ukázali, jak je možné v podobné analýze výhodně využít některých v lingvistice méně uplatňova-

ných statistických metod, v tomto případě klasifikačních a regresních stromů a náhodných lesů. V naší analýze jsme ukázali míru variace v postavení posesora vůči řídícímu substantivu v rámci adnominální posesivní konstrukce ve staré a střední češtině a pokusili jsme se tuto variaci vysvětlit postulováním dvou částečně lexikálně obsazených konstrukcí *N Kristův* a *syn N-ův*, které jako silné exempláře mohly působit na ostatní posesivní konstrukce, a poukázali jsme na jejich spojení jak s faktory sociálními a kulturními, tak s faktory jazykovými a kognitivními. Nutno ovšem dodat, že studie má i své nedostatky a poskytuje prostor pro další šetření. Především by bylo třeba prošetřit chování posesivních zájmen, která jsou z hlediska životnosti a referenčnosti nadřazena apelativům i propriím a která by svým kategoriálním charakterem a související možnou vyšší odolností vůči jazykovému kontaktu zároveň mohla lépe odrážet obecný stav v jazyce. Z hlediska zmíněného jazykového kontaktu postrádají naše data informace o překladovosti či vlivu cizojazyčných předloh; tuto složitou problematiku bude třeba v navazujících studiích ošetřit. Zároveň by bylo žádoucí na základě širšího kontextu připojit anotaci kontextového referenčního statusu posesora. Konečně by na základě našich zjištění bylo vhodné na větším vzorku deapelativních posesivních adjektiv ověřit, zda má naše interpretace širší platnost.

PODĚKOVÁNÍ

Tato studie vznikla za podpory projektu Univerzity Karlovy Progres č. 4, Jazyk v proměnách času, místa, kultury. Při vzniku práce byly využity zdroje Výzkumné infrastruktury pro diachronní bohemistiku (RIDICS, http://vokabular.ujc.cas.cz).

LITERATURA

Aikhenvald, A. Y. (2013): Possession and ownership: a cross linguistic perspective. In: A. Y. Aikhenvald (Ed.), *Possession and ownership: A cross-linguistic typology* (s. 1–64). Oxford: Oxford University Press.

Baayen, R. H. (2008): *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Behaghel, O. (1909): Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, *25*, 110–142.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M., Croft, W., Ellis, N., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009): Language is a complex adaptive system. *Language Learning*, *59*(Supplement), 1–26.

Breiman, L. (2001): Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984): Classification and Regression Trees. Belmont, CA: Wadsworth.

Bybee, J. L. (1985): *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam: Benjamins.

Bybee, J. L. (2006): From usage to grammar: the mind's response to repetition. *Language*, *82*, 711–733.

Daneš, F. (1985): *Věta a text: studie ze syntaxe spisovné češtiny*. Praha: Academia.

Diessel, H. (2017): Usage-Based Linguistics [online]. In M. Aronoff (Ed.), *Oxford Research Encyclopedia of Linguistics*. New York, NY: Oxford University Press. Dostupné z: <https://doi.org/10.1093/acrefore/9780199384655.013.363>.

Gebauer, J. (2007/1929): *Historická mluvnice jazyka českého IV. Skladba*. (2. vyd. 2007;

1. vyd. 1929, ed. F. Trávníček.) Praha: Academia.

Gries, S. T., & Hilpert, M. (2012a): Variability-based neighbor clustering: a bottom-up approach to periodization in historical linguistics. In T. Nevalainen & E. C. Traugott, (Eds.), *The Oxford Handbook on the History of English* (s. 134–144). Oxford: Oxford University Press.

Gries, S. T., & Hilpert, M. (2012b): Variability-based Neighbor Clustering: A bottom-up approach to periodization in historical linguistics. Companion website [online]. T. Nevalainen & E. C. Traugott, (Eds.), *The Oxford Handbook on the History of English.* Oxford: Oxford University Press. Dostupné z: <http://global.oup.com/us/companion.websites/fdscontent/uscompanion/us/static/companion.websites/nevalainen/Gries-Hilpert_web_final/vnc.individual.html>.

Harrell, F. E. (2017): *Hmisc: Harrell Miscellaneous. R package version 4.0-3*. Dostupné z WWW: <https://CRAN.R-project.org/package=Hmisc>.

Haspelmath, M. (2008): Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, 19, s. 1–33.

Hawkins, J. A. (2004): *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hilpert, M. (2013): *Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge University Press.

Hopper, P. J. (1987): Emergent grammar. In J. Aske, N. Beery, L. Michaelis & H. Filip (Eds.), *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, *13* (s. 139–157). Berkeley, CA: Berkeley Linguistics Society.

Hosmer, D. W., & Lemeshow, S. (2000): *Applied Logistic Regression*. New York, NY: Wiley.

Hothorn, T., Hornik, K. & Zeileis, A. (2006): Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Křivan, J. (2014): The role of information structure in Czech possessive constructions. In J. Emond & M. Janebová (Eds.), *Language Use and Linguistic Structure: Proceedings of the Olomouc Linguistics Colloquium 2013* (s. 211–227). Olomouc: Univerzita Palackého.

Kučera, K., Řehořková, A., & Stluka, M. (2015): *DIAKORP: Diachronní korpus, verze 6 z 18. 12. 2015*. Praha: Ústav Českého národního korpusu FF UK. Cit. 9. 7. 2016. Dostupné z WWW: <http://www.korpus.cz>.

Langacker, R. W. (1988): A Usage-Based Model. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics* (s. 127–161). Amsterdam: Benjamins.

Levshina, N. (2015): *How to do Linguistics with R. Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins.

Milin, P., Divjak, D., Dimitrijević, S., & Baayen, R. H. (2016): Towards cognitively plausible data science in language research. *Cognitive Linguistics*, *27*(4), 507–526.

O'Connor, C., Maling, J. & Skarabela, B., (2013): Nominal categories and the expression of possession: A cross-linguistic study of probabilistic tendencies and categorical constraints. In K. Börjars., D. Denison & A. Scott (Eds.), *Morphosyntactic categories and the expression of possession* (s. 89–121). Amsterdam: John Benjamins.

Poplack, S., & Cacoullos, R. T. (2015): Linguistic emergence on the ground: A variationist paradigm. In B. MacWhinney & W. O'Grady (Eds.), *The Handbook of Language Emergence* (s. 267–291). Chichester: Wiley-Blackwell.

R Core Team (2017): *R: A language and environment for statistical computing* [online]. R Foundation for Statistical Computing: Wien. Dostupné z: <http://www.R-project.org/>.

Rezaee, A. A., & Golparvar, S. E. (2016): The Sequencing of Adverbial Clauses of Time in Academic English: Random Forest Modelling of Conditional Inference Trees. *Journal of Language Modelling*, *4*(2), 225–244.

RStudio Team (2016): RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. Dostupné z WWW: <http://www.rstudio.com>.

Seiler, H. (1983): *Posession as an operational dimension of language*. Tübingen: Günter Narr.

Silverstein, M. (1976). Hierarchy of Features and Ergativity. In R. M. W. Dixon (Ed.),

Grammatical Categories in Australian Languages (s. 112–171). Canberra: Australian National University.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., & Zeiles, A. (2008): Conditional variable importance for Random Forests. *BMC Bioinformatics*, *9*(307). Dostupné z WWW: <https://doi.org/10.1186/1471-2105-9-307>.

Tagliamonte, S. A., & Baayen, R. H. (2012): Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change*, *24*(2), 135–178.

Vachek, J. (1972): Ještě k osudu českých posesívních adjektiv (Glosa k pohybu v českém tvarosloví). *Slovo a slovesnost*, *33*, 146–148.

Vokabulář webový (2006–2016a): *Staročeská textová banka*. Praha: Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR. Dostupné z: <http://vokabular.ujc.cas.cz/banka.aspx>.

Vokabulář webový (2006–2016b): *Středněčeská textová banka*. Praha: Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR. Cit. Dostupné z: <http://vokabular.ujc.cas.cz/banka.aspx>.

Wickham, H. (2009): *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer.

Widmer, M., Auderset, S., Nichols, J., Widmer, P., & Bickel, B. (2017): NP recursion over time: evidence from Indo-European. *Language*, *93*(4), 799–826.

Wolk, C., Bresnan, J., Rosenbach, A., & Szmrecsanyi, B. (2013): Dative and Genitive Variability in Late Modern English: Exploring Cross-constructional Variation and Change. *Diachronica*, *30*(3), 382–419.

Jan Křivan | Ústav pro jazyk český, v. v. i.
<krivan@ujc.cas.cz>

Michal Láznička | Ústav Blízkého východu a Afriky FF UK
<michal.laznicka@ff.cuni.cz>