Univerzita Karlova

Filozofická fakulta

Ústav obecné lingvistiky

Bakalářská práce

Eliška Konývková

Výpočet gramatického statusu: kvantitativní analýza čínských textů

Calculation of grammatical status: quantitative analysis of Chinese texts

Praha 2024

Vedoucí práce: PhDr. Jiří Milička, PhD.

Poděkování: Děkuji vedoucímu práce PhDr. Jiřímu Miličkovi, PhD. za čas a cenné připomínky k práci.

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, že jsem řádně citovala 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 15. července .....

Abstrakt:

Tato studie opakuje výzkum Linlin Sun & Davida C. Saavedry zaměřený na určování gramatického statusu jednotek v čínštině pomocí kvantitativních metod. V návaznosti na jejich práci používáme binární logistický regresní model pro výpočet skóre gramatického statusu vybraných lexikálních jednotek. Jako zdroj dat nám slouží Lancasterský korpus čínštiny (LCMC), který obsahuje současné standardní čínské texty.
V práci podrobně popisujeme metriky a modelovací přístup použitý k odvození skóre gramatického stavu a porovnáváme naše výsledky s výsledky autorů původní studie.
V naší analýze dále posuzujeme vhodnost vybraných metrik, hodnotíme přesnost predikce binárního logistického regresního modelu a analyzujeme, kde se po přiřazení gramatického statusu jednotky nachází na lexikálně - gramatické škále, a zda tomu odpovídá jejich „tradiční“ zařazení do slovních kategorií.

Klíčová slova: gramatický status, korpus, čínština, kvantitativní lingvistika, replikační studie

Abstrakt:

This study replicates Linlin Sun & David C. Saavedra's research on determining the grammatical status of units in Chinese using quantitative methods. Following their work, we use a binary logistic regression model to calculate the grammatical status scores of selected lexical units. As a source of data, the Lancaster Corpus of Chinese (LCMC), which contains contemporary standard Chinese texts, is used.
In this paper, we detail the metrics and modeling approach used to derive the grammatical status scores and compare our results with those of the authors of the original study.
In our analysis, we further assess the appropriateness of the selected metrics, evaluate the accuracy of the binary logistic regression model's prediction, and analyze where units are on the lexico-grammatical scale after grammatical status assignment, and whether this corresponds to their "traditional" placement in word categories.

Key words: grammatical status, corpora, Chinese, quantitative linguistics, replication study

Obsah

Seznam zkratek

LCMC Lancasterský korpus čínštiny
ČLR Čínská lidová republika
POS part of speech, slovní druh

1. Úvod

Práce staví na článku Sun a Saavedry (Sun & Saavedra, 2020), ve kterém autorka a autor za pomoci modelu binární logistické regrese stanovili pro vybrané lexikální jednotky tzv. gramatický status. Jako zdroj dat použili Lancasterský korpus čínštiny (LCMC), který obsahuje texty psané poměrně současnou standardní čínštinou. V článku detailně představují použité metriky a způsob modelování, kterým k jednotlivým skóre gramatického statusu dospěli. Výsledky pak porovnávají s výsledky studie David C. Saavedry, která měla podobný cíl, avšak byla zaměřena na angličtinu, autor pracoval s daty získanými z Britského národního korpusu (BNC). O tomto procesu pojednává Saavedra ve své dizertační práci (Correia Saavedra, 2019), která byla též jedním ze zdrojů pro tuto práci, jak pro převzetí praktických nástrojů (scriptu) k výpočtu metrik, tak pro teoretické pozadí této bakalářské práce – replikační studie.

Autor a autorka v článku nejdříve krátce pojednávají o fenoménu gramatikalizace, v další části představují metodu. V této kapitole se soustředí hlavně na popis jednotlivých použitých metrik, na uvedení čtenářů do binární logistické regrese, a také nastiňují parametry binárního logistického modelu. Dále charakterizují Lancasterský korpus čínštiny a uvádí, jakým způsobem sestavovali databázi lexikálních a gramatických jednotek k ohodnocení. Při diskuzi nad výsledky autoři vyhodnocují vhodnost použití vybraných metrik, evaluují přesnost predikce modelu binární logistické regrese a následně uvádí seznam jednotek s přiřazeným gramatickým statusem. Ten slouží jako podklad pro širší analýzu, ve které se dotýkají jak konkrétních slovních jednotek, tak i např. problematiky slovních druhů v čínštině, či hodnotí, zda jimi navržené skóre gramatického statusu vykazuje dobré výsledky, a zda mohou tak v otázce lexikálního – gramatického dělení, tradičně binárního, jejich gradientní hodnoty přinést nové odpovědi.

Tato práce částečně kopíruje strukturu článku. Na začátku stručně uvedeme teoretická východiska autorů a shrneme teze a cíle studie. Poté se budeme podrobněji věnovat metrikám a základním naměřeným hodnotám potřebným pro výpočet. Provedeme srovnání výchozích hodnot s autory studie. Než přejdeme k výsledkům výpočtu, popíšeme jiná data, tedy konkrétně LCMC a databázi osmi set slovních jednotek, kterým bylo přiřazeno skóre gramatického statusu. Za poskytnutí této databáze autorům Linlin Sun a Davidovi C. Saavedrovi moc děkuji. Následně uvedeme přehled dělení slovních druhů v čínštině a porovnáme je s kategorizací slov podle autorů článku. V kapitole 4. se budeme věnovat hlavnímu tématu studie, a tedy možnosti výpočtu gramatického statusu. Výsledky analyzujeme a porovnáme s autory. V závěrečné části se dotkneme dalších možných způsobů modelování dat pro výpočet gramatického statusu.

1.2 Východiska autorů studie

Pojmu gramatikalizace, jak ve smyslu fenoménu, tak ve smyslu jazykového procesu bylo přiřazeno mnoho významů a definicí. Sun a Saavedra tedy hned na začátku článku poměrně jasně vymezují pro svoji studii užší rámec. S odkazem na Hoppera & Traugottovou (Hopper & Traugott, 2002) pojem gramatikalizace chápou v širší a více synchronní perspektivě, jako rámec, ve kterém je v jazyce možné zkoumat způsoby užití gramatických morfémů a konstrukcí, a je také možné zkoumat chování těchto gramatických jednotek a jejich vliv na jazyk obecně. Již zmínění Hopper & Traugottová, ve své knize „Gramatikalizace" otevírají polemiku nad striktním oddělováním synchronního a diachronního přístupu k jazyku. Konkrétně odkazují ke konceptualizaci synchronní dimenze jazyka, jako k něčemu stabilnímu a homogennímu, se zaběhlým systémem jazykových pravidel, k něčemu, co stojí v protikladu k diachronní dimenzi jazyka, která je vnímána jako soubor změn, které tvoří pojítko mezi synchronní fází jazyka a jeho fázemi následnými. Takový přístup je blízký právě i autorům, dále v článku uvádějí, že jejich pohnutky k takovéto studii jsou založeny na předpokladu postupného přesunu jednotky z jedné kategorie do druhé na lexikálně-gramatickém kontinuu, která prochází postupným procesem zvyšující se gramatičnosti. Na této škále tvoří výrazy jako např. sufixy, klitika a částice její gramatické zakončení, zatímco prototypické lexikální výrazy jako např. zájmena a slovesa jsou umístěna na ten opačný, lexikální konec škály.

Autoři vyslovují více cílů, sepíšeme je zde pod sebe:

- skrze synchronní na korpusu založenou analýzu určit, jaké jednotky jsou v čínštině považovány za gramatické a jaké nikoli
- jak jsou gramatické jednoty uspořádány na základě svého gramatického statusu
- navrhnout, jak skrze kvantitativní korpusovou analýzu měřit gramatický status jednotek v čínštině
- porovnat výsledky s podobnou studií, ve které se analyzovala angličtina
- předestřít, do jaké míry jsou proměnné modelu[1] užitečné v určování stupně gramatického statusu napříč těmito jazyky (čínštinou a angličtinou)

V této práci je primárním cílem replikovat metodu autorů, zároveň se do tohoto cíle bakalářské práce nutně promítají i ty jejich. Zprostředkovaně se tedy budeme soustředit pouze na první dva vytyčené cíle, tedy se pokusíme pomocí metrik jako nezávislých proměnných a modelu binární logistické regrese určit, jaké jednotky jsou v čínštině považovány za gramatické a jaké nikoli. Pokud se tento krok podaří, podíváme se také na uspořádání na základě jejich gramatického statusu.

[1] Metoda v aplikované matematice, která zahrnuje, numerická data, rovnice a statistické metody pro předpovídání výsledků, https://www.studysmarter.co.uk/explanations/math/applied-mathematics/quantitative-modeling/

1.3 Popis základních metrik

Autorka a autor článku staví současný přístup na kvantitativním modelu pro určování indexu gramatikalizace, který vyvinul Saavedra (Correia Saavedra, 2019) na angličtině. Tento model logistické regrese pracoval při predikci s pěti kritérii. Vzhledem k tomu, že v čínštině není fonémická délka relevantní, autoři se rozhodli ve studii použít čtyři z nich. Jedná se o následující parametry a z nich odvozené proměnné.

a) frekvence tokenů (*token frequency*)

Určování frekvence, tedy četnosti výskytu zkoumané jednotky v korpusu můžeme považovat za jednu ze základních metrik v kvantitativní a korpusové lingvistice (Zipf, G.K., 1932). Avšak například Gries (2008) upozorňuje, že je k této metrice potřeba přistupovat obezřetně, protože obzvlášť vysoká frekvence může zkreslovat výsledky (Gries, 2008). Pokud používáme tuto metriku pro porovnání rozdílných korpusů je nutné aplikovat její „relativní variaci" – frekvenci tokenů vydělenou celkovým rozsahem korpusu.
V článku autoři frekvenci nijak neupravují (ve smyslu škálování). Motivací zapojení této metriky k vyhodnocení gramatického statusu jednotky je předpoklad, že jednotky s vysokou frekvencí bývají častěji gramatické. V textu odkazují na Bybee (2003), která uvádí, že „jednou z nejvíce nápadných vlastností gramatických morfémů a konstrukcí ve kterých se vyskytují, je jejich, v porovnání s typickými lexikálními jednotkami, mimořádně vysoká frekvence v textu" (str. 1).

b) rozmanitost kolokátů (*collocate diversity*)

Vedle frekvence se k základním metrikám dále řadí četnost kolokátů (Gries & Ellis, 2015), zde autoři používají termín „rozmanitost kolokátů". Vychází z předpokladu, že volnost jednotky k zapojení do různých kontextů můžeme vyhodnotit podle toho, kolik různých kolokátů se vedle ní vyskytuje.
Tato metrika je zařazena do vyhodnocení gramatického statusu, protože autoři předpokládají, že gramatické jednotky mají díky svojí větší významové obecnosti možnost vyskytovat se ve společnosti vyššího počtu různorodých kolokátů než lexikální jednotky (Sun & Saavedra, 2020). Z tohoto parametru „rozmanitosti kolokátů" vyvozují autoři tři konkrétní metriky.

Naměřené skóre (počet unikátních kolokátů jednotky) autoři dělí celkovou frekvencí jednotky. U *CollocDiv4-4* je skóre dále děleno osmi, tak aby vyjadřovalo skóre pouze pro jednu pozici.

- *CollocDiv1L* = počet kolokátů na pozici 1 vlevo/ celková frekvence jednotky
- *CollocDiv1R* = počet kolokátů na pozici 1 vpravo/ celková frekvence jednotky
- *CollocDiv 4-4* = počet kolokátů na pozicích 1– 4 vpravo i vlevo/ celková frekvence jednotky/ 8

c) rozmanitost koligátů (*colligate diversity*)

Hodnoty rozmanitosti koligátů počítali autoři odlišně než hodnoty výše zmíněných kolokátů. Kolokáty a koligáty jsou si podobné, s tím rozdílem, že u koligátů nebereme v potaz slovní význam jednotky, ale pouze jeho slovní druh (POS, *part of speech*). Pro popis jednotlivých značek slovních druhů viz kapitolu (2.1 Korpus). Motivací autorů zapojit tuto metriku do vyhodnocení gramatického statusu jednotky byl předpoklad, že co se týká vazby s koligáty, gramatické jednotky naopak nemají tolik volnosti jako lexikální jednotky. Jejich počet koligátů by tedy měl být nižší, a následné přepočítané skóre rozmanitosti koligátů by tedy mělo být vyšší. U koligátů autoři nevyhodnocují koligáty v rozsahu 1-4 vlevo a vpravo od zkoumané jednotky, pouze na pozici 1 vlevo a vpravo.

Pro výpočet je nutné znát hodnotu nejčastěji se vyskytujícího koligátu, tato hodnota se pak dále dělí celkovým počtem kolokátů (na příslušné pozici) jednotky.

- *MaxColliPercent1L* = počet nejfrekventovanějšího koligátu na pozici 1 vlevo/ počet kolokátů na pozici 1 vlevo
- *MaxColliPercent1R* = počet nejfrekventovanějšího koligátu na pozici 1 vpravo/ počet kolokátů na pozici 1 vpravo

d) proporční odchylka (*deviation of proportions*)

Skrze proporční odchylku, kterou jako metriku navrhl lingvista Gries (2008), můžeme zkoumat pravidelnost výskytu slov v textu. Autoři ji zařadili na základě předpokladu, že gramatické jednotky jsou v textu distribuované rovnoměrněji než jednotky lexikální (Sun & Saavedra, 2020). Zároveň dodávají, že rovnoměrnější distribuci budou mít také slova s vyšší frekvencí.

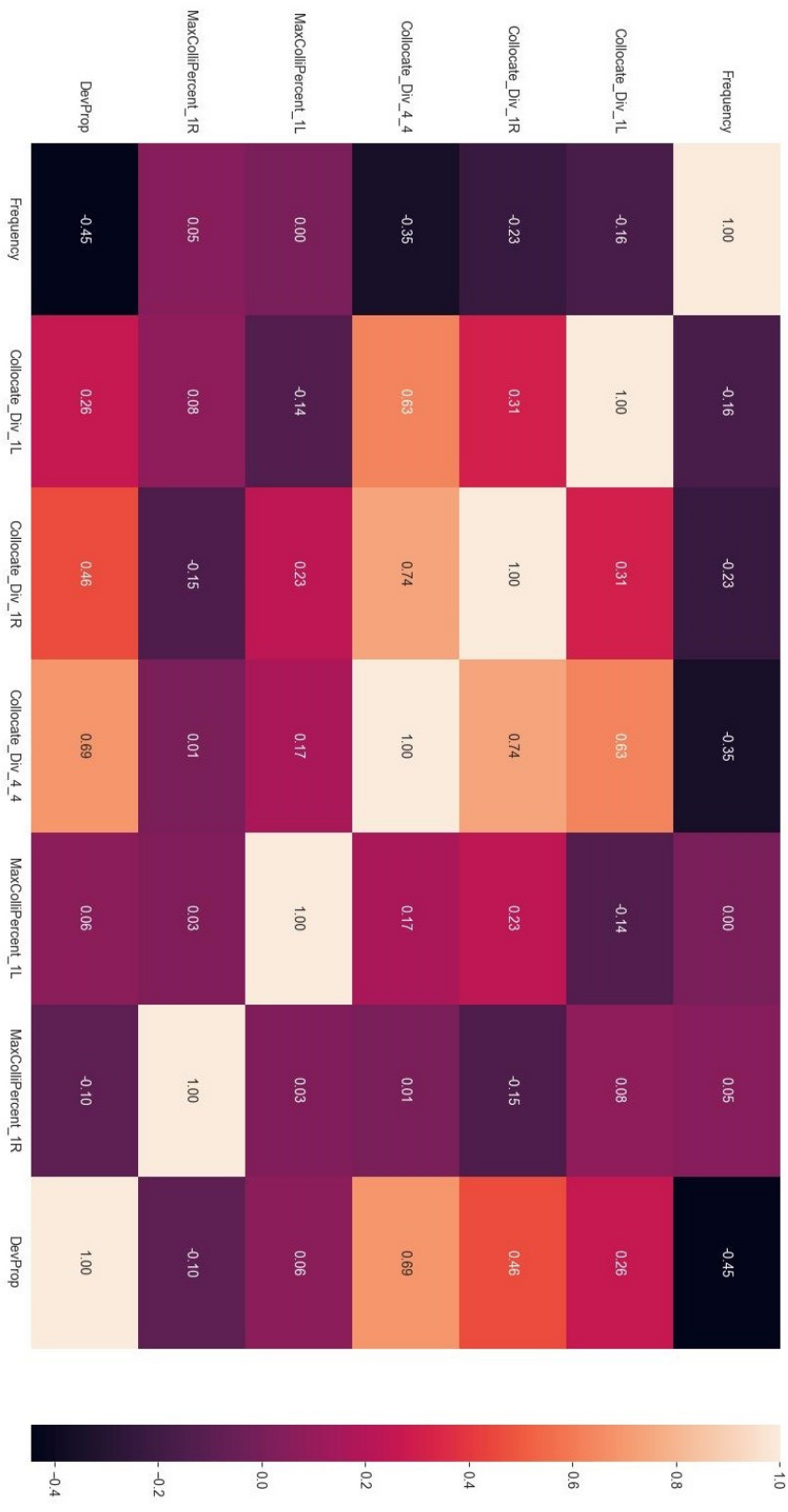
Pro vypočítání finální hodnoty je potřeba nejdříve určit absolutní rozdíl mezi teoretickou distribucí slova a reálnou distribucí slova v předem stanoveném kousku (*chunk)* korpusu. Tyto absolutní rozdíly v jednotlivých kusech pak sečteme a vydělíme dvěma.

- *DevProp* = součet (absolutní distribuce – teoretická distribuce) pro každý kus korpusu/ 2

Například pokud by věta „*I am very very happy I study linguistics, I am*.“ představovala náš korpus, a ten bychom rozdělili na dva kusy „*I am very very happy*“ a „*I study linguistics I am*“. Pak by součet absolutních rozdílů mezi teoretickou a reálnou distribucí slova „*very*“ vycházel 2. První kus korpusu: 2 (reálná distribuce) – 1 (teoretická distribuce) = 1. Druhý kus korpusu: 0 (reálná distribuce) – 1 (teoretická distribuce) = 1. Součet následně vydělíme dvěma. Výsledek pro slovo „very“ = 1, neboli „nejvíce nevyrovnaná distribuce slova“.

Autoři ve svojí studii stanovili, že jeden *chunk* je rovný tisíci tokenů. Hodnota této metriky se pohybuje mezi 0 a 1, pokud má jednotka naměřenou hodnotu *DevProp* blízko nule, znamená to, že je v korpusu distribuovaná rovnoměrně a naopak.

V grafu č. (1) můžeme pozorovat, že jako jediné jsou spolu korelované metriky *CollocDiv4-4* a *CollocDiv1L, CollocDiv4-4* a *CollocDiv1R*. Stejně tak to vychází i autorům. Tuto korelaci očekáváme, protože metrika rozmanitosti kolokátů, která vypočítává pozice 1 – 4 na každé straně klíčového slova v sobě zahrnuje i kolokáty na pozici 1 od klíčového slova zleva i zprava.



Graf č. 1 – korelace metrik

1.4 Analýza základních hodnot

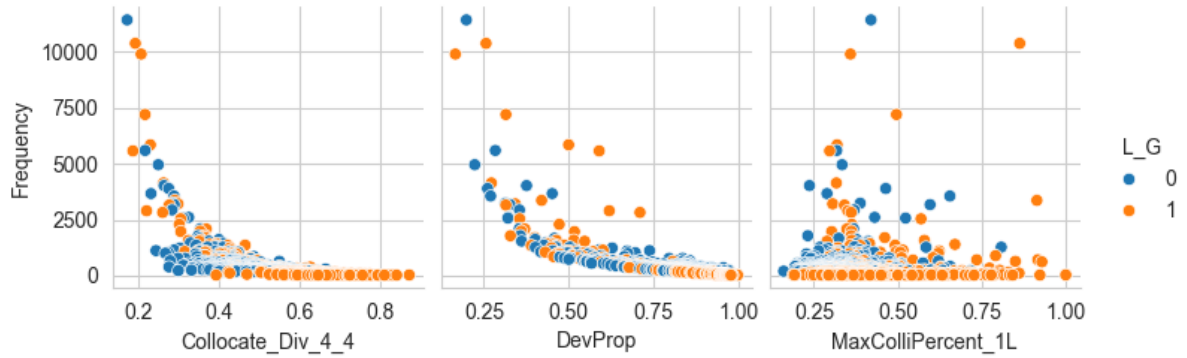
Autoři předkládají u několika metrik tezi (viz výše u popisu každé metriky), jakou tendenci by gramatické jednotky měly mít. Konkrétně předpokládají, že gramatické jednotky budou mít:

- vyšší frekvenci
- vyšší skóre rozmanitosti kolokátů
- nižší skóre rozmanitosti koligátů
- nižší skóre proporční odchylky

Pokud nahlédneme do grafu č. (2), tak zjistíme, že souvislosti nejsou tak přímočaré, něco z něj však vyčíst můžeme. Z dat do grafu byla dočasně odstraněna částice 的 (*de*, přivlastňovací/atributivní částice), která svojí frekvencí (50 831 tokenů) dalece převyšuje ostatní zkoumané jednotky a graf tak nebyl přehledný. Na ose *y* je vynesená frekvence, na ose *x* pak hodnoty metriky rozmanitosti kolokátů (4 pozice na každé straně), proporční odchylky a rozmanitosti koligátů (1 pozice vlevo). Body jsou obarveny podle lexikální (0) nebo gramatické (1) kategorie, do které jednotka spadá.

Ve všech třech částech grafu č. (2) vidíme úplně nahoře modře obarvené sloveso identifikace 是 (shì), přestože je v základní databázi zařazeno do lexikální kategorie, při analýze výsledků gramatického statusu uvidíme, že jeho extrémně vysoká frekvence posouvá při hodnocení logistickou regresí sloveso do kategorie gramatické. Poté následuje 了 ve funkci slovesné přípony (*le*, slovesná přípona), 在 ve funkci předložky (*zài*, „v“) a spojka 和 (*hé*, „a“). Můžeme tedy říci, že čtyři z pěti (když započítáme i vyňatou částici 的) nejfrekventovanějších jednotek spadají do gramatické kategorie. Poté už se začínají objevovat i lexikální jednotky.

V grafu č. 2 vpravo jsou na ose x vyneseny hodnoty rozmanitosti koligátů, u odlehlé slovesné přípony 了 se můžeme pokusit odvodit, proč dosahuje tak vysokého skóre. U *MaxColliPercent_1L* vidíme částici vpravo nahoře, vzhledem k tomu, že se na pozici koligátů 1 vlevo vyskytuje u této částice výhradně sloveso, je její skóre rozmanitosti koligátů podle předpokladu velmi vysoké (vysoké skóre rozmanitosti koligátů odpovídá nízkému počtu koligátů).



Graf č. 2 – frekvence s dalšími proměnnými

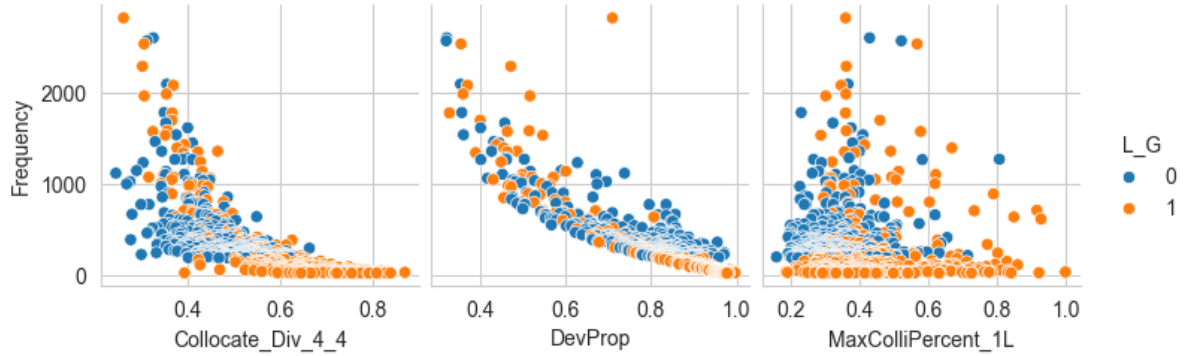
Data v grafu č. (3) jsou zobrazena bez dvaceti nejfrekventovanějších jednotek, díky tomu můžeme lépe porovnat distribuci jednotek s nižší frekvencí.
Jak bylo zmíněno výše, co se týká rozmanitosti koligátů, předpoklad je takový, že gramatické jednotky budou dosahovat vyššího skóre. Vidíme (v interaktivním grafu, zde bohužel nejsou body naprosto zřetelné), že nad hranici skóre 0, 6 se u koligátů dostalo šest lexikálních jednotek. Při bližším prozkoumání je zřejmá jejich specifičnost, a zkusíme vysvětlit, z jakého důvodu se dostaly na opačný konec škály, než bylo předpokládáno.

Jedná se dva lokativy 里 (lǐ, „v, uvnitř“) a 之间 (zhījiān, „mezi“), tři slovesa 完（wán, „spotřebovat, skončit“), 出来 (chūlái, „vyjít, objevit se“), 起来 (qǐlái, stoupnout si, povstat) a jedno substantivum 年代 (niándài, „dekáda století“).

Pokud si představíme, který slovní druh se nejčastěji vyskytuje na pozici 1 vlevo u výše zmíněných lokativů, tak zjistíme, že se jedná téměř výhradně o substantivum. Zmíněná slovesa najdeme často na pozici výsledkových (完) nebo směrových (出来, 起来) modifikátorů slovesa, které jim předchází. Jejich koligáty na pozici 1 vlevo budou tedy do velké míry slovesa. Se substantivem 年代 se z levé strany nejčastěji pojí číslovky. V grafu č. (3) se lexikální jednotky (modrá barva) seskupují nevíce mezi hodnotami 0, 2 až 0,5 metriky *MaxColliPercent_1L*.

Dále v grafu č. (3) (s odstraněnými 20 nejfrekventovanějšími jednotkami) vidíme, že u proporční odchylky (*DevProp*) je v levé části grafu (nízké skóre proporční odchylky značí rovnoměrnou distribuci) vykreslena směs lexikálních i gramatických jednotek. Popíšeme prvních pět s nejnižším skóre. Lexikální jednotky s nejnižším skóre jsou příslovce 都 (dōu, „všichni, všechno“), 又 (yòu, „znovu“) a sloveso 到 (dào, „přijet, přijít“), z gramatických jednotek mezi nimi najdeme předložku 从 (cóng, „z (místa)“ a numerativ 个 (ge, obecný numerativ). Kategorie adverbií je (autory studie) zařazena do lexikální kategorie, tedy by podle předpokladu měla mít vyšší skóre proporční odchylky. V případě uvedených adverbií se jejich distribuce zdá být rovnoměrná.

Na straně vysokého skóre proporční odchylky se nachází skupina lexikálních jednotek (mezi 0,7 a 1) a také gramatické jednotky s velmi nízkou frekvencí.



Graf č. 3

V grafu č. (3) na levé straně jsou vykresleny hodnoty pro metriku rozmanitosti kolokátů v rozsahu 1 – 4 pozice vlevo i vpravo. Na levé straně, kde jsou vyneseny jednotky s nejnižším skóre, najdeme gramatické i lexikální jednotky. Jednotkou s nejvyšší frekvencí a nízkým skóre rozmanitosti kolokátů je zájmeno 他 (tā, „on“). Po nahlédnutí přímo do korpusu vidíme, že jako kolokáty se s tímto zájmenem nejčastěji pojí další zájmena a také frekventovaná slovesa (např. 说, shuō, „mluvit, říkat“) nebo (知道, zhīdao, „vědět“). Dále má nízké skóre skupina substantiv 经济 (jīngjì, „ekonomika“), 社会 (shèhuì, „společnost“), 社会主义 (shèhuìzhǔyì, „socialismus“) a sloveso 发展 (fāzhǎn, „rozvíjet, „vyvíjet““). Při nahlédnutí do korpusu a prozkoumání těchto substantiv a jejich širších kolokací zjistíme, že se tyto tři substantiva a sloveso vyskytují poměrně často společně (v různých kombinacích) jako frekventované kolokace. Např. 经济发展 („ekonomický rozvoj“), 社会发展 („společenský rozvoj“), případně spolu s kolokáty ustavují konkrétní termín, např. 中国特色的社会主义 (zhōngguó tèsè de shèhuìzhǔyì, „socialismus s čínskými rysy“), jednotka se tak často pojí přesně s těmito kolokáty, což ji při výpočtu determinuje k nízkému skóre rozmanitosti kolokátů, opačného, než bychom u substantiv očekávali.

Jako další, snad více „přímočará“ možnost analýzy tezí o tendencích jednotlivých metrik, se nabízí vypočítat průměrné hodnoty metrik pro lexikální a gramatickou kategorii. V tabulce č. (1) vidíme, že průměrné hodnoty metrik tyto teze přímočaře nepotvrzují.

	Frequency	**Collocate_Div_1L**	**Collocate_Div_1R**	**Collocate_Div_4**
LEX	526.9875	0.52963	0.52962	0.460802
GRAM	510.85	0.64887	0.69257	0.597390
	DevProp	**MaxColliPer_1L**	**MaxColliPer_1R**	
LEX	0.75718	0.33381	0.37942	
GRAM	0.84006	0.44561	0.39190	

Tabulka č. 1

Průměrná frekvence je oproti očekávání vyšší pro lexikální jednotky a skóre proporční odchylky je oproti očekávání vyšší pro gramatické jednotky. Předpokládáme, že se zde opět projevila role frekvence, která hodnoty posunula mimo očekávaný rozsah. Na grafech č. (2) a č. (3) výše jsme mohli vidět, že v databázi je několik gramatických slov s velmi vysokou frekvencí a mnoho gramatických slov, jejichž hodnota frekvence je menší než 50. To by mohlo vysvětlovat opačnou tendenci v hodnotách proporční odchylky – jednotka, která má velmi nízkou frekvenci těžko v rozsáhlém korpusu dosáhne na pravidelnou distribuci.
Průměrné výsledky hodnot u rozmanitosti kolokátů odpovídají tezím o dost lépe. Skóre rozmanitosti kolokátů je, nehledě na počet pozic či směr od klíčového slova, vyšší u skupiny gramatických jednotek.
U průměrných hodnot koligátů, které odpovídají vzneseným tezím (gramatické jednotky by měly mít vyšší skóre rozmanitosti koligátů, tzn. nižší počet koligátů), můžeme pozorovat typologický vliv čínštiny na výslednou hodnotu koligátů na pozici 1 vlevo a na pozici 1 vpravo. Na pozici vlevo od klíčového slova se nachází podstatně méně koligátů než na pravé straně.

1.5 Porovnání základních hodnot

První krok směrem k výpočtu gramatického statusu spočíval ve získání základních hodnot – frekvence, kolokátů, koligátů a proporční odchylky. Přesnost naměření těchto hodnot byla zásadní pro získání uspokojivého skóre gramatického statusu.
V článku autorů jsou uvedeny tyto základní hodnoty jako příklad u výpočtu metrik pouze u tří slov. Jedná se o numerativ 元 (yuán, numerativ pro peníze), částici 的 (de, atributivní částice) a substantivum 方面 (fāngmiàn, „aspekt“). V tabulce č. (2) a (3) níže uvádím porovnání těchto hodnot v článku a hodnot, ze kterých vycházím v této BP. Jak můžeme v tabulce č. (2) vidět, výsledky se stoprocentně nestýkají.

元	**Freqency**	**Kolok_1L**	**Kolok_1R**	**Kolok_4**	**Kolig_1L**	**Kolig_1R**
Autoři	604	402	336	2033	389 (číslovky)	78 (substantiva)
BP	604	359	335	1920	305 (číslovky)	80 (substantiva)
Absolutní Rozdíl	0	43	1	113	84	2

Tabulka č. 2 - porovnání základních hodnot

元	**Frequency**	**Collocate_Div_1L**	**Collocate_Div_1R**	**Collocate_Div_4**
Autoři	604	0.628	0.525	0.397
BP	604	0.560	0.523	0.375
元	**DevProp**	**MaxColliPer_1L**	**MaxColliPer_1R**	
Autoři	---	0.968	0.232	
BP	0.806	0.85	0.239	

Tabulka č. 3 - porovnání základních hodnot

	Dev_Prop 的	**Dev_Prop 方面**
Autoři	0.128	0.659
BP	0.129	0.665

Tabulka č. 4 - porovnání základních hodnot

Ještě kromě naměřené frekvence slovesa identifikace 是, která odpovídá, nemáme k dispozici další základní hodnoty k porovnání. Co se týká neodpovídajících hodnot pro numerativ 元, domníváme se, že by mohla být způsobena striktněji nastaveným rozsahem pro vyhledávání jiných znaků než čínských znaků. Důvodem je první testování logistické regrese, které kromě numerativu 元 přiřadilo opačnou (lexikální) kategorii i znaku 年 (nián, numerativ pro roky, „rok“), v korpusu opatřeným značkou numerativu „q“. Když jsme tedy vyhledávání kolokátů rozšířili o možnost shody s čísly s desetinou čárkou a čísly kombinovanými se znaky, počty kolokátů výrazně vzrostly jak pro znak 元, tak i pro znak 年. Další úpravy vyhledávání už přesnější výsledek nepřinesly. Domníváme se tak, že by numerativ 元 mohl být ve své odchylce od hodnot naměřených autory jako jednotka ojedinělý.

2. Data

2.1 Korpus

Lancasterský korpus čínštiny (McEnery, T., & Xiao, R., 2003) představuje vyvážený korpus textů, který byl sestaven z textů psaných standardní čínštinou, publikovaných v ČLR. Co se týká velikosti korpusu, LCMC celkem obsahuje 1. 003. 289 tokenů a 45. 501 typů. Zařazené texty vyšly v období mezi lety 1991–1992, a zahrnují ve vyváženém poměru texty z 15 různých žánrových kategorií, např. tisk, mysteriózní a detektivní literatura, akademické texty, sci-fi, bojová umění, náboženství nebo rukodělné činnosti. Korpus je označen pomocí XML značení, každé slovo a každá věta jsou z obou stran opatřeny značkami, která určují jejich začátek a konec. Například slovo z naší databáze s nejnižší frekvencí v korpusu (frekvence = 23) je 那边 (nàbian, „tam“). Přímo v textovém souboru, ve kterých je korpus k dispozici je zaneseno následovně:

- <w POS="r">那边</w>

Po špičaté závorce, která označuje začátek slova je slovo opatřeno značením slovního druhu, v tomto případě značkou „zájmeno“, poté následuje samotné slovo a pak opět značka oznamující konec slova. Sada značek v LCMC obsahuje na 50 *tagů*, které označují i podkategorie jednotlivých slovních druhů.

2.2 Databáze slov k ohodnocení

Druhým zdrojem dat byla databáze osmi set vybraných jednotek (slov), kterou sestavili Sun & Saavedra. Autoři uvádějí, že při výběru slov sledovali dvě podmínky. První podmínkou bylo, že do databáze vybrali přesně čtyři sta lexikálních a čtyři sta gramatických jednotek. Důvodem tohoto dělení je výhoda, že se poté data nemusí při trénování modelu logistické regrese vyvažovat (uměle vytvořit hodnoty pro domnělé jednotky, které vyváží rozdíl v datech). Autoři článku nejdříve sestavili seznam různých slov, těm poté přiřadili značku (*tag*), která slovu odpovídala v LCMC. Na stránkách lancasterského korpusu čínštiny jsme nedohledali, zda/ podle jaké normy se autoři korpusu řídili při *tagování* jednotlivých slov. Jak bylo zmíněno, sada značek v LCMC umožňuje detailnější dělení příslušnosti do určité kategorie. Autoři, z pochopitelných důvodů, do databáze z velké části vybrali slova, která byly označené „hrubozrnnou“ kategorií slovního druhu (substantivum, adjektivum, numerativ, předložka apod). Při rozřazování vybírali slova s těmito slovními druhy a řídili se podle následujícího klíče v tabulce č. (5).

Druhou podmínkou, kterou si autoři stanovili, byl minimální počet frekvence slova, určili, že ta by neměla být menší než 50, aby se při výpočtu vyhnuli velmi ojedinělým slovům a také výskytům, které mohly obsahovat chybu. Pokud se podíváme na hodnoty slov vypočítané

podle LCMC, najdeme mnoho slov s frekvencí menší než 50. První slovo, které nabývá hodnoty frekvence 51, se nachází až na místě 124, počítáno od spodu.
Je otázkou, čím je tento rozdíl způsobený. Při porovnání několika málo základních hodnot, které autoři v článku uvádí se zdá, že počty frekvencí ve výpočtech pro BP odpovídají počtům frekvencí, ke kterým dospěli autoři článku. Šance, že ve výsledcích pro BP jsou chybné hodnoty, protože bylo jinak nastavené vyhledávání nebo jiná nesrovnalost je samozřejmě vysoká. Z dalších možností se ještě nabízí, zda v čase mezičase nedošlo k aktualizaci LCMC, vydání jiné verze. Jak uvádí Gries (Gries & Ellis, 2015), proces tokenizace slov je zvláště v čínštině nelehkou záležitostí a stává se, že odlišné softwarové nástroje použité k tokenizaci rozdělují slova odlišným způsobem a mohou tím ovlivňovat i hodnoty jejich frekvence. V posledku v databázi víc slov chybí, například v kapitole o „multifunkčních“ slovech (zde viz 3.4 Multifunkční slova) uvádíme seznam multifunkčních slov, převzatý z článku od autorů, pro mnohá slova uvedená v článku jsme v databázi nenašli ekvivalent, proto je v kapitole neuvádíme.

Lexikální kategorie	**Označení (*tag*)**	**Gramatická kategorie**	**Označení (*tag*)**
Substantivum	n	Numerativ	q
Sloveso	v	Pomocné slovo	u
Adjektivum	a	Předložka	p
Příslovce	d	Spojka	c
Časové slovo	t	Zájmeno	r
Prostorové slovo	s	Modální částice	y
Lokativ	f	Přípona	k
		Předpona	h

Tabulka č. 5 - slovní kategorie a jejich značky dle autorů

2.3 Dělení slovních druhů v čínštině

Předkládáme stručný přehled vývoje slovních druhů jako kategorií, dle Encyklopedie čínského jazyka a jazykovědy (Sybesma et al., 2017).
Až do 19. století nebyl koncept přidělování slov do kategorií podle jejich druhu v čínské lingvistice příliš známý. Toto dělení do tříd podle západního vzoru poprvé v gramatické příručce představil v roce 1898 čínský lingvista 马建忠 Mǎ Jiànzhōng.
V roce 1986 došlo ke shodě na vytvoření jednotného standardu pro výuku gramatiky. 暂拟汉语教学语法系统 Zànnǐ Hànyǔ jiàoxué yǔfǎ xìtǒng „Návrh systému výuky čínské gramatiky“ tedy rozdělil slovní jednotky do jedenácti kategorií, jednalo se o:

- substantiva, numerativy, zájmena, slovesa, adjektiva, číslovky, příslovce, předložky, spojky, částice a citoslovce

V roce 1984 byla při revizi tohoto systému doplněna kategorie zvukomalebných slov, onomatopoií.
Do současnosti se těchto dvanáct kategorií slovních druhů zachovalo, někdy s drobnými úpravami. U některých kategorií se například vyčleňují další podtypy hlavních kategorií. Takovým příkladem detailního dělení může být i Lancasterský korpus čínštiny. V jeho repertoáru značek pro slovní druhy najdeme na 50 různých *tagů*.

Na tomto místě uvádíme přehled, jakým způsobem mohou být slovní druhy v čínštině rozděleny.
Jako referenci odkazujeme dělení podle čínského jazykovědce jménem Zhū Déxi 朱德熙, které uvádí v knize Yǔfǎ Jiǎngyì 语法讲义 („Materiály pro výuku gramatiky“) (Zhu, 1999). Jeho dělení nejdříve zahrnuje dvě kategorie, kategorii shízì 实字 („plná slova“) a xūzì 虛字 („prázdná slova“). Dle jeho názoru, slova, která spadají do kategorie „plných slov“ mohou zastupovat funkci subjektu, objektu nebo predikátu. Těchto funkcí gramatická slova nabývat nemohou. Další charakteristiky dělení podle Zhū Déxiho jsou:

- velká většina „plných slov“ je samostatná (může samostatně tvořit větu), „prázdná slova“ jsou vázaná (nemohou samostatně tvořit větu)
- velká většina „plných slov“ nemá fixní pozici ve struktuře věty, velká většina „prázdných slov“ má fixní pozici ve struktuře věty
- „plná slova“ tvoří otevřenou kategorii, „prázdná slova“ tvoří uzavřenou kategorii

2.2.3 Lexikální kategorie

Kategorii „plných slov“ 实词 dále dělí na tǐcí 体词 (slova substantivní povahy) a wèicí 谓词 (predikativa). Ve výčtu za každou kategorií uvádíme tag, kterým je určena v LCMC. Do skupiny slov substantivní povahy řadí:

- míngcí 名词 substantiva značka N
- chùsuǒcí 处所词 slova místa značka NS
- fāngwèicí 方位词 lokativy značka F
- shíjiāncí 时间词 slova času značka T
- qūbiécí 区别词 atributiva značka B
- shùcí 数词 číslovky značka M
- liàngcí 量词 numerativy značka Q
- dàicí 代词 zájmena substantivní povahy značka R

Do skupiny predikativ zařazuje:

- dàicí 代词 zájmena predikativní povahy značka R

- dòngcí 动词 slovesa značka V
- xíngróngcí 形容词 adjektiva značka A

Lancasterský korpus čínštiny obsahuje značky pro všechny zmíněné kategorie. Co se týká kategorie zájmen 代词, nijak nerozlišuje zájmena substantivní povahy a zájmena predikativní povahy, obě kategorie označuje stejnou značkou, a to *tagem* „R“. Slova místa (NS) autoři v tabulce č. (5) nevydělují z kategorie substantiv, v databázi se však taková slova vyskytují (Zhōngguó 中国 „ČLR“, Shànghǎi 上海, „Shanghai“). V tabulce navíc uvádějí kategorii „prostorových slov“, značenou *tagem* „S“. V databázi osmi set jednotek k analýze se objevují dvě jednotky označené tímto *tagem,* jedná se o xīnli 心里 („v mysli, na srdci“) a yīqǐ 一起 („na stejném místě, společně“).

Hlavním rozdílem je však v tomto dělení zařazení skupiny numerativů a zájmen. Tyto kategorie, které Zhū Déxi uvádí jako příslušící do třídy „plných slov“, tedy lexikální kategorie (řečeno termínem autorů studie), uvádí autoři článku jako skupiny náležící do kategorie gramatické.

2.2.4 Gramatická kategorie

Funkční kategorii „prázdných slov“ 虚词 pak dále dělí do těchto pěti skupin:

- fùcí 副词 příslovce značka D
- jiècí 介词 předložky značka P
- liáncí 连词 spojky značka C
- zhùcí 助词 pomocná slova značka U
- yǔqìcí 语气词 modální částice značka Y

Pod kategorií funkčních uvádí Zhū Déxi ještě:

- nǐshēngcí 拟声词 onomatopeia značka O
- gǎntàncí 感叹词 citoslovce značka E

Kromě onomatopeí a citoslovcí autoři v tabulce č. (5) uvádějí všechny výše popsané gramatické kategorie. Navíc ještě přidávají dvě kategorie, které v Zhū Déxiho členění výše nenajdeme, tím jsou kategorie sufixů (značka „K“) a prefixů (značka „H“). Třetí kategorií, která se v korpusu také objevuje, je kategorie „adjektivního příslovce“, označná *tagem* „AD“. Takovým slovem je v korpusu například 一般 (yībān, „stejný“, „běžný“, ale též v yībān lái shuō, „obecně řečeno“).

Hlavním rozdílem je však zařazení kategorie příslovcí. Zde vidíme, že jsou zařazena do kategorie „prázdných slov“, gramatické kategorie. Autoři studie je však řadí do kategorie lexikální.

2.2.5 Diskuze o zařazení jednotek do kategorií

Jakým způsobem autoři určili, zda konkrétní kategorii slov umístí dále do kategorie lexikální nebo gramatické autoři v článku přesně neuvádí. Podle jejich uchopení konceptu gramatikalizace (viz 1.2 Východiska autorů studie) však můžeme usuzovat, že velmi netrvají na přesném a neměnném ukotvení slovních jednotek do těchto dvou kategorií.
Dále v článku explicitně uvádějí, že rozdělení jednotek do dvou kategorií je spíše „praktické rozhodnutí“, které je nutné pro vytvoření seznamu lexikálních a gramatických jednotek a také pro otestování kvality modelu. Co se týká stanovení přesné hodnoty gramatického statusu (0.500), která bude tvořit takto jednoznačnou linii mezi oběma kategoriemi, autoři opět zdůrazňují, že byla stanovena pouze z praktických důvodů.

3. Vlastní výpočet

Opakování výpočtu gramatického statustu podle Linlin Sun a Davida C. Saavedry mělo v realitě strukturu spíše chaotickou, s mnoha slepými uličkami a opravami dat. V této kapitole však budeme sledovat následující strukturu.
Nejdříve představíme pojem „binární logistické regrese“ vůbec, následně přistoupíme k srovnání evaluací modelu autorů a modelu pro bakalářskou práci. V následné analýze výsledků porovnáme, jaké gramatické jednotky model vyhodnotil jako lexikální a naopak. Porovnáme také výsledky pro jednotky, které spadají do více než jedné slovní kategorie, v korpusu je jim přidělen více než jeden *tag*.

K výpočtu základních hodnot jsme převzali script dostupný v doktorské práci Davida C. Saavedry (Correia Saavedra, 2019), který bylo nutné upravit tak, aby bylo možné vyhledávat čínská slova. Jako programovací jazyk byl použit *Python* (Python, b.r.) , explorace a vizualizace dat proběhla za pomoci softwaru *Jupyter notebook* (Kluyver, T. et al., 2016). Pro trénování a evaluaci všech tří modelů (logistické regrese, random forests, PCA) jsme využili knihovnu scikit-learn[2] jazyka *Python*.
Pracovali jsme také s korpusovým nástrojem AntConc (Anthony, L., 2024), který sloužil k rychlé orientaci v korpusu, zkoumání kontextu dané jednotky a také k orientačnímu ověřování naměřených hodnot. Postup práce byl přibližně následující:

Za pomoci převzatého scriptu jsme vypočítali hodnoty frekvence, kolokátů, koligátů a proporční odchylky pro každé jednotlivé slovo z databáze slov, kterou poskytli autoři. Po výpočtu těchto vstupních hodnot, bylo nutné tyto „syrové“ hodnoty převést na sedm proměnných (viz 1.3 Popis základních metrik), které následně sloužily jako nezávislé proměnné (*features*) při predikci modelem binární logistické regrese. Model z poskytnutých hodnot predikoval závislou proměnnou, v tomto případě, zda jednotka náleží do lexikální nebo gramatické kategorie. Po natrénování modelu (v této fázi má model k dispozici spolu s nezávislými proměnnými i závislou proměnnou – příslušnost k lexikálním nebo gramatickým jednotkám), ohodnotí model data v plném rozsahu a bez toho, že by měl k dispozici závislou proměnnou. Hodnocení vypadá tak, že každé jednotce přiřadí, s jakou pravděpodobností spadá do gramatické kategorie, kategorie, kterou kódujeme jako jedna. Výsledkem je pro každou jednotku z databáze skóre v rozsahu 0 až 1. Jednotky, které budou nabývat hodnot blízko nule, by měly spadat do lexikální kategorie, naopak jednotky na opačné škále kontinua, blížící se svým skóre jedné, by měly spadat do gramatické kategorie (viz rozdělení slov do kategorií v tabulce č. (5).

Autoři uvádějí, že jejich primárním cílem není dosáhnout maximální možné přesnosti s jakou natrénovaný model odhadne příslušnost jednotky ke kategorii (i když i přesnost modelu je pro proces důležitá), avšak konkrétní skóre „gramatického statusu“, které modelem ohodnocené jednotky získají.

[2] Pojmem „knihovna“ myslíme sadu nástrojů pro programovací jazyk Python, dostupné z: https://scikit-learn.org/stable/

3.1 Binární logistická regrese

Logistická regrese[3], na rozdíl od regrese lineární, pracuje s kategorickou závislou proměnnou. Výsledek logistické regrese, její závislá proměnná, tedy bude nabývat dvou vzájemně se vylučujících hodnot (lexikální – gramatické, vítězství – prohra, koupí – nekoupí). Nezávislé proměnné mohou nabývat numerických i kategorických hodnot (Correia Saavedra, 2021). Kategorické hodnoty je potřeba převést jak u závislých, tak u nezávislých proměnných na hodnoty binární, v tomto výpočtu jsme pracovali s nulou pro lexikální kategorii a jedničkou pro kategorii gramatickou, stejně jako autoři článku. Zde však výsledkem není „pouze" binární vyhodnocení, do které ze dvou kategorií jednotka pasuje, za pomoci logistické regrese hodnotíme, s jakou pravděpodobností jednotka spadá do gramatické kategorie (této kategorii je přisouzeno číslo 1). Výsledkem takového hodnocení je číslo v rozsahu nula až jedna. Právě toto číslo autoři studie nazývají gramatickým statusem jednotky. U takto ohodnocené jednotky pak můžeme zkoumat, kde se nachází na škále mezi lexikálním a gramatickým podle svého gramatického statusu.

3.2 Porovnání modelů

V článku autorka a autor zdůrazňují, že snažit se dosáhnout co nejvyšší přesnosti modelu není primárním cílem studie. Přesto tyto hodnoty v článku uvádějí, protože je pro proces přisouzení gramatického statusu důležité, aby model fungoval v rámci možností přesně. V tabulce č. (6) níže uvádíme koeficient determinace (R^2) a hodnocení přesnosti modelu (prosté vydělení počtu správných predikcí počtem celkových predikcí).

元	**Nagelkerke R^2**	**Cox & Shnell R^2**	**Přesnost modelu**
Autoři	0.553	0.415	80.3 %
BP	0.5256	0.3942	80 %

Tabulka č. 6 - evaluace modelů

Nyní se zaměříme na koeficienty jednotlivých proměnných. Ve studii autoři při trénování modelu logistické regrese použili „postupnou" (*stepwise*) logistickou regresi. Tento postup může být užitečný, pokud jsou nezávislé proměnné mezi sebou silně korelované, model totiž bude takové prediktory hodnotit jako redundantní. V případě „postupné" logistické regrese jsou nezávislé proměnné po jedné buď přidávány nebo ubírány z celkového počtu. Po každém takovém „kroku" dojde k vyhodnocení, zda konkrétní proměnná model vylepšila. Takto se postupně dojde k výsledným proměnným (Correia Saavedra, 2021), které mají v modelu největší váhu.

[3] https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

V tabulce č. (7) předkládáme porovnání β koeficientů jednotlivých metrik. Pozitivní koeficient naznačuje, že pokud se zvýší hodnota proměnné, ke které koeficient odkazuje, zvýší se i pravděpodobnost, že bude jednotka zařazena do gramatické kategorie. V tabulce č. (7) jsou z naší strany uvedeny koeficienty, když model predikoval výsledky na základě všech sedmi metrik. Autoři koeficienty proměnných, které byly v rámci „postupné“ logistické regrese odstraněny, neuvádí.
Z tabulky č. (7) jsou pozorovatelné hodnoty β koeficientů vynechaných metrik, tedy diverzita kolokátů 1 vlevo (0.284), diverzita koligátů 1 vpravo (0.082) a proporční odchylka (- 0.0103).

	Frequency	**Collocate_Div_1L**	**Collocate_Div_1R**	**Collocate_Div_4**
Autoři	2.171	---	0.309	1.943
BP	1.73681611	0.284671	0.4863128	1.49063922
	MaxColliPer_1L	**MaxColliPer_1R**	**DevProp**	
Autoři	0.997	---	---	
BP	0.91775345	0.08289136	- 0.01037233	

Tabulka č. 7 - koeficienty proměnných

Zde (tabulka č. 8) uvádíme, jakých β koeficientů proměnné dosahují, pokud jsou vynechány výše zmíněné metriky, které mají malý vliv na predikci závislé proměnné. V koeficientech naměřených autory a koeficientech vypočtených pro tuto bakalářskou práci jsou rozdíly, není jednoznačné, čím jsou přesně dány. Jednou z možností je odchylka v naměřených základních hodnotách (1.5 Porovnání základních hodnot), dalším důvodem by mohlo být, že autoři k výpočtu používají jiný software (IBM SPSS Statistics for Windows), případně možné rozdíly ve verzích LCMC korpusu (též 1.5).

	Frequency	**Collocate_Div_1R**	**Collocate_Div_4**	**MaxColliPer_1L**
Autoři	2.171	0.309	1.943	0.997
BP	2.6363499	0.32079518	2.00602422	0.9907774

Tabulka č. 8 - koeficienty proměnných po redukci

3.3 Analýza výsledků

V tabulce č. 9 níže uvádíme srovnání výpočtů autorů s výpočty provedenými pro tuto bakalářskou práci. Do tabulky jsme zařadili pouze třicet pět jednotek s nejvyšším skóre gramatického statusu, a stejný počet jednotek z opačné škály, tedy s nízkým skóre gramatického statusu. Jednotlivé sety výsledků si neodpovídají přesně, většinou se však pohybují ve vzdálenosti několika pozic od sebe. Autoři jich v článku sice uvádějí sedmdesát z obou konců kontinua, pro představu by však mohl tento výčet stačit. V tomto výčtu (tabulka č. 9) výsledky obsahují u obou výpočtů pouze jedno „chybné“ zařazení, je jím již zmíněný znak 是 (shì, sloveso identifikace, „být“), který je v tabulce označen hvězdičkou (více viz 3.2.2).

Tabulka č. 9 - analýza výsledků logistické regrese

35 jednotek s nejvyšším skóre					35 jednotek s nejnižším skóre			
Autoři		BP			Autoři		BP	
Jednotka	Skóre	Jednotka	Skóre		Jednotka	Skóre	Jednotka	Skóre
('u', '的')	1.000	('U', '的')	1		('n', '球')	0.044	('N', '政治')	0,061
('u', '了')	1.000	('U', '了')	0,999		('n', '情况')	0.043	('VN', '管理')	0,060
('q', '团')	0.999	('Q', '团')	0,999		('n', '历史')	0.042	('N', '价值')	0,058
('q', '斤')	0.999	('Q', '斤')	0,998		('n', '国家')	0.042	('N', '问题')	0,055
('q', '艘')	0.999	('R', '之类')	0,998		('n', '牌')	0.042	('N', '部门')	0,053
('r', '此事')	0.998	('Q', '艘')	0,998		('n', '工程')	0.040	('A', '基本')	0,053
('v', '是')*	0.998	('R', '此事')	0,998		('n', '部门')	0.038	('VN', '贸易')	0,051
('q', '对')	0.998	('K', '家')	0,996		('n', '基础')	0.037	('VN', '教育')	0,051
('k', '家')	0.998	('V', '是')*	0,995		('vn', '贸易')	0.036	('N', '人民')	0,050
('c', '也')	0.997	('C', '也')	0,995		('n', '问题')	0.036	('N', '宗教')	0,048
('y', '罢')	0.997	('C', '以免')	0,995		('v', '想')	0.034	('N', '工程')	0,047
('r', '之类')	0.996	('C', '就')	0,994		('n', '政治')	0.034	('N', '新闻')	0,045
('c', '就')	0.996	('P', '每当')	0,994		('n', '目标')	0.033	('N', '目标')	0,043
('c', '到')	0.996	('Q', '丝')	0,994		('n', '事')	0.031	('N', '环境')	0,043
('q', '丝')	0.996	('Y', '罢')	0,994		('n', '计划')	0.030	('VN', '工作')	0,041
('p', '顺着')	0.996	('Q', '对')	0,993		('n', '新闻')	0.028	('V', '发展')	0,039
('y', '极了')	0.996	('Y', '极了')	0,993		('n', '环境')	0.027	('V', '想')	0,039
('c', '以')	0.996	('R', '这项')	0,993		('vn', '工作')	0.027	('VN', '改革')	0,039
('q', '出')	0.995	('P', '顺着')	0,993		('v', '发展')	0.027	('N', '计划')	0,038
('q', '所')	0.995	('C', '到')	0,993		('n', '企业')	0.024	('A', '重要')	0,037
('c', '以免')	0.995	('R', '那边')	0,992		('n', '语言')	0.024	('N', '意义')	0,037
('q', '面')	0.995	('Q', '出')	0,992		('vn', '改革')	0.023	('N', '事')	0,036
('p', '每当')	0.994	('Q', '面')	0,991		('a', '基本')	0.023	('V', '知道')	0,035
('q', '枚')	0.993	('C', '与其')	0,990		('n', '人民')	0.023	('VN', '发展')	0,031
('q', '处')	0.993	('Q', '处')	0,990		('n', '标准')	0.022	('N', '社会主义')	0,030
('r', '那边')	0.993	('C', '以')	0,990		('n', '意义')	0.021	('N', '民族')	0,028
('r', '之')	0.992	('K', '界')	0,989		('vn', '发展')	0.020	('N', '社会')	0,026
('q', '眼')	0.992	('Q', '枚')	0,988		('n', '民族')	0.020	('N', '语言')	0,026
('p', '在')	0.992	('R', '此人')	0,988		('n', '系统')	0.018	('N', '企业')	0,025
('q', '架')	0.992	('RG', '何')	0,987		('n', '社会')	0.015	('N', '牌')	0,021
('k', '界')	0.991	('P', '对着')	0,987		('v', '知道')	0.015	('N', '球')	0,020
('q', '页')	0.991	('C', '不然')	0,987		('n', '技术')	0.012	('N', '标准')	0,019
('k', '制')	0.991	('R', '上述')	0,987		('a', '重要')	0.012	('N', '技术')	0,018
('q', '节')	0.991	('H', '超')	0,986		('n', '经济')	0.008	('N', '系统')	0,018
('r', '这项')	0.990	('U', '似的')	0,986		('n', '社会主义')	0.006	('N', '经济')	0,013

3.3.1 Gramatické jednotky v lexikální kategorii

V prvních sto padesáti slovech s nejnižším hodnocením najdeme dvanáct slov, které by měly spadat do gramatické kategorie, byly však ohodnoceny nízkým skóre. Je velmi nápadné, že téměř všechny jsou označeny *tagem* pro zájmeno. Zbylé dvě jednotky jsou předložky. Čtyři zájmena klasifikoval chybně i model autorů článku, nachází se mezi prvními sedmdesáti jednotkami s nejnižším skóre uvedenými ve studii. Jedná se o tyto zájmena (zájmena shodná s autory článku zvýrazněna):

- 你 (nǐ, „ty“), **我们** (wǒmen, „my“), 她 (tā, „ona“), 它 (tā, „to, ono“), 自 (zì, „sám“, v korpusu označeno jako zájmenný morfém, značka „rg“)
- 谁 (shěi, „kdo“), **什么** (shénme, „co“), 怎么 (zěnme, „jak“)
- 那 (nà, „tamto“), **这样** (zhèyàng, „takto, takovým způsobem“)

A tyto předložky:

- 跟 (gēn, „s“)
- 给 (gěi, značí dativ, jiāo gěi tā yī fēng xìn, „předat mu dopis“)

Výše jsme uvedli, že „tradičně“ (viz 2.2.3 Lexikální kategorie) se zájmena řadí do kategorie „plných slov“, tedy do kategorie lexikální. Jak už jsme zmiňovali, autoři k řazení slovních druhů do jedné z velkých binárních kategorií přistupovali spíše prakticky. Avšak v článku poskytují vodítka, proč jsou zájmena řazena právě takto. K rozhodnutí zařadit zájmena do funkční kategorie autoři uvádějí hlavně dva důvody. První z nich je důvod, že zájmena tvoří uzavřenou slovní třídu a jejich význam je spíše schematický (Sun & Saavedra, 2020, str. 329). Podobně o povaze zájmen (konkrétně osobních zájmen) uvažují i Heine & Song (Heine & Song, 2011), kteří také k tématu dodávají, že „gramatický status osobních zájmen je předmětem kontroverzních diskuzí“. Jejich přiřazení do funkční, gramatické kategorie podkládají těmito tezemi:

- mají schematický význam, poměrně jednoduše konceptualizovaný (osobní deixe a číslo)
- mají o mnoho omezenější možnosti např. co do přibírání částic nebo afixů
- jsou obecně kratší než substantiva a slovesa, (eroze)

Již při popisu základních hodnot jsme mohli pozorovat, že velmi frekventovaná zájmena mají zároveň nízkou diverzitu kolokátů (viz 1.4 Analýza základních hodnot). Jak už jsme uváděli, např. u zájmena 他 (tā, „on“) po nahlédnutí přímo do korpusu vidíme, že jako kolokáty se s tímto zájmenem nejčastěji pojí další zájmena a také frekventovaná slovesa (např. 说, shuō, „mluvit, říkat“ nebo 知道, zhīdao, „vědět“).

3.3.2 Lexikální jednotky v gramatické kategorii

V tabulce č. 9 vidíme chybně zařazené sloveso identifikace 是, které bylo zařazeno do gramatické kategorie v případě výpočtu autorů i výpočtu pro tuto práci.
Autoři v článku primárně přisuzují zařazení slovesa identifikace 是 jeho vysoké frekvenci (11.427 výskytů). V korpusu je hned po částici 的 druhým nejfrekventovanějším slovem, a jak můžeme vyčíst z tabulky koeficientů (3.2 Porovnání modelů), vyšší frekvence má na zařazení jednotky do gramatické kategorie silný vliv. Dalšími důvody, které autoři uvádějí, je jeho užívání ve funkci spony, dále také jeho role ve zdůrazňovací konstrukci. Tyto jeho možná užití tak zřejmě přispěli k jeho zařazení na tento konec škály.

Pokud nepočítáme sloveso identifikace, tak první slovo z třinácti chybně zařazených slov do gramatické kategorie má pořadí 173., *tag* „příslovečné adjektivum, „ad“, a jedná se o jednotku 一般 (yībān). Další výčet je takový:

- 完（wán, „spotřebovat, skončit“), 出来 (chūlái, „vyjít, objevit se“), 下来 (xiàlái, „sestoupit“), 达 (dá, „dosáhnout, rovnat se“), V 即 (jí, „znamenat“, „rovnat se něčemu“)
- 即 (jí, „v současnosti“), 仍 (réng, „stále“), 甚至 (shènzhì, „dokonce“)
- 间 (jiān, „mezi“), 里 (lǐ, „v, uvnitř“)
- 的 (de, atributivní slovo)
- 一起 (yīqǐ, „spolu, na jednom místě“)

Jak vidíme, tyto jednotky netvoří stejně homogenní skupinu, jako gramatické jednotky zařazené na lexikální konec škály, přesto se je pokusíme rozklíčovat. První skupinu tvoří jednotky označené *tagem* pro sloveso, výlučnost těchto jednotek (konkrétně 下来, 出来, 完) byla viditelná už při rozboru základních hodnot, nebudeme se jim zde již věnovat (viz popis a analýza proměnných). Při prozkoumání kolokátů slovesa 达 (dá, „dosáhnout, rovnat se“) v korpusu můžeme vidět, že se velice často pojí s číslovkou, to může být důvodem jeho nízké diverzity koligátů, tím pádem zařazení do lexikální kategorie. U slovesa 即 (jí, „znamenat, rovnat se něčemu“) je možná důvodem jeho zařazení též nízká diverzita koligátů, v pohledu do korpusu na místě jeho levého koligátu často najdeme substantivum.

3.3.3 Hodnoty na pomezí

Po vyhodnocení okrajů lexikálně – gramatické škály, můžeme odkrýt, jaké lexikální jednotky stojí na pomezí těchto dvou kategorií, tedy řečeno slovy čísel logistické regrese. Již jsme zde víckrát opakovali, že toto náhlé rozdělení kategorií hodnotou 0,500 je dost hrubozrnné, a vlastně to tak potvrzují i výsledky, protože v samotné středu se objevují jednotky rozdílných slovních druhů i lexikálně-gramatických kategorií. Přesto se domníváme, že jednotky vyhodnocené jako na půl cesty mezi lexikálním a gramatickým můžou být zajímavé. Těchto šest jednotek se nachází kolem přelomu mezi lexikálním a gramatickým:

Gramatická „polovina“:

- 顿 (dùn, numerativ pro porce“) značka Q, numerativ
- 的话 (dehuà, pomocné slovo vyjadřující podmínku) značka U, pomocné slovo
- 非 (fēi, prefix s významem záporu) značka H, prefix

Lexikální „polovina“:

- 别人 (biérén, „ostatní lidé) značka R, zájmeno
- 通过 (tōngguò, „ skrze“) předložka značka P, předložka
- 少 (shǎo, „málo“) značka A, adjektivum

Z tabulky č. (9) můžeme vyčíst, že na spodku výsledků modelu binární logistické regrese se nachází až na dvě adjektiva (jīběn 基本 „základní“) a (zhòngyào 重要 „důležitý“), pouze substantiva a slovesa. Na opačném konci zase převládají numerativy (14 výskytů z 35 uvedených), dále sem byly zařazeny pomocná slova, spojky, modální části a také sufixy. Vidíme, že ve střední části seznamu jednotek se míchají různorodé kategorie slov, nejen ve smyslu lexikálního – gramatického zařazení, ale i ve smyslu mnohosti kategorií slovních druhů.

3.4 Multifunkční slova

Autoři zdůrazňují, že pokud má jednotka více než jednu značku slovního druhu, pro každý slovní druh je skóre vypočítáno individuálně. Do studie zařazují tyto „multifunkční“ jednotky, na kterých můžeme vidět, jak lexikální jednotka nabyla jeden nebo více gramatických významů a zároveň si zachovala svůj původní lexikální význam.
Autoři navrhují, že by model logistické regrese mohl přispět k řešení otázky, jestli a do jaké míry se tato čínských „multifunkčních“ slova od sebe liší, co do stupně gramatického statusu.

Následuje tabulka č. (10), ve které jsou taková „multifunkční slova“ seřazená podle hodnoty gramatického statusu.

Jednotka	Pinyin	Frequency	Colloc_4	Collig_1L	Status	Skóre
('Y', '了')	le	2294	0,30231	0,363	GRAM	0,154398
('U', '了')	le	10379	0,19366	0,863	GRAM	0,999471
('C', '就')	jiu	40	0,809375	0,538	GRAM	0,994019
('P', '就')	jiu	197	0,585025	0,392	GRAM	0,701956
('D', '就')	jiu	3236	0,287044	0,388	LEX	0,293901
('C', '到')	dao	45	0,816667	0,488	GRAM	0,992688
('P', '到')	dao	153	0,651144	0,241	GRAM	0,705093
('V', '到')	dao	2606	0,325691	0,431	LEX	0,353129
('P', '在')	zai	9899	0,208064	0,359	GRAM	0,983432
('U', '在')	zai	474	0,482859	0,504	GRAM	0,563842
('Q', '下')	xia	36	0,625	0,786	GRAM	0,979014
('V', '下')	xia	273	0,536172	0,292	LEX	0,385379
('F', '下')	xia	1065	0,426408	0,358	LEX	0,271629
('Q', '把')	ba	97	0,554124	0,769	GRAM	0,936745
('P', '把')	ba	1989	0,354324	0,362	GRAM*	0,232657
('Q', '头')	tou	29	0,711207	0,421	GRAM	0,944519
('N', '头')	tou	263	0,529943	0,328	LEX	0,415231
('Q', '回')	hui	60	0,525	0,5	GRAM	0,619574
('V', '回')	hui	241	0,525934	0,363	LEX	0,449796
('P', '为')	wei	1702	0,366113	0,461	GRAM*	0,34261
('V', '为')	wei	859	0,486758	0,252	LEX	0,283433
('P', '用')	yong	986	0,433063	0,357	GRAM*	0,275822
('V', '用')	yong	452	0,510232	0,198	LEX	0,214383

Tabulka č. 10 – hodnoty gramatického statusu „multifunkčních“ jednotek

Při pohledu na tabulku č. 10 vidíme vždy dvojici nebo trojici znaků, které se liší v tagu, který je přiřazuje do konkrétní slovní kategorie. Chybně jsou zařazeny tři jednotky, které jsou v tabulce označeny hvězdičkou. Co se však týká jejich „relativního“ zařazení v rámci setu, skóre gramatického statusu odpovídá velmi dobře.
Podívejme se blíže na příklady, kde je v setu numerativ zároveň se slovem/slovy patřící do lexikální kategorie, tabulka č. (11).

	Pinyin	**Slovní druh**	**Frequency**	**Colloc_Div_4**	**MaxColli Per_1L**	**Status**	**Skóre**
下	xià	Numerativ	36	0,625	0,786	GRAM	0,979014
下	xià	Sloveso	273	0,536172	0,292	LEX	0,385379
下	xià	Lokativ	1065	0,426408	0,358	LEX	0,271629
头	tóu	Numerativ	29	0,711207	0,421	GRAM	0,944519
头	tóu	Substantivum	263	0,529943	0,328	LEX	0,415231
回	huí	Numerativ	60	0,525	0,5	GRAM	0,619574
回	huí	Sloves	214	0,525934	0,363	LEX	0,449796

Tabulka č. 11 – porovnání multifunkčních jednotek, II.

Při relativním porovnání jednotek v setu vychází skóre gramatického velmi dobře.
U skupiny 回 je sloveso ohodonceno nízkým skóre, numerativ dosahuje skóre daleko vyššího. To samé můžeme pozorovat i u skupiny 头, o mnoho nižším skóre je označeno substantivum, následuje numerativ s vysokým skóre gramatického statusu.
U skupiny 下 má nejnižší ohodnocení lokativum, následuje sloveso a pak až s vysokým skóre numerativ.

Dle Janet J. Xing (Xing, 2012) zájem jazykovědců o numerativy kvůli jejich jedinečné gramatické roli, trvá již velmi dlouho. Sémantická změna je pro gramatikalizaci klíčová, numerativy při ní přecházejí od specifických, konkrétních významů k abstraktním, obecným funkcím. Například se ukazuje, že numerativy vznikly z běžných konkrétních podstatných jmen používaných při počítání, a postupně se staly abstraktními numerativy v početních konstrukcích.
Po přidělení gramatického skóre dostáváme k dispozici nástroj, který by nám mohl dát nápovědu, jak tyto jednotky dobře rozlišit, a posoudit tak jejich stupeň gramatikalizace.

4. Odlišné modelování

Kromě binární logistické regrese, se dají k vypočítání gramatického statusu použít i jiné způsoby modelování. Zajímalo nás, jestli a jak se výsledky promění, pokud pro výpočet použijeme jiný model.

4.1 Random Forests

Random Forests[4] představuje kvantitativní model, který můžeme použít jak pro klasifikační úlohy (kategorizaci dat), tak pro regresní úlohy. Pro naši úlohu jsem tento model vybrali, protože dokáže stejně jako binární logistická regrese na základě vstupních proměnných predikovat závislou proměnnou, respektive její pravděpodobnost.
Pro představu o tom, jak model rozhoduje si představme, že Random Forests je model složený z mnoha menších modelů (= rozhodovacích stromů), které vzájemně spolupracují. Stromy operují každý na jiné části souboru dat a poté, co všechny stromy provedou své předpovědi, je model Random Forests zkombinuje zase všechny dohromady a vytvoří konečnou predikci.
Model Random Forests vyhodnocoval závislou proměnnou na základě všech základních nezávislých proměnných (u logistické regrese byly ponechány pouze čtyři proměnné).

Zde pro přehled uvádíme *feature importance* „důležitost nezávislých proměnných“, nejedná se o koeficienty, které jsme uváděli u modelu logistické regrese (evaluace modelu), tato „důležitost“ vystihuje, jak moc konkrétní proměnná přispívá k predikci. I u Random Forests jsou však jako vlivné proměnné, stejně jako u logistické regrese vyhodnoceny frekvence, diverzita kolokátů s rozsahem 4 pozice na každé straně a rozmanitost koligátů na pozici 1 vlevo.

RF	**Frequency**	**Collocate_Div_1L**	**Collocate_Div_1R**	**Collocate_Div_4**
	0.2938551	0.11257716	0.14062914	0.14620013
RF	**MaxColliPer_1L**	**MaxColliPer_1R**	**DevProp**	
	0.10446952	0.07067255	0.1315964	

Tabulka č. 12 - feature importance

Model byl oproti modelu logistické regrese velmi přesný, do „správné“ kategorie se trefil v 91 % případů. Mnohým jednotkám (zvláště gramatickým), přisoudil jejich hodnotu se stoprocentní pravděpodobností. Jak už bylo řečeno výše, přesnost modelu není primárním cílem ani autorů, ani této BP. Zde mnoho jednotek ohodnocených stoprocentní pravděpodobností naopak zastírá „pořadí“ jednotek, co do gramatického statusu. Může být však zajímavé podívat se na jednotky, které tento model ohodnotil chybně, a porovnat, zda se stýkají s chybně ohodnocenými jednotkami logistickou regresí.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

První chybně zařazené slovo, pokud postupujeme na škále od gramatických jednotek směrem k lexikálním, je sloveso 达 (dá, „dosáhnout, rovnat se“), které bylo takto zařazeno i modelem logistické regrese, model přisoudil této jednotce skóre 0.95.
Dále se stejně jako u logistické regrese opakují 甚至 (shènzhì, „dokonce“) a 里 (lǐ, „v, uvnitř“). Je zajímavé, že slovesu identifikace 是 (shì) s velmi vysokou frekvencí, model Random Forests přidělil skóre 0.18, a to i přesto, že je frekvence pro model při rozhodování nejvíce důležitá (tabulka č. 12).

Na opačné straně najdeme zájmena 每 (méi, „každý“), 别 (bié, „jiný“), 他们 (tamēn, „oni“) a 那 (nà, „tamto“).
Dále numerativ 句 (jù, numerativ pro věty) a také dvě modální částice 吧 (ba) a 吗 (ma). Kontroverzní pozici zájmen jsme tu již zmiňovali.
Může být zajímavé, že model logistické regrese i model Random Forests opakovaně přiřazují zájmenům nízký gramatický status. Random Forests zařazuje zájmena do lexikální kategorie, přestože se zdá, že se i v případě jednotky s vysokou frekvencí dokáže rozhodnout správně. Nabízí se otázka, zda bychom nemohli tvrdit, že zájmena, přestože je lingvistky a lingvisté mají tendenci řadit minimálně konceptuálně do gramatické kategorie (pokud je takové explicitní dělení třeba, např. pro potřeby tohoto výpočtu), nemohou paralelně vykazovat takové chování, které je specifické spíše pro jednotky z kategorie lexikální.

4.2 PCA

PCA[5] neboli analýza hlavních komponent (*principle component analysis*) je způsob, jak mnoho nezávislých proměnných redukovat na několik komponent tím, že je odstraněna redundance z datasetu. Nejdříve model vytvoří korelační matrici pro všechny dvojice proměnných, aby zjistil, které proměnné jsou silně korelované. Na základě těchto korelací vytvoří nový set proměnných (komponent), kterými původní proměnné nahradí (Correia Saavedra, 2021). První komponenta se snaží obsáhnout maximum variability v datech, druhá komponenta vysvětlí maximum variability, kterou první komponenta nezachytila, třetí komponenta pojme, co nevysvětlily první dvě atd.

Tuto metodu tedy nelze využít přímo k ohodnocení jednotek „gramatickým statusem“, může být však nápomocná při exploraci proměnných. Z grafu č. (2) se zobrazením korelovaných proměnných v kapitole 1.4 je zřejmé, že proměnné mezi sebou příliš nekorelují (kromě *Collocate_Div_4* s *Collocate_Div_*1R a *Collocate_Div_1L*), není v nich tedy velmi mnoho dat, které by PCA mohla redukovat. Představujeme ji hlavně z toho důvodu, že k proměnným staví jiným způsobem, než binární logistická regrese a Random Forests, tím pádem by tato metoda mohla přinést nové pohledy.

Spíše pro dodržení struktury přikládáme tabulku č. (13), ve které je uvedené, v jakém poměru vysvětluje první nebo druhá komponenta variabilitu u jednotlivých proměnných, takzvané *eigenvalues*.

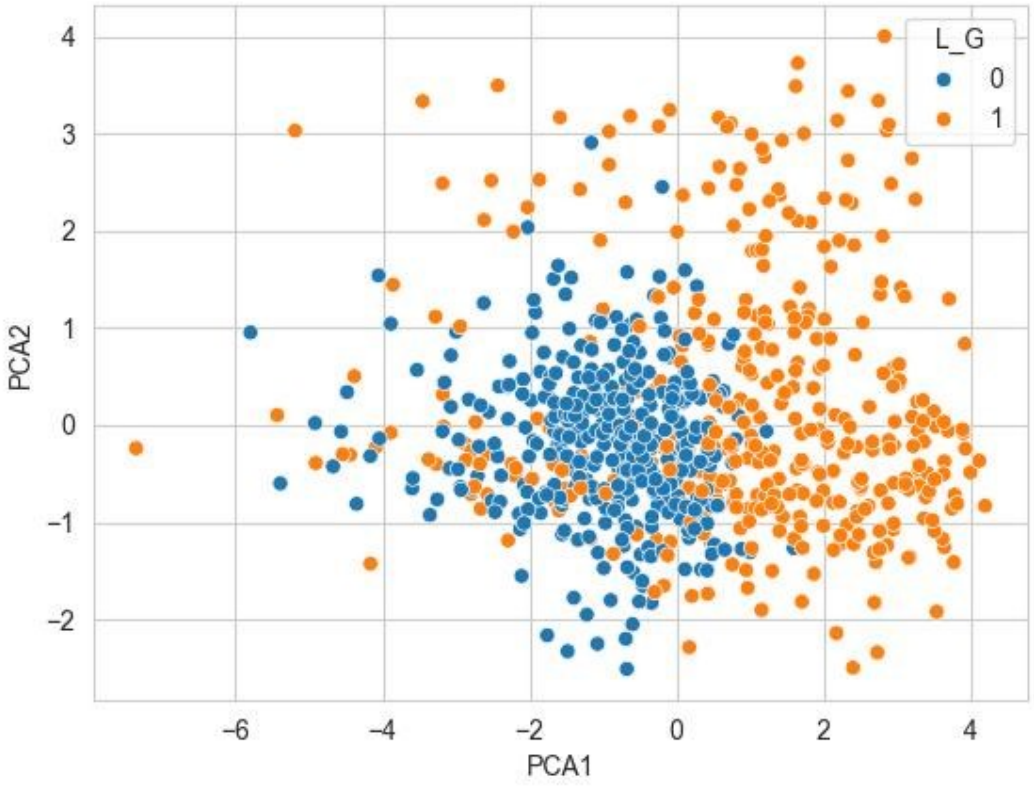
	Frequency_log2	**Collocate_Div_1L**	**Collocate_Div_1R**	**Collocate_Div_4**
PCA1	-0.50716222	0.305577	0.41699144	0.51609831
PCA2	-0.0540994	-0.56172563	0.20123355	-0.09261626
	MaxColliPer_1L	**MaxColliPer_1R**	**DevProp**	**Součet**
PCA1	0.10452693	-0.03813533	0.44360982	1,59113
PCA2	-0.38470456	0.04697859	-0.0540994	0, 12632

Tabulka č. 13 - PCA poměr vysvětlené variability

V řádku PCA1-2 je tedy zanesen poměr vysvětlené variability pro každou proměnnou.

V následujícím bodovém grafu č. 4, který je barevně rozdělený na jednotky zařazené do lexikální (modrá barva) a gramatické (oranžová) kategorie můžeme vidět, jakým způsobem PCA data transformovala. Pomocí takto obarveného grafu můžeme popsat odlehlé hodnoty a také pozorovat, jakým způsobem se data shlukují.

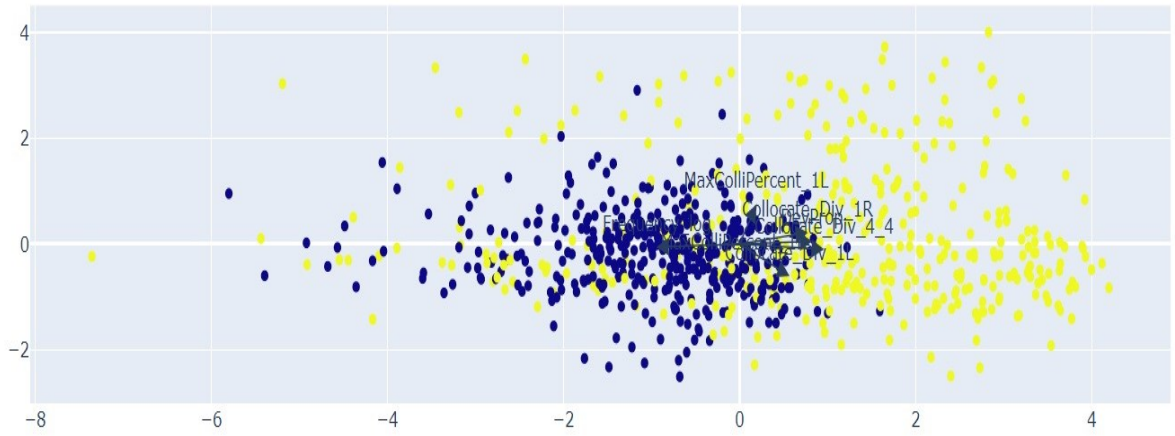
[5] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html



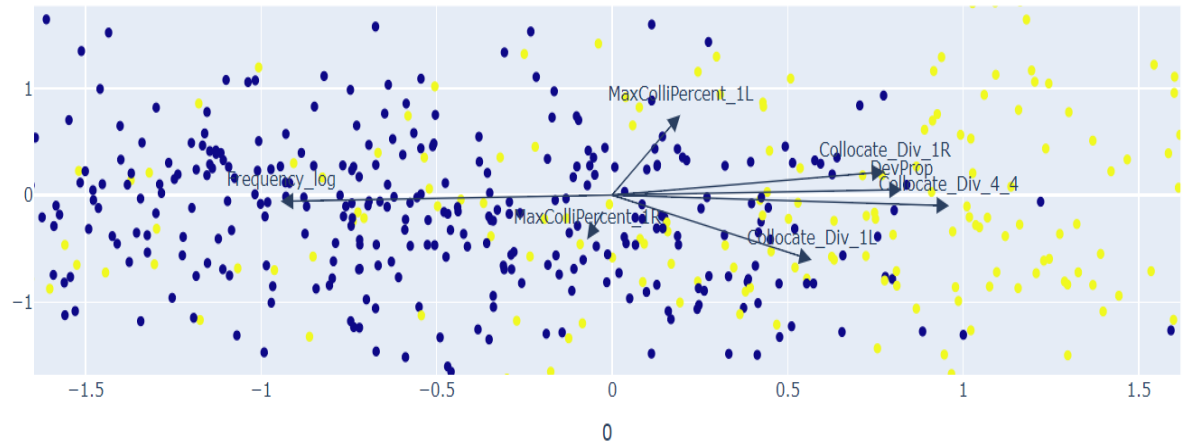
Graf č. 4 – PCA, ze sedmi proměnných

Na grafu je vyneseno všech sedm proměnných, kteří autoři na začátku studie navrhovali zapojit k hodnocení jednotek. První komponenta (PCA1) je vykreslená na ose x, druhá komponenta na ose y. U první komponenty vidíme data shluknutá do dvou skupin, tyto dvě skupiny, lexikální a gramatická jsou přibližně rozděleny hodnotou nula.

Pro podrobnější přehled poslouží graf č. (5) a graf č. (6), na kterých jsou zanesené hodnoty původních proměnných, tedy jejich *eigenvektory*. Graf č. (6) je přiblížená varianta grafu č. (5), aby byly eigenvektory dobře čitelné.
Jejich délka se odvozuje právě podle hodnot eigenvalues, respektive jejich součtu. Na první pohled z toho tedy můžeme odvodit, poměry vysvětlené variability u jednotlivých komponent. Hodnoty frekvence byly pro tento graf logaritmizovány.



Graf č. 5 - eigenvektory sedmi proměnných



Graf č. 6 – přiblížený graf s eigenvektory

Na grafu č. (6) vidíme eigenvektory „Frequency_log“ (vlevo od centra) a „Collocate_Div_4_4“ (vpravo od centra). Jsou to proměnné, které měly největší podíl na vysvětlení variability první proměnné (odpovídá, pokud porovnáme s výsledky v tabulce č. 13). Nahoře od centra vidíme výrazný eigenvektor „MaxColliPercent_1L“, který se výrazněji podílí na vysvětlení variace druhé komponenty (též viz tabulka č. 13).

Nejvýraznější hodnoty eigenvektorů (Frequency_log, Collocate_Div_4_4, MaxColliPercent_1) jsou ve shodě metrikami, které na základě koeficientů vybrali autoři jako určující při predikci gramatického statusu modelem binární logistické regrese. Při modelování výsledků pro bakalářskou práci i náš model binární logistické regrese vykazoval

podobné koeficienty. Zobrazení dat pomocí PCA tedy může být dalším způsobem, jak určit, které metriky budou pro predikci závislé proměnné určující.
Domníváme se, že metoda PCA může být vhodnou metodou k prozkoumání dat, můžeme také s takto transformovanými daty na dvě základní komponenty dále pracovat.

5. Závěr

Výsledky ve smyslu konkrétního skóre gramatického statusu, které v bakalářské práci uvádíme, jsou srovnatelné s výsledky, které ve své studii uvádí Linlin Sun a David C. Saavedra. Toto vyvozujeme na základě skóre gramatického statusu pro prvních sedmdesát jednotek s nejvyšším gramatickým statusem a prvních sedmdesát jednotek s nejnižším gramatickým statusem. Skóre jednotek na pomezí autoři neuvádí.

V práci jsme se potýkali s několika faktory, které mohly při opakování výpočtu gramatického statusu modelem binární logistické regrese mít vliv na finální výsledky. Za prvé se jedná o naměřené základní hodnoty (frekvence, kolokáty, koligáty, proporční odchylka). Při porovnání s několika hodnotami, které autoři ve studii uvádí jsme zjistili, že se naměřené hodnoty u některých jednotek liší. Vzhledem k tomu, že je však skóre pro jednotky, které se nacházejí na opačných koncích lexikálního a gramatického kontinua srovnatelné s tím, které uvádí autoři studie, domníváme se, že by rozdíl v naměřených hodnotách nemusel být tak výrazný, případně se vyskytl pouze u několika jednotek, tím pádem výsledky výrazně neovlivnil.

Odlišný způsob modelování (Random Forests), který jsme v práci použili jako alternativu k modelu binární logistické regrese, kteroý používali autoři studie, přidělil jednotkám srovnatelné skóre. Model Random Forests byl v predikci gramatického statusu jednotek velmi přesný, mnoha jednotkám, které se řadí na „gramatický konec“ škály přisoudil hodnotu pravděpodobnosti, že se jednotka řadí do „gramatické kategorie, celých 1.0. Ve výsledku usuzujeme, že takové hodnocení není pro posuzování gramatického statusu naprosto vhodné, protože tato přesná predikce smazává vítanou „škálu“, a rozřazení jednotek se tak vrací k binárním kategoriím.

Ve shodě s autory se domníváme, že zajímavým ukazatelem byly sety „multifunkčních jednotek“, jejich hodnoty nastínily postupný vývoj každé jednotky v setu na škále od lexikálnímu ke gramatickému.

Co se týká modelování pomocí PCA, domníváme se, že může být vhodnou metodou k základní exploraci dat, může také přinést zajímavé informace o proměnných v datech. Metoda PCA nám potvrdila, které proměnné (potažmo metriky) obsahují nejvíce informace, a jsou tak určující pro predikci závislé proměnné.

K vytyčeným cílům této práce, konkrétně určit pomocí metrik jako nezávislých proměnných a modelu binární logistické, jaké jednotky jsou v čínštině považovány za gramatické a jaké nikoli a následně se zaměřit na uspořádání jednotek na základě jejich gramatického statusu. V práci jsme představili výsledky, které jsou celkově srovnatelné s výsledky autorů práce, pomocí modelování jsme určili, jaké jednotky se řadí do lexikální a jaké do gramatické kategorie a následně jsme porovnávali uspořádání jednotek na základě jejich gramatického statusu. Následně jsme testovali vhodnost vybraných metrik, a zjišťovali (např. pomocí PCA), které metriky mají určující vliv na závislou proměnnou.

Ve shodě s autory se domníváme, že skóre gramatického statusu přiřazené modelem binární logistické regrese může být vhodným způsobem, jak určit na jaké části škály mezi lexikálním a gramatickým se slova pohybují.

Seznam literatury

Anthony, L. (2024). *AntConc[Computer Software]* (Version 4.3.0) [Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Correia Saavedra, D. (2019). *Measurements of grammaticalization: Developing a quantitative index for the study of grammatical change*. Neuchâtel : Université de Neuchâtel, Faculté des lettres et sciences humaines, Institut de langue et littérature anglaises , 2019.

Correia Saavedra, D. (2021). *Measurements of grammaticalization: Developing a quantitative index for the study of grammatical change*. De Gruyter Mouton.

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, S. Th., & Ellis, N. C. (2015). Statistical Measures for Usage-Based Linguistics. *Language Learning*, *65*(S1), 228–255. https://doi.org/10.1111/lang.12119

Heine, B., & Song, K.-A. (2011). On the grammaticalization of personal pronouns. *Journal of Linguistics*, *47*(3), 587–630. https://doi.org/10.1017/S0022226711000016

Hopper, P. J., & Traugott, E. C. (2002). *Grammaticalization* (Repr). Cambridge Univ. Press.

Kluyver, T. et al. (2016). *Jupyter Notebooks* [Software]. https://jupyter.org

McEnery, T., & Xiao, R. (2003). *The Lancaster Corpus of Mandarin Chinese*. European Language Resources Association. http://www.lancs.ac.uk/fass/projects/corpus/LCMC/

Python (version 3.6.15). (b.r.). [Software]. Software Foundationversion 2.7. Available at http://www.python.org

Sun, L., & Saavedra, D. C. (2020). Measuring grammatical status in Chinese through quantitative corpus analysis. *Corpora*, *15*(3), 317–342. https://doi.org/10.3366/cor.2020.0202

Sybesma, R. P. E., Behr, W., Gu, Y., Handel, Z. J., Huang, C.-T. J., & Myers, J. (Ed.). (2017). *Encyclopedia of Chinese language and linguistics*. Brill.

Xing, J. Z. (2012). *Newest trends in the study of grammaticalization and lexicalization in Chinese*. De Gruyter Mouton.

Zhu, D. (1999). *Zhu De xi wen ji*. Shang wu yin shu guan.

Zipf, G.K. (1932). *Selected studies of the principle of relative frequency in language*.