Filozofická fakulta Univerzity Palackého v Olomouci

Katedra obecné lingvistiky

Jazyková arbitrárnost pohledem kvantitativních metod

magisterská diplomová práce

Autor: Bc. Klára Hájková

Vedoucí práce: Mgr. Vladimír Matlach, Ph.D.

Olomouc

2021

Prohlášení

Prohlašuji, že jsem magisterskou diplomovou práci „Jazyková arbitrárnost pohledem kvantitativních metod“ vypracovala samostatně pod odborným vedením Mgr. Vladimíra Matlacha, Ph.D. a uvedla jsem veškerou použitou literaturu a veškeré použité zdroje.

V dne Podpis

Poděkování

Ráda bych zde poděkovala Mgr. Vladimíru Matlachovi, Ph.D. za čas a energii věnované odbornému vedení této diplomové práce. Poděkování patří také doc. Mgr. Danu Faltýnkovi, Ph.D. za zkonzultování problematiky.

Obsah

Úvod

Ženevský lingvista Ferdinand de Saussure (1857–1913) ve třech cyklech přednášek na Ženevské univerzitě představil nové pojetí obecné lingvistiky. Z jeho myšlenek na základě poznámek sestavili jeho studenti Charles Bally a Albert Séchehaye dílo, které v roce 1916 vyšlo pod názvem *Cours de linguistique générale (Kurs obecné lingvistiky)*.

De Saussure je považován za zakladatele strukturalismu. Definuje jazyk jako systém, ve kterém každý jazykový jev zastává určitou funkci. Kromě řady dichotomií (např. *synchronie* a *diachronie*, *langue* a *parole*, vztahů *paradigmatických* a *syntagmatických*) představuje také teorii jazykového znaku, který se skládá ze dvou složek – *signifiant* a *signifié* (*označující* a *označované*). Základními vlastnostmi jazykového znaku jsou jeho *arbitrárnost*, *lineárnost* a *diskontinuita*.[1] Všechny tyto vlastnosti byly v minulosti a jsou až dodnes předmětem mnoha diskuzí i kritik.

V této práci se budeme zabývat právě arbitrárností jazykového znaku, která znamená, že neexistuje žádný vnitřní vztah mezi *signifié* a sledem hlásek, který je jeho *signifiant*. Tato vlastnost je dnes obecně přijímána, přestože se v průběhu času proti arbitrárnosti znaku objevovala celá řada námitek. Kromě těch, kteří de Saussura kritizovali proto, že jeho koncept špatně pochopili nebo zaměňovali jazykovou nearbitrárnost s motivací (vnitřní vztahy v rámci jednoho jazyka, např. *voda* + *vést* = *vodovod*),[2] byly námitky nejčastěji spojeny s onomatopoickými slovy – např. *haf* imituje zvuk, který vydává pes, ale také slova jako *šeptat, šišlat, chraptět* mají základ v napodobování zvuků, které jsou pro tyto jevy typické.

Vznikl dokonce celý proud fonosymbolismu, který se zabývá vztahy mezi zvuky a významem, které konkrétní zvuky mají.[3] Již v roce 1955 provedli Brown, Black a Horowitz experiment, ve kterém byla posuzována synonyma a antonyma pouze na základě jejich zvukové podoby. Význam slov byl na základě jejich zvukové stránky správně určen s přesností vyšší než 50 %, načež autoři jako jedno z možných vysvětlení nabízejí předpoklad existence univerzálního

[1] Srov. Saussure, F., Bally, C., Sechehaye, A. (1996). *Kurs obecné lingvistiky*. Přeložil Čermák, F. Praha: Academia.

[2] Příklad viz Černý, J. (1996). *Dějiny lingvistiky*. Votobeia, s. 142.

[3] Rozycki, W. (1997) Phonosymbolism and the Verb. *Journal of English Linguistics*, 25/3, s. 202–206, https://doi.org/10.1177/007542429702500303. Dostupné z: https://journals.sagepub.com/doi/abs/10.1177/00754242970250030?journalCode=enga [7. 7. 2021].

fonetického symbolismu.[4] Podobné pokusy byly mnohokrát opakovány a rozšiřovány (např. Maltzman, Morrisett a Brooks, 1956), arbitrárnost znaku však zůstává stále přijímaným faktem.

V této práci bychom se chtěli na problematiku arbitrárnosti podívat z pohledu metod kvantitativní lingvistiky. Na příkladech vybraných toponym a hydronym, konkrétně především na jménech měst a vodních toků, budeme kvantifikovat různé vlastnosti jmen i objektů samotných a hledat mezi nimi souvislosti. Jsou-li jména arbitrární (tedy pokud neexistuje žádný vnitřní vztah mezi objektem a sledem hlásek, který ho označuje), očekáváme, že žádné statisticky prokazatelné vztahy neobjevíme, popřípadě že objevíme vztahy, které jsou vysvětlitelné dalšími empirickými lingvistickými zákony (např. principem jazykové ekonomie).

V první části se budeme zabývat jmény měst v ČR a tím, zda existuje nějaká souvislost mezi velikostí města a délkou jeho jména, a budeme zjišťovat, zda lze natrénovat model strojového učení tak, aby dokázal rozlišovat mezi názvy hlavních a periferních měst (jinými slovy, ptáme se, zda existují nějaké charakteristické rysy, které určují „hlavnost“ města). Ve druhé části se budeme zabývat hydronymy a budeme, obdobnými metodami jako v případě toponym, hledat korelace mezi jejich jmény a vlastnostmi. Vše bude podrobeno diskuzi o zjištěných jevech a o tom, zda tyto jevy opravdu souvisejí s jazykovou arbitrárností, nebo zda pro ně existuje nějaké jiné vysvětlení (ať už z hlediska jazykové ekonomie, kulturně-historického apod.). Veškeré výpočty a simulace budeme provádět v jazyce Python a R, zdrojové kódy budou k dispozici na CD, které bude přílohou této práce.

[4] Brown, R. W., Black, A. H., & Horowitz, A. E. (1955). Phonetic symbolism in natural languages. *The Journal of Abnormal and Social Psychology, 50*(3), 388–393, https://doi.org/10.1037/h0046820.

1. Ikonicita v názvech měst

Názvy měst byly podrobeny zkoumání již v řadě studií, ať už z hlediska jejich původu, etymologie a lidové etymologie, jazykové ekonomie, politického a ideologického, problematiky překladů a dalších. Jen ojedinělé studie se však zabývají názvy měst/míst z hlediska ikonicity. Ikonicitou zde rozumíme vztah podobnosti mezi objektem a jeho reprezentamenem, jak jej definuje Charles Sanders Peirce.[5] Jedná se tedy o hledání podobnosti mezi názvy měst a městy samotnými nebo jejich vlastnostmi. Pokud by taková podobnost existovala, bylo by to v rozporu s de Saussurovou definicí arbitrárnosti znaku.

Mezi texty zabývající se tímto tématem můžeme zařadit například práce polské autorky M. Rutkiewicz-Hanczewské, která publikovala v roce 2012 studii s názvem *Iconicity in urban place naming (with examples of names from places in Poland)*.[6] Autorka se zabývá ikonicitou nikoli z hlediska jazykové arbitrárnosti ve smyslu významu jednotlivých jednotek (hlásek/písmen apod.) či jejich kombinací, nýbrž z pohledu vzájemného odkazování mezi formami a významy názvů jednotlivých míst a místy samotnými. Jedná se tedy spíše o popisování *motivovanosti* nežli *ikonicity*.

Ve studii se autorka zabývá problematikou pojmenovávání míst ve městech, např. restaurací, obchodů či kaváren. Na samotné názvy měst se však autorka nesoustředí. Jména míst ve městech rozděluje na základě ikonicity na *endoforické* a *exoforické*, které dále dělí do subkategorií podle toho, s jakým typem objektu si jsou jejich forma nebo obsah podobné. Mezi exoforické podle ní můžeme zařadit např. pojmenování *Galeria Drukarnia* [Galerie Tiskárna], *Stary Browar* [Starý pivovar] nebo *Stara Papierna* [Stará papírna].[7] Ve všech uvedených případech se jedná o nákupní centra, v jejichž názvech nebyla reflektována změna funkce budovy. Z příkladů endoforické ikonicity můžeme uvést podnik *Wypas Czarnej Owcy* [Pastva Černé ovce], který se nachází jen několik metrů od hospody pojmenované *Czarna Owca* [Černá ovce].[8] V tomto příkladu se tedy jedná o intertextové odkazování mezi názvy podniků.

---

[5] Peirce, C. S. (1894). *The Art of Reasoning. Kapitola II. What is a sign?* MS [R] 404; MS [R] 1009. Dostupné z: https://peirce.sitehost.iu.edu/ep/ep2/ep2book/ch02/ep2ch2.htm [2. 8. 2021].

[6] Rutkiewicz-Hanczewska, M. (2012). Iconicity in urban place naming (with examples of names from places in Poland). *Semiotica, 2012*(189), 49-64, https://doi.org/10.1515/semi.2011.077. Dostupné se souhlasem autorky z: https://www.researchgate.net/publication/272264080_Iconicity_in_urban_place_naming_with_examples_of_names_from_places_in_Poland [10. 2. 2021].

[7] Srov. *Ibid.*, s. 5.

[8] Srov. *Ibid.*, s. 12.

Odlišným způsobem je ikonicita pojímána ve studii *What's in a Name? Linguistics, Geography, and Toponyms.*[9] Její autoři, L. Radding a J. Western, se v ní zabývají problematikou významu názvů měst i toponym obecně. Toponyma dle *Akademického slovníku cizích slov* definujeme jako „vlastní jméno než. přírodního objektu n. jevu i takového člověkem vytvořeného objektu, kt. je v krajině pevně fixován, geografické (zeměpisné) jméno."[10] Nad těmito tématy se zamýšlejí spíše z intuitivního pohledu, uvažují, proč jsou toponyma pro lidi důležitá. Vybírají také příklady z etymologie a diachronního vývoje názvů. Jména měst pro ně nejsou arbitrární, a to v tom smyslu, že je v rámci konkrétního jazyka nevybíráme náhodně, nýbrž na základě různých konotací, jak vnějších, tak i např. fonetických. V případové studii „New Orleans" demonstrují, jak se historicky měnil název tohoto města, jak změny souvisely s politickými a kulturními vlivy, jaké je spojení mezi toponymem a místem, které je jím označováno atp. Uvažují také nad otázkou, zda by město bylo stále týmž městem, kdyby bylo s veškerými budovami i obyvateli geograficky přemístěno na jiné místo a dospívají k názoru, že nikoli. Také zjišťují, že názvy měst jsou obzvláště úzce spjaty s lidmi, kteří je užívají.[11]

Autoři ve studii zmiňují, že názvy měst nejsou arbitrární kvůli jejich kontextu mezi arbitrárními slovy daného jazyka a jejich zasazení do dané kultury.[12] Toto však není arbitrárnost spojená s principem dvojí artikulace, který arbitrárnost neodmyslitelně doprovází, protože autoři nepracují se samotnými zvuky nebo grafémy v rámci jazykové struktury, nýbrž se zasazením topologických názvů do kontextu jazykového a kulturního. Autoři studie se vůbec nezabývají kvantitativními aspekty, neuvádí žádné statistiky, nýbrž vedou kvalitativní diskuzi, kterou otevírají celé téma arbitrárnosti toponym.

Na rozdíl od autorů obou výše zmíněných studií bychom se chtěli problematikou názvů měst zabývat ze synchronního hlediska. Kontext vzniku názvu je samozřejmě z etymologického pohledu zásadní, ale o to zajímavější bude zjistit, zda i přes nepřeberné množství vnějších faktorů, které volbu a proměny názvu města ovlivňují, nebudou pozorovány určité projevy

---

[9] Radding, L. a Western, J. (2010). What's In the Name? Linguistics, Geography, and Toponyms. *Geographical Review*, 100: 394–412, https://doi.org/10.1111/j.1931-0846.2010.00043.x. Dostupné z: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1931-0846.2010.00043.x?saml_referrer [1.12.2020].

[10] *Internetová jazyková příručka* [online] (2008–2021). Heslo *toponyma.* Praha: Ústav pro jazyk český AV ČR. Dostupné z: https://prirucka.ujc.cas.cz/?slovo=toponyma [1. 8. 2021].

[11] Srov. Radding, L. a Western, J. (2010). What's In the Name? Linguistics, Geography, and Toponyms. *Geographical Review*, 100: 394–412, https://doi.org/10.1111/j.1931-0846.2010.00043.x. Dostupné z: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1931-0846.2010.00043.x?saml_referrer [1.12.2020].

[12] Srov. *Ibid., s.6.* „*We have already noted that place-names, because of their context among the arbitrary words of a given language and their situation in a given culture, are not arbitrary.*"

nearbitrárnosti, například korelace mezi počtem obyvatel města a délkou jeho názvu. Jedná se tedy o zkoumání arbitrárnosti v de saussurovském slova smyslu, že neexistuje žádný vnitřní vztah mezi ideou slova a sledem hlásek, které jí slouží jako označující.[13]

1.1. Korelace mezi počtem obyvatel obce a délkou jejího názvu

V následující části této práce popíšeme sérii menších pokusů, ve kterých bude testována možnost odmítnutí arbitrárnosti na základě korelací mezi různými kvantifikovatelnými vlastnostmi měst a jinými kvantifikovatelnými vlastnostmi jejich názvů a jejich statistického vyhodnocení. Jako základní dataset názvů měst byla použita data Českého statistického úřadu, konkrétně Tab. 3 – *Počet obyvatel v obcích České republiky k 1. 1. 2020*.[14]

1.1.1. Korelace mezi počtem obyvatel obce a počtem grafémů v jejím názvu

Jako první byla zkoumána korelace mezi počtem obyvatel obce a délkou jejího názvu, přičemž délka byla určena počtem grafémů tvořících název obce. Na základě de Saussurovy definice lingvistického znaku jako arbitrárního by mělo platit, že také názvy obcí budou arbitrární. Jak již bylo zmíněno výše, zabýváme se možností, že i přesto může být v názvech nějakým způsobem skryta nearbitrárnost ve formě latentní tendence, která by se mohla manifestovat formou statisticky průkazných korelací.

Pro 6 258 obcí ČR byly z dat Českého statistického úřadu získány počty obyvatel. Pro každé jméno jsme na základě počítačového softwaru zjistili jeho délku, tj. počty grafémů v názvu obce.[15] Ze získaných dat byl následně spočítán Spearmanův korelační koeficient[16] a jeho testové statistiky a pro vizuální kontrolu byl vytvořen bodový graf. Byla zjištěna p-hodnota, která udává pravděpodobnost, s jakou by stejná nebo ještě extrémnější korelace mohla vzniknout pouze vlivem náhody, přičemž hladinu alfa zvolíme na typickou hodnotu 0,05. Následuje přehled výsledků a jejich interpretace.

---

[13] Saussure, F., Bally, C., Sechehaye, A. (1996). *Kurs obecné lingvistiky*. Přeložil Čermák, F. Praha: Academia, s. 98.

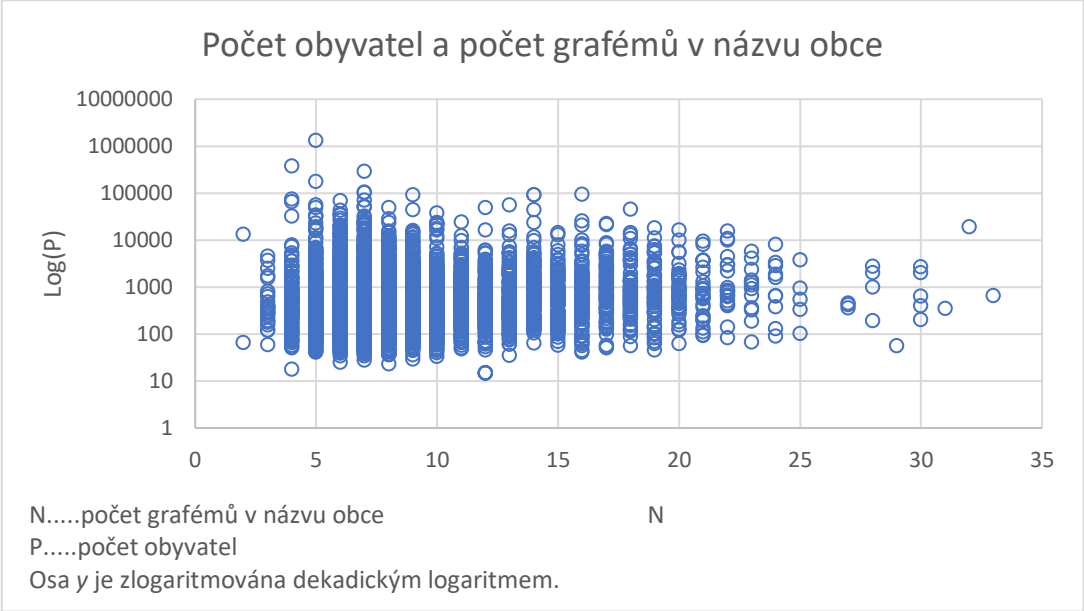
[14] Český statistický úřad (2020). *Tab. 3 Počet obyvatel v obcích České republiky k 1. 1. 2020*. Dostupné z: https://www.czso.cz/csu/czso/pocet-obyvatel-v-obcich-k-112019 [2. 12. 2020].

[15] Název jsme chápali jako jeden dlouhý řetěz grafémů, z tohoto důvodu jsou do jejich počtu zahrnuty i mezery, např. pro *České Budějovice* tak počítáme s délkou 16.

[16] Funkce vytvořená podle Best, D. a Roberts, D. (1975). Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 24*(3), 377-379, https://doi.org/10.2307/2347111 a Hollander, M. a Wolfe, D. A. (1973). *Nonparametric Statistical Methods.* New York: John Wiley & Sons, s. 185–194 (Kendall and Spearman tests).

Na první pohled není z bodového grafu (*Graf 1*) žádná výrazná korelace mezi počtem obyvatel obce a počtem grafémů v jejím názvu zřejmá. Po zlogaritmování osy *y* vidíme, že tvar grafu přibližně odpovídá na pravé straně nedokončenému a uprostřed zúženému písmenu *U*. Hodnota Spearmanova korelačního koeficientu 0,096 je slabá, přesto je třeba pomocí p-hodnoty statisticky ověřit, zda by stejná nebo ještě méně pravděpodobná korelace mohla vzniknout jen vlivem náhody. Kladná hodnota Spearmanova koeficientu znamená, že čím více obyvatel v obci žije, tím delší je její jméno, a naopak.

Stanovme nulovou hypotézu, kterou je statisticky nevýznamná korelace mezi délkou názvu obce a počtem jejích obyvatel. Alternativní hypotézou je potom statisticky významná korelace mezi těmito jevy.



Graf 1: Závislost počtu obyvatel obce a délky jejího názvu.

Výsledkem testu signifikance korelace je p-hodnota je $2{,}021\times10^{-14}$. Vzhledem ke zvolené hladině alfa = 0,05 budeme hodnotu $2{,}021\times10^{-14} < 0{,}05$ považovat za statisticky významnou, proto můžeme odmítnout nulovou hypotézu a přijmout alternativní hypotézu, podle které existuje statisticky významná korelace mezi počtem obyvatel obce a počtem grafémů v jejím názvu.

Přestože tedy Spearmanův koeficient (0,096) ukazuje velmi slabou korelaci, ukazuje se, že stejná nebo extrémnější korelace by ve stejném případě vznikla pouze vlivem náhody s pravděpodobností $2{,}021\times10^{-14}$, tedy s pravděpodobností blížící se nule, což je výsledek v rozporu s naším očekáváním.

1.1.2. Diskuze nad zjištěnou korelací

Existence signifikantní korelace nevylučuje přítomnost pravidla, které by definovalo, jak má být dlouhé jméno ve vztahu k počtu obytvatel. To můžeme upřesnit na pět různých možností z hlediska případné kauzality:

1) Počet obyvatel způsobuje délku názvu obce.
2) Délka názvu obce způsobuje počet obyvatel.
3) Existuje třetí faktor způsobující obojí.
4) Obce s takovou korelací máme jen vlivem štěstí, které se děje v $2{,}021 \times 10^{-14}$ případů.
5) Selhání statistických postupů.

Ačkoliv se uvedené varianty mohou zpočátku jevit téměř bizarně, je třeba provést diskuzi, proč by k takové kauzalitě mohlo dojít. Zároveň je možné, že najdeme souvislost s jazykovou ekonomií nebo jiné behaviorální vysvětlení.

Podívejme se blíže na první možnost, tedy jak by počet obyvatel mohl způsobovat délku názvu obce. Jedním z nejznámějších konceptů operujících s délkami slov a jejich využitím (v analogii s počtem obyvatel) je Zipfův princip jazykové ekonomie. Zipf jej popisuje ve své knize *The Psycho-Biology of Language*, kde říká, že délka slova má tendenci mít inverzní vztah k jeho relativní frekvenci,[17] a dále jej rozpracovává v knize *Human behaviour and the principle of the least effort*. V kontextu jím pozorovaného principu nejmenšího úsilí[18] definuje Zákon o zkracování slov.[19] Na začátku autor předpokládá inverzní vztah mezi délkou slov a frekvencí jejich užívání z důvodu vykonání co nejmenší práce.[20] Po provedení pokusu, ve kterém zkoumá frekvence slov na dvou datasetech, analýze amerických novin od R. C. Eldridge a Plautových latinských slovech, považuje tento vztah za dokázaný.[21]

---

[17] Srov. „In view of the evidence of the stream of speech we may say that the length of a word tends to bear an inverse relationship to its relative frequency.“ Zipf, G. K. (1936). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton-Mifflin, Boston, s. 38.

[18] The Principle of Least Effort, Zipf, G. K. (1949). *Human behaviour and the principle of the least effort*. Addison-Wesley Press, s. 18.

[19] The Law of Abreviation of Words, viz *Ibid.*, s. 88.

[20] Srov. „From our foregoing argument, we should anticipate mutatis mutandis that there will be an inverse relationship between the lengths of words and the frequencies of their usage if we assume – quite correctly, I believe – that, under otherwise constant conditions, the work of uttering a longer word is greater than the work of uttering a shorter one.“ Viz *Ibid.*, s. 88.

[21] Srov. „In both sets of data of Table 3-1 we find an unmistakable inverse relationship between the lengths of words and their frequency of occurrence as we have anticipated theiretically on the basis of our Tool Analogy, which is thereby confirmed.“ Viz *Ibid.*, s. 94.

Odtud pak můžeme uvážit obdobný vztah. Vezmeme-li v úvahu, že čím více osob bydlí v obci, tím více je používáno její jméno – např. při adresování poštovních zásilek, vyplňování formulářů, uvádění sídel firem, sledování politického či kulturního dění nebo při běžném hovoru jak mezi místními, tak mezi lidmi žijícími v jiných obcích, můžeme usoudit, že počet jejích obyvatel by měl navyšovat frekvenci, se kterou se jméno obce používá, a ta by tak měla přibližně proporčně odpovídat tomu, jak často je název obce užíván. Dle Zipfova zákona tedy očekáváme, že Spearmanův koeficient bude záporný, aby bylo vyhověno pravidlu, že čím častěji je slovo používáno, tím je kratší.

Zde se však dostáváme do sporu. G. K. Zipf tvrdí, že mezi délkou slov a jejich frekvencí se jedná o vztah *záporné* korelace, zatímco v našem pokusu vyšla korelace *signifikantně pozitivní*. Možné příčiny tohoto rozporu jsou tři:

a) Zipfovo tvrzení o vztahu délky slova a frekvence neplatí obecně.
b) Výsledek prezentovaný v této práci není správný.
c) Zipfovo tvrzení nelze aplikovat na jména měst.

Možnost a) můžeme vyloučit, protože by byla v rozporu s empirií, jakož i s matematickou tendencí při náhodném generování slov (je jednoduché opětovně náhodně vygenerovat krátká slova shodná s těmi, která se již objevila, zatímco dlouhá již dříve vygenerovaná slova opakovaně vygenerujeme méně často). Tento zákon byl podroben mnoha zkoumáním a například Sigurd, Eeg-Olofsson a van de Weijer ve svém článku *Word Length, Sentence Length and Frequency – Zipf Revisited*[22] nejenže nevyvracejí jeho platnost v anglických, švédských a německých textech, ale dokonce nacházejí vzorec, kterým je možné popsat vztah mezi délkou slov a jejich frekvencí $f_{exp} = a * L^b * c^L$.[23]

Autoři Bentz a Ferrer-i-Cancho v článku *Zipf's law of abbreviation as a language universal*[24] z roku 2016 testují přítomnost negativní korelace na vzorku 1 262 textů v 986 různých jazycích.

[22] Sigurd, B., Eeg-Olofsson, M. a Van Weijer, J. (2004), Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica*, 58: 37-52. https://doi.org/10.1111/j.0039-3193.2004.00109.x Dostupné z: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.0039-3193.2004.00109.x?saml_referrer [7. 4 2021].

[23] Srov. *Ibid.*, s. 1.

[24] Bentz, C. a Ferrer-i-Cancho, R. (2016). *Zipf's law of abbreviation as a language universal*. http://dx.doi.org/10.15496/publikation-10057. Dostupné z: https://upcommons.upc.edu/bitstream/handle/2117/178845/Bentz_Ferrer.pdf?sequence=1&isAllowed=y [7. 4. 2021].

Ve všech z nich je platnost zákona potvrzena a autoři článku dospívají k závěru, že se jedná o jednu z jazykových univerzálií.

Možnost b), že náš výsledek není správný, nepředpokládáme, přestože ji nelze s jistotou vyloučit. Chyby nebo opomenutí se mohly vyskytnout hned v několika oblastech. Mohli jsme pracovat se špatnými daty. Tuto chybu s ohledem na Český statistický úřad nepředpokládáme, navíc uvážíme-li, že v tomto případě bylo nemožné vybrat špatný vzorek, protože jsme pracovali se všemi obcemi ČR. To by mohlo představovat problém pouze v případě, že bychom výsledky pokusu chtěli zobecnit, protože v rámci všech obcí na světě nejsou všechny obce ČR reprezentativní.

Další chyba v práci s daty mohla vzniknout tím, že v běžném jazyce se neužívají celé názvy měst, nýbrž zkrácené tvary. Podívejme se na příklad *Českých Budějovic*. Dle dat Českého národního korpusu v mluveném projevu 56 % mluvčích užívá výraz *Budějovice*[25] a 44 % mluvčích výraz *Budějice*, zatímco v psaném projevu se téměř ve 100 % případů užívá výrazu *Budějovice*. Navíc v mluveném projevu je první část názvu *České* zachována pouze ve 12,5 % případů.[26] Existuje tedy rozpor mezi standardizovanou formou a běžným užíváním jazyka.

Chybné, nebo přinejmenším problematické, by mohlo být také určení délky slova pomocí grafémů. Vzhledem k tomu, že každý znak byl chápán jako grafém, bylo např. písmeno *ch* počítáno jako dva grafémy, přestože se jedná pouze o jeden fonetický znak. Podobně tomu bude v případech, kdy bude namísto dvou fonémů vysloven pouze jeden (např. skupinu *ts* v názvu *Rohatsko* je možné vyslovit jako [c]). Do délky byla počítána také mezera, aby bylo nějakým způsobem zohledněno, kolika slovy je název tvořen. Pracujeme-li však výhradně s psanou podobou jazyka, nepovažujeme tyto překážky za tolik zásadní, aby ovlivnily celkový výsledek.

Chybný mohl být konečně i samotný výpočet. To však s ohledem na výše popsaný postup a výše citované zdroje nepředpokládáme.

Přejděme k možnosti c), že Zipfovo tvrzení platí a náš výsledek je správný, avšak Zipfovo tvrzení nelze na seznam jmen měst aplikovat. Zde je třeba položit si otázku, proč by tomu tak mohlo být.

---

[25] Existují i *Moravské Budějovice*, které by mohly naše data zkreslit, dle dat Českého národního korpusu jsou však v 95,73 % zmiňovány České Budějovice, nikoli Moravské.

[26] Srov. Cvrček, V. a Vondřička, P. (2011). SyD – Korpusový průzkum variant. FF UK. Praha 2011. Dostupné z: https://syd.korpus.cz/ [4. 4. 2021].

Můžeme uvážit, že slova v jazyce nejsou bezkontextová, Zipfovo tvrzení tedy nelze použít na jednotlivá slova vytržená z kontextu. Jména měst by bylo třeba chápat v rámci celé jazykové struktury jako ta slova, která nepotřebují být tak často používána ve srovnání s nejfrekventovanějšími částmi lexika daného jazyka. Budeme-li tento argument považovat za správný, při hlubším uvážení stále neřeší problém, proč existují i málo frekventované názvy, které jsou krátké (např. *Důl, Lom, Buk, Láz, Peč* nebo *Tuř*). Ověření vlivu vytržení z kontextu na korelace mezi délkami slov a jejich frekvencemi se budeme blíže zabývat v další podkapitole.

Vzhledem k tomu, že z formálního hlediska jazykové ekonomie nenacházíme odpověď, zaměříme se na data samotná. Tvar rozmístění bodů (*Graf 1*) přibližně odpovídá tvaru neúplného písmene U. Zde by se nabízelo vysvětlení, že zjištěný vztah délky názvu obce a počtu obyvatel může souviset s potřebou upřesňovat, o kterou konkrétní obec se jedná. Je-li obec malá a významná pouze pro velmi malý počet lidí, není zde potřeba upřesňovat, který Buk či Lom máme zrovna na mysli, protože jeho totožnost je malému počtu místních dobře známá, zatímco pro ostatní je její existence natolik nevýznamná, že o ní nemusí ani vědět. Taková tendence by zapříčinila levou dolní část grafu, kdy platí, že obce s malým počtem obyvatel mají krátký název. Větší obce již potřebují buď název, který žádná jiná obec nenese (např. *Neratovice*), nebo jméno upřesnit tak, aby bylo jasné, o kterou z několika stejnojmenných obcí se jedná (např. *Lomnice nad Popelkou* nebo *Klášterec nad Ohří*). Obě dvě možnosti automaticky implikují potřebu vyššího počtu grafémů. To by zapříčiňovalo pravou část grafu, kde platí, že obce se středním až vyšším počtem obyvatel mají delší názvy. Velká a významná města potom už mohou mít opět názvy kratší, protože jsou natolik unikátní, že zde záměna nehrozí (např. *Praha*, *Brno* nebo *Zlín*). Tomuto trendu by odpovídala levá horní část grafu, kdy velké obce mají opět kratší názvy, ale nikoli tak krátké, jako nejmenší obce.

Budeme-li však toto vysvětlení považovat za správné, narazili jsme na protipříklad jazykové arbitrárnosti, protože malé nevýznamné obce mají krátké názvy, zatímco větší obce mají delší názvy, až na ty úplně největší.

Vraťme se nyní k druhé nabízené možnosti, jak nahlížet na kauzalitu v tomto problému, a to že délka názvu obce způsobuje počet obyvatel. Jednalo by se o případy, kdy se lidé budou do určité obce stěhovat kvůli tomu, že se jim například líbí její název, nebo že se v ní bude rodit více dětí. Takovéto možnosti zní poněkud absurdně, avšak nemůžeme je vyvrátit, nebudeme-li mít k dispozici dostatečná data.

Proto z databáze demografických údajů za obce ČR Českého statistického úřadu získáme celkový přírůstek obyvatel za rok 2017[27], do kterého jsou započítáni narození, zemřelí, přistěhovalí a vystěhovalí obyvatelé pro 6258 obcí ČR. Získáme počty grafémů reprezentující jednotlivé obce a spočítáme korelaci mezi celkovým přírůstkem obce a délkou jejího názvu. Dle očekávání získáváme nesignifikantní korelaci –0,003 s p-hodnotou 0,812, která říká, že takováto nebo ještě větší korelace by s 81,2% pravděpodobností vznikla pouze vlivem náhody. Možnost, že délka názvu obce způsobuje počet obyvatel, můžeme tudíž vyloučit, což je v souladu s intuitivním očekáváním.

Třetí možností z hlediska případné kauzality mezi počtem obyvatel města a délkou názvu je existence třetího faktoru, který způsobuje oba dva jevy. Tím by hypoteticky mohla být nějaká vlastnost obce, která ovlivní to, že v ní lidé budou chtít žít, a zároveň bude reflektována v jejím názvu, čímž implicitně ovlivní jeho délku. Příkladem by mohl být *Havířov*, ve kterém byly vybudovány doly, a proto se tam stěhovali horníci za prací. Zároveň je tato vlastnost města reflektována v jeho názvu (havíř = horník) a tím implicitně ovlivňuje i jeho délku. Při hlubším uvážení je to však velice slabý argument hned ze dvou důvodů:

1) V Havířově žijí i jiní lidé než ti, kteří tam přišli kvůli dolům.
2) Takovouto motivovanost najdeme jen u velmi malého počtu obcí. Názvy sice často odrážejí nějakou vlastnost města, ta však nijak nesouvisí s tím, proč by v ní lidé chtěli žít.

Může však existovat nějaký jiný třetí faktor, o jehož existenci nevíme.

Čtvrtou možností pak je, že korelace vznikla jen vlivem náhody (např. štěstí), která nastane v $2{,}021\times10^{-14}$ případů. Jinými slovy, tato možnost s 99,99999999998% pravděpodobností nenastává.

Pátou možností je selhání statistických postupů. Statistické metody mohly z nějakého nám neznámého důvodu na tomto konkrétním případě selhat, mohly se vyskytnout systematické drobné chyby v datech apod.

1.1.3. Korelace mezi počtem obyvatel obce a počtem slov v jejím názvu

Cílem dalšího pokusu je zjistit korelaci mezi počtem obyvatel v obci a počtem slov v jejím názvu, aby tím bylo ukázáno, zda je vztah mezi nimi náhodný, či zda se zakládá na případné

[27] Český statistický úřad (2018). *Stav a pohyb obyvatelstva v ČR – rok 2017*. Dostupné z: https://www.czso.cz/csu/czso/stav-a-pohyb-obyvatelstva-v-cr-rok-2017 [14. 5. 2020].

ikonicitě. Očekáváme, že žádná průkazná korelace nebude zjištěna, protože vztah mezi označujícím (název obce) a označovaným (obec samotná nebo její vlastnost, zde počet obyvatel), má být arbitrární.

Jedná se o podobný případ jako v předchozím pokusu (viz kapitola 1.1.1.), ale pracujeme s rozdílnými jednotkami. Problém je zde zredukován z grafémů na počty slov. Pro porovnání dat předchozího a tohoto pokusu změříme korelaci mezi počtem grafémů a počtem slov v názvu obce. Výsledkem je signifikantně pozitivní korelace s hodnotou Spearmanova koeficientu 0,645 a p-hodnotou $p < 2{,}2\times10^{-16}$. To znamená, že počet slov a počet grafémů spolu korelují, nicméně test je třeba provést, protože tato korelace automaticky neimplikuje, že čím více grafémů je v názvu, tím víc v něm je i slov. Např. obce *Postoloprty* a *Bor* mají velice rozdílný počet grafémů (11 a 3 grafémy), zatímco počet slov je stejný. Výsledky pokusu by se tak mohly lišit.

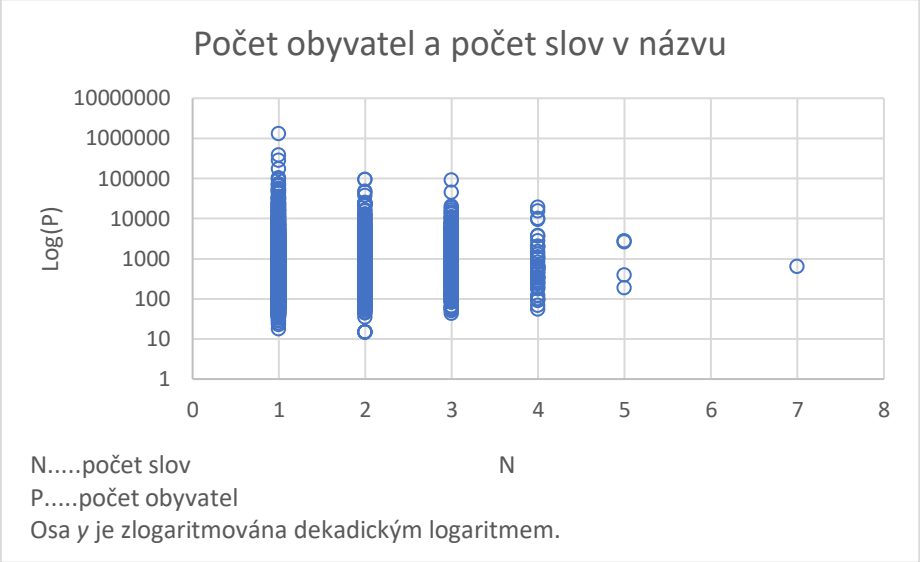
Obdobně jako v předchozím pokusu spočítáme Spearmanův koeficient na počtu obyvatel a počtu slov v názvu obce. Získáme hodnotu Spearmanova koeficientu 0,118, která značí, že se jedná o pozitivní korelaci, a p-hodnotu $p < 2{,}2\times10^{-16}$, která představuje extrémně nízkou hodnotu blížící se nule, která jinými slovy říká, že pravděpodobnost, že by takováto korelace vznikla pouze vlivem náhody, se blíží nule.

Všimněme si, že korelace je opět signifikantně pozitivní, jak tomu bylo v předchozím pokusu, nikoli signifikantně negativní, jak by to vyžadovala výše zmíněná Zipfova definice, ani nesignifikantní, jak by to vyžadovala de Saussurova definice arbitrárnosti znaku. Interpretace těchto výsledků je vzhledem ke shodné metodice a blízké povaze dat prakticky totožná s interpretací vztahu délky názvů měst v grafémech a počtu obyvatel (viz kapitola 1.1.2.) jen s tím rozdílem, že je diskuze o počtech slov, nikoli o počtu grafémů.

Data pro větší názornost znázorníme v grafu (*Graf 2*). Na rozdíl od grafu, který znázorňoval vztah mezi počtem grafémů názvu a počtem obyvatel obce (*Graf 1*) a v němž rozložení bodů odpovídalo písmeni *U*, zde vidíme rozložení ve tvaru ostrého *V*[28], které je po logaritmizaci osy *y* symetrické. Z tvaru je patrné, že města s největším a nejmenším počtem obyvatel, tj. dva extrémy, mají název tvořený jedním slovem. S přibývajícím počtem slov v názvu obce se

[28] Vrchol písmene *V* vznikl díky české obci s největším počtem slov v názvu (7). Jedná se o *Novou Ves u Nového Města na Moravě.*

postupně zmenšuje vzdálenost mezi oběma extrémy počtu obyvatel v množině obcí, jejichž názvy obsahují stejný počet slov.



Graf 2: Závislost počtu obyvatel obce a počtu slov v jejím názvu.

1.1.4. Vliv kontextu

Během diskuze nad zjištěnou korelací jsme narazili na otázku, zda je možné uplatnit Zipfův zákon o zkracování slov na seznam slov vytržených z kontextu. Jazyk totiž existuje a přirozeně se vyvíjí a mění tím, že jej lidé používají, proto uvažme, že by zákonitosti, které platí v přirozeném jazyce, nemusely platit v uměle vytvořených seznamech bezkontextových slov.

Podívejme se proto na tři analogické problémy. Obdobným způsobem, který byl použit výše, zkusíme spočítat korelaci mezi dalšími jmény, u kterých můžeme měřit popularitu nebo frekvenci. Vzhledem k množství dat, která můžeme získat, to jsou:

1) korelace mezi délkou názvu státu a počtem jeho obyvatel,
2) korelace mezi délkou názvu filmu a jeho oblíbeností,
3) korelace mezi délkou názvu seriálu a jeho oblíbeností.

V prvním případě jako dataset použijeme 235 států (a závislých území)[29] světa.[30] Pracujeme se jmény států v anglickém jazyce. Podobně jako u měst získáme počty grafémů, ze kterých se jména států skládají, a spočítáme korelaci mezi počtem grafémů jména státu a populací státu.

[29] Vzhledem k tomu, že se zabýváme problematikou lingvistickou a nikoli politickou, nebereme v potaz problémy uznání samostatnosti státu, politickou situaci či právní status daného území.

[30] Worldmeters.info (2021). *Countries in the world by population.* Dover, Delaware, USA. Dostupné z: https://www.worldometers.info/world-population/population-by-country/ [31. 3. 2021].

Získáváme signifikantně zápornou korelaci s hodnotou Spearmanova koeficientu –0,294 a p-hodnotu $4{,}505\times10^{-6}$.

Ve druhém a třetím případě zkoumáme korelaci mezi délkou názvu filmu (resp. seriálu) a jeho oblíbeností, která byla reprezentována počtem jeho fanoušků na základě databáze ČSFD.[31] Oba datasety obsahovaly 300 filmů (resp. seriálů). Jejich názvy byly uvedeny v českém jazyce. Pro filmy získáváme Spearmanův koeficient 0,032 s p-hodnotou 0,578. Taková nebo extrémnější korelace by tedy s více než 57% pravděpodobností vznikla pouze vlivem náhody. Pro seriály získáváme Spearmanův koeficient s hodnotou 0,02 a p-hodnotu 0,73. Jedná se opět o nesignifikantní korelaci. Přehled korelací je uveden v následující tabulce (*Tab. 1*).

Přehled korelací mezi délkami názvů objektů a jejich vlastnostmi

Vlastnost 1	**Vlastnost 2**	**Spearmanův koeficient**	**P-hodnota**	**Korelace**
Počet grafémů v názvu obce	Počet obyvatel obce	0,096	$2{,}021\times10^{-14}$	Signifikantně pozitivní
Počet grafémů v názvu státu	Počet obyvatel státu	–0,294	$4{,}505\times10^{-6}$	Signifikantně negativní
Počet grafémů v názvu filmu	Oblíbenost filmu (počet fanoušků)	0,032	0,578	Nesignifikantně pozitivní
Počet grafémů v názvu seriálu	Oblíbenost seriálu (počet fanoušků)	0,02	0,73	Nesignifikantně pozitivní

Tab. 1: Srovnání korelací mezi různými vlastnostmi názvů a objektů čtyř datasetů.

Zjistili jsme, že vztah mezi počtem grafémů v názvu státu a počtem jeho obyvatel není nahodilý (resp. vlivem náhody by vznikl s pravděpodobností $4{,}505\times10^{-6}$). Mezi těmito dvěma jevy existuje signifikantně negativní korelace, která říká, že čím více lidí v daném státě žije, tím kratší je jeho název. Tento jev lze vysvětlit Zipfovým tvrzením o jazykové ekonomii. Vztah mezi počtem grafémů a oblíbeností filmu (resp. seriálu) je čistě arbitrární, což odpovídá jak intuitivnímu očekávání, tak i de Saussurově definici arbitrárnosti jazykového znaku.

Oběma pravidlům se nadále vymyká pouze původní vztah mezi délkou názvu obce a počtem jejích obyvatel, na který nelze aplikovat ani de Saussurovu definici arbitrárního znaku, ani Zipfův princip jazykové ekonomie. To by mohlo nastiňovat možnost, že jména obcí jsou v rámci jazykového systému něčím specifickým. Zároveň ale nepřítomnost záporné

[31] Česko-Slovenská filmová databáze (2001–2021). *Žebříčky – nejlepší filmy.* Dostupné z: https://www.csfd.cz/zebricky/filmy/nejlepsi/ a Česko-Slovenská filmová databáze (2001–2021). *Žebříčky – nejlepší seriály.* Dostupné z: https://www.csfd.cz/zebricky/serialy/nejlepsi/ [12. 3. 2021].

signifikantní korelace u filmů a seriálů potvrzuje, že Zipfův princip jazykové ekonomie nelze automaticky aplikovat na námi zvolený seznam dílčích slov vytržených z kontextu, jak jsme to udělali u obcí.

1.1.5. Limity uvedených pokusů

Výše uvedené experimenty mají své limity. První úskalí se týká univerzálnosti zjištěných výsledků. V těchto dílčích pokusech jsme pracovali s názvy všech současných obcí České republiky, tedy s 6 258 vzorky. Přestože se jedná o všechny obce ČR, je tento počet v rámci počtu všech obcí na celém světě zcela nepatrný. Pro srovnání, jen v celé Francii bylo ke dni 1. 1. 2018 dle oficiálních dat 35 357 obcí.[32] Pokud bychom chtěli tvrdit, že existuje univerzálně platná závislost například mezi délkami názvů měst a počty jejich obyvatel, bylo by třeba tuto hypotézu ověřit s mnohem větším a stratifikovaným vzorkem dat, ve kterém budou jednotlivé země zastoupeny v poměru podle toho, kolik obcí se v nich nachází.

Dalším limitem by mohla být nejednoznačnost jednotek měřených především v pokusu s počtem slov. Narážíme zde na zásadní lingvistický problém, kterým je neexistence jasné definice toho, co je to slovo.

Ve *Slovníku spisovného jazyka českého* nalezneme definici, že slovo je „sled hlásek (zř. jen jedna hláska), kt. tvoří v jazyce ustálený a ve větě přemístitelný celek mající zřejmý význam věcný i mluvnický.“[33] Tato definice nám nicméně nepomůže v některých případech přesně rozlišit, zda se pro účely jazykové analýzy jedná o jedno slovo či nikoli. Proto se podíváme na texty různých autorů, kteří se touto problematikou zabývají.

V úvodu knihy *Word Frequency Studies* shrnují její autoři největší problém výzkumu v oblasti frekvencí slov, a to, že neumíme říci, co je to slovo. Abychom se vyhnuli esencialismu, měli bychom se spíše ptát, co můžeme za text nebo slovo považovat. Je třeba si uvědomit, že *slovo* je koncepčně vytvořeným kritériem. Není obsaženo ve sledovaných entitách, nýbrž slouží jako prostředek pro získání dat. Data tedy nejsou nalezena, nýbrž zkonstruována.[34]

---

[32] *Srov.* Collectivités locales. *Les chiffres-clés des collectivités locales* (2018). Dostupné z: https://www.collectivites-locales.gouv.fr/files/files/statistiques/brochures/chapitre_1_-_les_chiffres_cles_des_collectivites_locales_1.pdf [18. 2. 2021].

[33] Ústav pro jazyk český ČSAV (2011). *Slovník spisovného jazyka českého*. Dostupné z: https://ssjc.ujc.cas.cz/search.php?db=ssjc [2. 8. 2021].

[34] Srov. „We must be aware of the fact that our definitions are not reflections of reality. They are conceptually constructed criteria which are not contained in the observed entities but serve as means for constructing data.

Slova v psaném textu můžeme jednoduše definovat jako posloupnost písmen mezi oddělovači slov. Tyto oddělovače však nejsou přítomné ve všech jazycích, a navíc toto grafické oddělení ne vždy odpovídá slovu. Autoři vyjmenovávají řadu příkladů, kdy je minimálně sporné, zda se jedná o jedno či více slov, např. stažené tvary (*we're* [my jsme] v angličtině), nespojité jednotky (např. *ne ... pas* tvořící ve francouzštině zápor), složená slova, která se v některých případech píšou v jednom celku a jindy zvlášť (např. německé *Sie will etwas abschreiben* [chce něco opsat] X *Sie schreibt etwas ab* [něco opisuje]), tvoření otázek přidáváním morfémů (např. v maďarštině) nebo samostatných slov (např. v japonštině).

Autoři se dále věnují dalším problémům frekvenční analýzy, např. lemmatizaci, přiřazování slovních druhů a mnoha dalším. Tato témata už však s naším problémem určení počtu slov přímo nesouvisí.

Grefenstette a Tapanainen se ve své studii *What is a word, What is a sentence? Problems of Tokenization*[35] z roku 1994 zabývají definicí slova z hlediska komputační lingvistiky. Zaměřují se na tokenizaci, která bývá podle autorů v učebnicích lingvistiky rychle odbyta jako „relativně nezajímavý krok předzpracování provedený před samotnou lingvistickou analýzou,“[36] zatímco ve skutečnosi se jedná o netriviální problém, a cílem autorů je nalézt vhodnou jednotku pro tokenizaci. Nejprve se zabývají vymezením hranic věty, která je relativně jasně ukončena vykřičníkem nebo otazníkem, avšak u tečky se mohou vyskytnout ambiguidní případy. Tečku mohou obsahovat např. i některé zkratky (*m.p.h.*), alfanumerické reference (*T-1-AB.1.2*), desetinná čísla, data a podobně. Některé z nich lze však pomocí regulárních výrazů relativně snadno rozpoznat a určit, že se nejedná o tečku ukončující větu.

V případě slov je problém složitější. Je třeba rozhodnout např. o tom, zda bude řetězec písmen oddělený apostrofy považován za jeden, nebo dva tokeny (např. *governor's*, ale také *it's, that's*). Rozhodnutí se ale bude lišit pro konkrétní jazyky. Třeba ve francouzštině bude určení pravidel ještě složitější vzhledem k širšímu využití apostrofu (např. *l'addition* [*součet* – člen

---

Consequently, data are not found but constructed.“ Popescu, I. (2009). Word Frequency Studies. Berlin, New York: De Gruyter Mouton, https://doi.org/10.1515/9783110218534, s. 5.

[35] Grefenstette, G. a Tapanainen, P. (1997). What is a word, What is a sentence? Problems of Tokenization. Dostupné z: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.8947&rep=rep1&type=pdf [3. 8. 2021].

[36] Srov. „a relatively uninteresting pre-processing step performer before linguistic analysis is undertaken.“ *Ibid.,* s. 1.

určitý + substantivum], *m'appelle* [*jmenuji se* – zvratné se + sloveso], *presqu'île* [*poloostrov* – slovo složené]).

V závěru autoři vyjadřují názor, že proces tokenizace by měl být koncipován jako řada jednotlivých filtrů, kterými může dle okolností text projít či nikoli.[37] Nejednoznačné případy mohou být vyřešeny pomocí analýzy struktury vloženého textu. Řešení je vždy třeba upravit v závislosti na kontextu a neexistuje jednoznačná univerzální odpověď na otázky, které znaky považovat za oddělovače slov a které nikoli.

Význam *slova* byl předmětem diskuze také v československém prostředí. Schwanzer v příspěvku *Slovo a text* (1978) zdůrazňuje, že „analýzou textu získané nadvětné, větné a slovní jednotky mají charakter (lingvistou) subjektivně stanovených a intersubjektivně uznaných částí celého textu, vymezených s jistým zřetelem na jejich postavení v celku, na samotný celek a na účely a možnosti s nimi operovat."[38] Slova se vyznačují sémantickými a gramatickými vlastnostmi, které je třeba brát v úvahu při analýze textu. Pokud lingvista pracuje v analýze se slovy jako jednotkami, musí si uvědomit, že vynětím slova z textu ztrácí slovo konkrétní denotační charkter.[39]

Přestože uvažování slova jako řetězce písmen odděleného mezerou je, jak jsme viděli ve výše citovaných textech, vysoce problematické, v našem případě jsme kvůli povaze dat a z praktických důvodů toto řešení zvolili. Lze však spekulovat o tom, zda například předložky lze považovat za slova, či nikoli.

1.1.6. Shrnutí první kapitoly

V první kapitole jsme se zabývali vztahem mezi délkou názvu obce a počtem jejích obyvatel. Na základě de Saussurovy definice arbitrárnosti znaku bylo očekáváno, že vztah mezi těmito dvěma vlastnostmi bude nahodilý, což by se mělo manifestovat prostřednictvím nesignifikantních korelací, popřípadě že budou v souladu se Zipfovou definicí zákona o zkracování slov signifikantně negativní. Namísto toho byly odhaleny signifikantně pozitivní korelace mezi počty obyvatel obcí a délkami jejich názvů měřenými v počtech grafémů a

[37] Srov. *Ibid.*, s. 9

[38] Srov. „*Analýzou textu získané nadvetné, vetné a slovné jednotky majú charakter (lingvistom) subjektívne stanovených a intersubjektívne uznaných častí celého textu, vymedzených s istým zreteľom na ich postavenie v celku, na celok sám a na účely a možnosti operovania s nimi.*" Schwanzer, V. (1978). Slovo a text. *Slovo a slovesnost, 39 (1978), 3-4, s. 259-261.* Dostupné z: http://sas.ujc.cas.cz/archiv.php?art=2548 [3. 4. 2021].

[39] Srov. *Ibid.*

počtech slov, což je v rozporu s naším očekáváním a nastínilo to možnost, že v jazyce, pokud všechny testy proběhly správně, nemusí být vše arbitrární.

Zabývali jsme se i myšlenkou, že slova v jazyce nejsou bezkontextová, a proto na seznam slov vytržených z kontextu nelze aplikovat lingvistické zákony. Abychom tuto myšlenku mohli rozvinout a ověřit, měřili jsme stejným způsobem korelace mezi délkami názvů filmů a seriálů a jejich oblíbeností reprezentovanou počtem fanoušků a mezi délkami názvů států světa a počty obyvatel. Ani v jednom případě jsme však neodhalili signifikantně pozitivní korelaci, jak tomu bylo v případě jmen obcí, což vede k myšlence, že názvy obcí by v rámci jazykové struktury mohly být specifickými slovy, na něž nelze automaticky aplikovat empirické lingvistické zákony.

1.2. Hlavní města: trénování SVM modelu

Od zjišťování korelací se nyní přesuňme k odlišným metodám. V následujícím pokusu budeme trénovat model metodou podpůrných vektorů (*Support Vector Machines*, SVM),[40] aby se naučil rozlišovat mezi hlavními a periferními městy, a následně budeme zjišťovat úspěšnost jeho predikcí pro zatím neviděné vzorky. Jinými slovy, klademe si otázku, zda mají názvy hlavních měst nějaké specifické vlastnosti, které názvy ostatních měst nemají, a podle kterých by byl počítač schopný rozeznat, že se jedná o hlavní město, a zda tyto vlastnosti dokáže odhalit. Jsou-li však jména měst arbitrární, SVM model by neměl být schopen rozlišit mezi hlavním a jiným městem lépe než náhodný model.

SVM je jednou z metod strojového učení s učitelem (*supervised learning*). Ty se obecně používají pro nalezení možnosti odlišovat dvě či více tříd objektů popsaných kvantifikovanými vlastnostmi. Data, která tak do SVM vkládáme, jsou ve formě vektoru číselně reprezentovaných vlastností objektů a požadovaných výstupů.

Nejjednodušším mechanismem SVM je lineární klasifikátor, jehož cílem je nalézt lineární oddělovač, který bude separovat jednotlivé třídy objektů reprezentované v *k*-rozměrném prostoru. Jednotlivým vlastnostem objektů je přitom přidělována váha. Pokud je vlastnosti přiřazena váha 0, SVM model ji bude ignorovat. Označíme-li počet vlastností objektů jako *n* a

[40] Viz Vapnik, V. (1963). *Pattern Recognition Using Generalized Portrait Method*. Automation and Remote Control, 774-780.

počet vlastností, kterým byla přiřazena nulová váha jako *m*, výsledný *k*-rozměrný prostor bude mít $k = n - m$ dimenzí.

Cílem učení SVM modelu je nalézt optimální umístění lineárního oddělovače tak, aby byl minimalizován počet chybných klasifikací a maximalizován okraj mezi oběma třídami. Toto umístění se hledá během tzv. trénování.

Pro optimalizaci separace ve složitějších situacích než v lineárně oddělitelných lze použít tzv. jádrovou transformaci (*kernel transformation*), která převede úlohu do prostoru o $k + 1$ dimenzích a lze tak separovat i třídy, které by nebylo jinak možné správně lineárně oddělit.[41]

V našem pokusu však použijeme lineární jádro, tj. bez jádrové transformace, protože se jedná o nejjednodušší model, a zároveň u něj existuje možnost získat váhy jednotlivých vlastností a tím zjistit, na základě kterých znaků model predikoval klasifikaci do první nebo druhé třídy.

Vstupní data se rozdělují na tzv. trénovací a testovací dataset. Trénovací data slouží k natrénování modelu. Zároveň na nich měříme i úspěšnost predikcí. Tím se totiž ověří, zda se model vůbec něco naučil. Testovací dataset se pak využívá k predikcím výstupu pro modelem dosud neviděná testovací data.[42] Toto je důležité, protože jinak bychom nebyli schopni odhalit, zda došlo k přeučení modelu (*overfit*).

Při učení modelu totiž mohou nastat dvě extrémní a negativní situace – model se naučí data až příliš „nazpaměť“, dojde k jeho přeučení (*overfit*), nebo se naopak nedoučí (*underfit*). Přeučení znamená, že model se příliš do detailu naučí trénovací data, není však schopný dostatečně zobecnit naučené trendy. V konečném důsledku dosahuje velice vysoké úspěšnosti při predikcích pro trénovací data, pro testovací data je však úspěšnost predikce výrazně nižší. Při nedoučení naopak model nedokáže správně modelovat trénovací data, ani zobecnit vypozorovaný trend na dosud neviděná data. Takový model bude mít velice nízkou úspěšnost predikce jak pro trénovací, tak pro testovací data. Cílem učení je dosáhnout toho, aby se model naučil „akorát“ (*good fit*), tedy aby správně modeloval obecné trendy, a nikoli přílišné detaily.[43]

---

[41] *Ibid.*

[42] Srov. Steinwart, I. a Christmann, A. (2008). Support Vector Machines, Springer-Verlag, New York.

[43] Srov. Brownlee, J. (2016). *Overfitting and Underfitting With Machine Learning Algorithms*. Machine Learning Algorithms. Dostupné z: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/ [1. 7. 2021].

1.2.1. Popis pokusu a výsledky

V tomto pokusu použijeme dataset 208 hlavních měst států a 18 943 ostatních měst. Data, která vznikla na základě zdrojů NGIA, US Geological Survey, US Census Bureau nebo NASA, byla získána z webu *simplemaps*.[44] Dle informací uvedených na tomto webu data zahrnují všechna význačná města světa, ať už z hlediska jejich rozlohy, počtu obyvatel, nebo jejich jiné významnosti. Jména měst jsou psána podle pravidel anglického pravopisu.

Jako kvantifikovatelnou vlastnost názvů použijeme Bag-of-Words model (dále BoW) jednotlivých grafémů slov. Jedná se o seznam všech grafémů, které se v názvech z trénovacího datasetu vyskytují, přičemž každý název, ať už jednoslovný či víceslovný, je reprezentován vektorem 0 a 1 podle toho, zda se v jeho názvu daný grafém vyskytuje, nebo nevyskytuje. Všechna písmena jsou přitom konvertována na malá a jsou započítány pouze alfanumerické znaky. Tímto postupem se ztrácí informace o pořadí znaků i jejich frekvenci. Je možné použít i frekvenční model BoW, ve kterém by namísto 1 a 0 byly uvedeny frekvence jednotlivých grafémů. Tuto možnost vyzkoušíme jako další.

Vzhledem k různě dlouhým datasetům je nejprve nutné data předpřipravit. Datasety hlavních a periferních měst načteme a ztokenizujeme. Poté bude seznam periferních měst, tedy delší z obou datasetů, náhodně zamíchán a zkrácen na délku datasetu hlavních měst tak, aby byly datasety vyvážené, tj. tak, aby byl při trénování počet hlavních měst stejný jako počet periferních měst, ale abychom zároveň ponechali co největší prostor náhodě. Následně jednotlivé vzorky hlavních a periferních měst rozdělíme na trénovací a testovací data, a to v poměru 2/3.[45] To znamená, že z 208 hlavních a 208 periferních měst, tedy celkem z 416 měst, připadnou 2/3 (276 jmen) na trénování modelu, zatímco zbylá 1/3 (140 jmen) bude použita na testování úspěšnosti predikce modelu. Ze jmen z trénovacího datasetu vytvoříme slovník (tzv. *globální slovník*), kterým definujeme BoW pro trénovací a testovací data (viz *Tab. 2*). V případě, že jména v testovacím datasetu obsahují znaky, které v trénovacím datasetu nejsou, budou při evaluaci ignorovány.

[44] Viz Pareto Software, SimpleMaps.com (2010–2021). *World Cities Database*. Dostupné z: https://simplemaps.com/data/world-cities [9. 4. 2021].

[45] Rozdělení na trénovací a testovací data viz např. Albrecht, J., Ramachandran, S. a Winkler, C. (2020). *Blueprints for Text Analytics Using Python*. O´Reilly Media.

Ukázka globálního slovníku a vektorů reprezentujících jednotlivá města

Název	**Popis**	**Reprezentace/vektor**
Globální slovník	Slovník obsahující všechna písmena vyskytující se alespoň v jednom ze slov trénovacího datasetu	{'b', 'd', 'v', 'u', 'i', 'y', 'w', 'j', 'h', 'z', 'm', 'r', 'f', 'c', 's', 'a', 'l', 'k', 't', 'p', 'x', 'o', 'e', 'n', 'g'}
Antananarivo	hlavní město Madagaskaru	[0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0]
Madrid	hlavní město Španělska	[0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Svitavy	město v České republice	[0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0]
Hambantota	město na Srí Lance	[1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0]

Tab. 2: Ukázka globálního slovníku a vektorů reprezentujících jednotlivá města.

Model natrénujeme na trénovacím datasetu a otestujeme na testovacím datasetu. Níže popsanými evaluačními metrikami[46] (viz *Tab. 3)* změříme úspěšnost predikce na obou datasetech zvlášť.

Evaluační metriky

Evaluační metrika	**Anglický název**	**Definice metriky**	**Vysvětlení**
Přesnost	*Accuracy*	$\frac{TP + TN}{TP + TN + FP + FN}$	počet správných predikcí ku celkovému počtu predikcí
Úplnost	*Recall*	$\frac{TP}{TP + FN}$	počet správně predikovaných pozitivních výsledků ku celkovému počtu pozitivních výsledků
Preciznost	*Precision*	$\frac{TP}{TP + FP}$	počet správně predikovaných pozitivních výsledků ku celkovému počtu pozitivních predikcí
F1	*F1*	$2 * \frac{\frac{TP}{TP + FN} * \frac{TP}{TP + FP}}{\frac{TP}{TP + FN} + \frac{TP}{TP + FP}}$	harmonický průměr úplnosti a preciznosti

Tab. 3: Evaluační metriky použité k měření úspěšnosti predikce modelu a jejich vysvětlení.

TP (True Positive): Počet správně predikovaných pozitivních výsledků.
TN (True Negative): Počet správně predikovaných negativních výsledků.
FP (False Positive): Počet nesprávně predikovaných pozitivních výsledků.
FN (False Negative): Počet nesprávně predikovaných negativních výsledků.

Kdybychom tento pokus provedli pouze jednou, měl by jeho výsledek velice malou váhu, protože by jej bylo možné označit pouze za dílo náhody kvůli specificky vybraným datům.

[46] Viz Powers, D. M. (2011): *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. Journal of Machine Learning Technologies, 2(1), s. 37–63.

Typicky se vyhodnocení dělá pomocí tzv. *n-folded testů*,[47] ale protože je tato úloha výpočetně relativně snadná a množství dat relativně malé, můžeme si dovolit opakovaně náhodně vzorkovat dataset. Celý pokus proto 1 000× zopakujeme a z výsledků pro každou metriku spočítáme aritmetický průměr, medián a konfidenční interval, do kterého bude spadat 95 % hodnot. Díky tomu, že jsou obě třídy vyrovnané z hlediska počtu, je kritériem nearbitrárnosti překročení 50% hranice úspěšnosti, tj. úspěšnosti dané náhodným modelem. Bude-li konfidenční interval zahrnovat hodnotu 0,5, můžeme považovat jména měst za arbitrární. Následuje tabulkový přehled výsledků pro trénovací (*Tab. 4*) a testovací (*Tab. 5*) dataset.

Přehled výsledků pro trénovací dataset s využitím binárního BoW

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	< 0,66; 0,75 >	0,703	0,703
Preciznost (*Average precision*)	< 0,6; 0,68 >	0,642	0,642
Úplnost (*Recall*)	< 0,66; 0,79 >	0,725	0,725
F1	< 0,66; 0,75 >	0,709	0,71

Tab. 4: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím binárního BoW.

Přehled výsledků pro testovací dataset s využitím binárního BoW

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	< 0,543; 0,686 >	0,62	0,621
Preciznost (*Average precision*)	< 0,523; 0,639 >	0,576	0,574
Úplnost (*Recall*)	< 0,529; 0,757 >	0,648	0,643
F1	< 0,545; 0,708 >	0,63	0,63

Tab. 5: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro testovací dataset s využitím binárního BoW.

Z výsledků uvedených v tabulkách je patrné, že model má statisticky prokazatelně vyšší úspěšnost predikce než náhodný model (tj. konfidenční interval nezahrnuje hodnotu 0,5 a nižší) jak na trénovacích, tak i testovacích datech, což je výsledek v rozporu s naším očekáváním.

[47] Viz např. Hastie, T., Tibshirani R. a Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2. vydání, Springer. Dostupné z: https://web.stanford.edu/~hastie/ElemStatLearn/ [8. 4. 2021].

V následující kapitole provedeme ještě pokus s využitím frekvenčního BoW, poté bude následovat shrnutí výsledků a jejich interpretace v kompletní diskuzi níže.

1.2.2. Binární a frekvenční BoW

Jak jsme již nastínili v popisu tohoto pokusu na začátku kapitoly, kromě binárního BoW lze použít také frekvenční BoW. Jméno města tak bude reprezentováno vektorem nikoli jedniček a nul, nýbrž absolutními frekvencemi výskytu daného grafému v názvu. V tabulce (*Tab. 6*) je uvedeno několik příkladů, jak taková vektorová reprezentace města vypadá.

Ukázka globálního slovníku a vektorů reprezentujících jednotlivá města

Název	**Popis**	**Reprezentace/vektor**
Globální slovník	Slovník obsahující všechna písmena vyskytující se alespoň v jednom ze slov trénovacího datasetu	{'f', 'x', 'o', 'c', 'v', 'k', 'm', 'b', 'p', 'w', 'a', 'r', 'e', 'd', 'l', 'u', 's', 'i', 'y', 'n', 'h', 't', 'z', 'q', 'j', 'g'}
Islamabad	hlavní město Pákistánu	[0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 3, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]
Nay Pyi Taw	hlavní město Myanmaru	[0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 0, 0, 0, 0, 0, 0, 1, 2, 1, 0, 1, 0, 0, 0, 0]
Tadepallegudem	město v Indii	[0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 2, 0, 3, 2, 2, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1]
Hamden	město v USA	[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0]

Tab. 6: Ukázka globálního slovníku a vektorů reprezentujících jednotlivá města. Použit je frekvenční BoW.

Kromě záměny binárního BoW za frekvenční jsme provedli zcela identický pokus se stejnými daty. V následujících tabulkách (*Tab. 7* a *Tab. 8*) vidíme, jakou úspěšnost měly predikce SVM modelu při užití frekvenčního BoW. Z tabulek je patrné, že model využívající reprezentaci dat na základě frekvenčního BoW dosahuje obdobných výsledků jako model s binárním BoW, a to jak u trénovacích, tak u testovacích dat.

Přehled výsledků pro trénovací dataset s využitím frekvenčního BoW

Evaluační metrika	**Konfidenční interval**	**Průměr**	**Medián**
Přesnost (*Accuracy*)	< 0,659; 0,757 >	0,709	0,71
Preciznost (*Average precision*)	< 0,603; 0,691 >	0,646	0,646
Úplnost (*Recall*)	< 0,681; 0,812 >	0,747	0,746
F1	< 0,676; 0,762 >	0,720	0,721

Tab. 7: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím frekvenčního BoW.

Přehled výsledků pro testovací dataset s využitím frekvenčího BoW

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	< 0,55; 0,7 >	0,625	0,621
Preciznost (*Average precision*)	< 0,572; 0,64 >	0,578	0,576
Úplnost (*Recall*)	< 0,543; 0,786 >	0,664	0,671
F1	< 0,554; 0,716 >	0,638	0,64

Tab. 8: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro testovací dataset s využitím frekvenčního BoW.

1.2.3. Diskuze nad výsledky

Ze zjištěných konfidenčních intervalů vyplývá, že natrénujeme-li SVM model pomocí BoW (ať už binárních, nebo frekvenčních) reprezentujících názvy hlavních a ostatních měst, necháme jej predikovat na dosud neviděných datech a budeme zkoumat úspěšnost správné predikce, pak v rámci 95 % případů bude model dosahovat takových výsledků, že hodnoty přesnosti, preciznosti, úplnosti a f1 budou ve všech případech vyšší než 0,5. To znamená, že natrénujeme-li SVM model uvedeným způsobem, úspěšnost jeho predikce bude lepší než náhodný model „hodu férovou mincí“ (s pravděpodobnostmi 0,5 a 0,5 pro obě strany). Z toho můžeme vyvodit, že mezi tím, jaké znaky obsahují názvy měst, a jejich „hlavností“, musí být skrytá jistá, byť nepříliš silná pravidelnost.

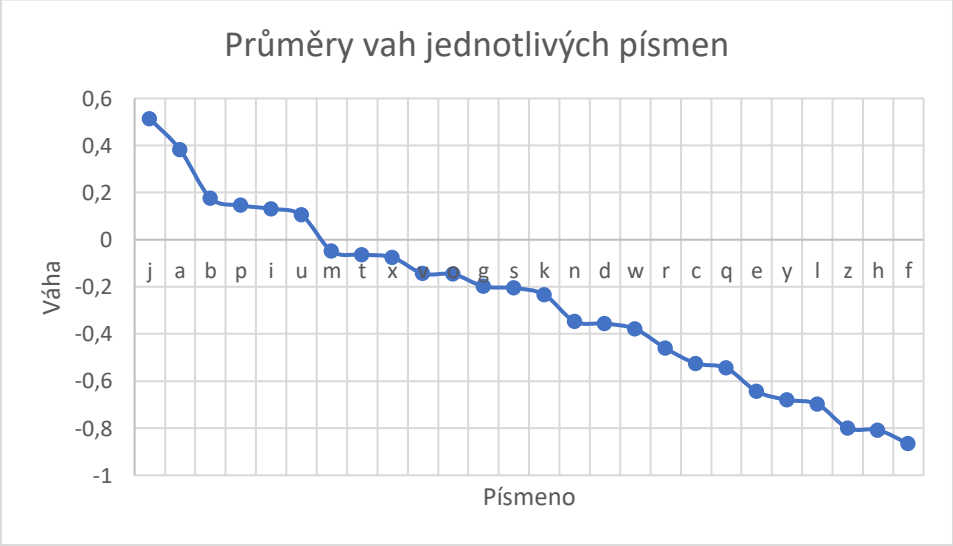
Vzhledem k tomu, že jsme použili lineární jádro, můžeme získat váhy jednotlivých vstupních vlastností a zjistit tak podíl jednotlivých písmen na predikci, což by nám mělo pomoci tuto pravidelnost odhalit. Získáváme následující přehled (*Tab. 9*).

V tabulce (*Tab. 9*) a názorněji také ve spojnicovém grafu (*Graf 3)* vidíme, že seřadíme-li hodnoty od nejvyšší po nejnižší, jsou průměry vah od sebe různě vzdálené. V některých případech jsou průměry vah jednotlivých písmen velice blízké hodnotám, kterých dosahují jejich sousední hodnoty (v grafu vidíme jako velmi mírný spád), zatímco mezi jinými vahami jsou větší rozdíly (v grafu vidíme jako skoky). Např. písmena *m, t, x* dosahují podoných hodnot a mají tak podobnou roli při určování hlavnosti města, zatímco např. mezi písmeny *m* a *u* je z grafu zřejmý výrazný rozdíl, což značí, že hrají odlišnou roli v určení hlavnosti města.

Přehled vah jednotlivých grafémů

Znak	Průměr	Medián	Znak	Průměr	Medián
j	0,514	0,518	**k**	-0,233	-0,220
a	0,382	0,382	**n**	-0,347	-0,352
b	0,177	0,177	**d**	-0,356	-0,353
p	0,146	0,147	**w**	-0,379	-0,366
i	0,131	0,139	**r**	-0,461	-0,472
u	0,105	0,087	**c**	-0,525	-0,514
m	-0,049	-0,049	**q**	-0,545	-0,551
t	-0,063	-0,062	**e**	-0,644	-0,648
x	-0,075	0,000	**y**	-0,680	-0,682
v	-0,144	-0,133	**l**	-0,698	-0,699
o	-0,146	-0,140	**z**	-0,800	-0,848
g	-0,198	-0,210	**h**	-0,808	-0,808
s	-0,205	-0,198	**f**	-0,866	-0,887

Tab. 9: Průměry a mediány váhy jednotlivých písmen. Řazeno podle průměru sestupně.



Graf 3: Vizualizace průměrů vah jednotlivých písmen ve spojnicovém grafu.

Z tabulky (*Tab. 9)* lze vyčíst, na základě kterých písmen se SVM model průměrně rozhodoval pro třídu hlavních měst (kladná hodnota) nebo třídu periferních měst (záporná hodnota), jinými slovy, podle kterých písmen model predikoval „hlavnost“ města.

Přitom platí, že čím vyšší je hodnota směrem od nuly (v kladém i záporném směru), tím více dané písmeno rozhoduje o dané třídě. Z tabulky je tedy zřejmé, že např. písmena *j, a, b, p*

přispívají nejvíce k určení města jako hlavního, zatímco např. písmena *f, h, z* přispívají nejvíce k určení města jako periferního. Dále pozorujeme, že průměry vah jednotlivých písmen nabývají stejných nebo podobných hodnot jako jejich mediány. Z toho lze vyvodit, že data jsou okolo průměru symetricky rozdělena.

Vidíme, že znaků, jejichž průměr je kladný a na jejichž základě se model rozhodoval pro třídu hlavních měst, je pouhých šest, zatímco znaků, podle kterých model průměrně predikoval třídu ostatních měst, je o 20 více. Můžeme uvážit, že periferních měst je na světě mnohonásobně více (v našich datasetech je pak konkrétně periferních měst 91× více než hlavních měst, datasety obsahují 208 hlavních měst a 18 943 periferních měst), proto se v jejich názvech projevuje větší rozmanitost a zároveň i větší pravidelnosti. Kdyby tedy hlavních měst bylo na Zemi více, je možné, že by jejich názvy mohly obsahovat více typických znaků. Vzhledem k malé velikosti datasetu se však tento trend nemůže projevit. Nebo je naopak možné, že kdyby bylo hlavních měst na Zemi více, tak by rozdíly zanikly, protože hlavních měst s charakteristickými rysy (např. s písmeny *j, a, b, p*) by bylo oproti novým hlavních městům méně a při testování úspěšnosti modelu bychom tak získali 50% přesnost (*accuracy*).

Nyní se pokusme interpretovat roli jednotlivých písmen. Proč písmena *j, a, b, p, i, u* rozhodují o tom, že město, jehož název je obsahuje, bude patřit spíše mezi hlavní města, nebo naopak, co způsobuje, že například písmena *f, h, z, l, y, e* v názvu města rozhodují o tom, že město bude spíše periferní.

Při interpretaci je třeba rozlišovat případnou ikonocitu grafickou a fonetickou. O grafické ikonocitě bychom mohli mluvit v případě, že by existovala souvislost mezi grafickou podobou jména města a jeho vlastností, např. kdyby se písmeno *A* vyskytovalo v názvech měst, ve kterých se vyskytuje objekt připomínající tvar tohoto písmene (např. *A* ve jménu *PAŘÍŽ* by souviselo s tvarem Eiffelovy věže, která je jedním z významných symbolů tohoto města). Fonetická ikonicita by pak zahrnovala případy, kdy zvuk vyslovené hlásky bude souviset s vlastností města. Tuto ikonicitu lze zkoumat např. prostřednictvím popisu charakteristik jednotlivých hlásek.

Analýza grafické ikonocity by vyžadovala samostatný rozsáhlý výzkum, proto se v rámci naší práce budeme zabývat pouze interpretací fonetickou. Před samotným pokusem o interpretaci je třeba uvážit, že jak grafika, tak i výslovnost se v průběhu času proměňovaly. Pracujeme-li tedy s dnešní podobou názvů, nemusí tato podoba odpovídat té originální.

Nejdříve se podívejme na vokály. Z tabulky vyplývá, že vokály *a, i, u* hovoří pro třídu hlavních měst, zatímco vokály *e, o, y* pro třídu měst periferních. Písmenem *y* se nyní zabývat nebudeme, protože mu odpovídají různé způsoby fonetické realizace v závislosti na pravidlech daného jazyka i na fonetickém okolí, např. v angličtině se *y* bude často vyslovovat jako [j].

Vokály [a], [i], [u] odpovídají vrcholům vokalického trojúhelníku,[48] jedná se tedy o extrémní případy z hlediska artikulace. Vokál [a] je otevřená střední nezaokrouhlená samohláska, [i] a [u] jsou potom samohlásky zavřené, přičemž [i] je přední a nezaokrouhlená a [u] je je zadní a zaokrouhlená. Naopak samohlásky [o] a [e] jsou samohlásky středové. Zajímavé je, že [a], [i], [u], tedy samohlásky rozhodující o tom, že město je hlavní, odpovídají samohláskám, které existovaly již v sanskrtu,[49] tedy v jednom z nejstarších dochovaných indoevropských jazyků.

Při fonetické realizaci písmen *b* a *p* se jedná o labiály, tedy hlásky, které mají nejlepší recepci z hlediska registrace retní koartikulace. U písmene *j* je obdobný problém jako u *y,* v závislosti na jazyce a fonetickém okolí se bude měnit jeho zvuková realizace (např. [ʒ] ve francouzštině, [x] ve španělštině apod.) a je tedy obtížné stanovit nějaké obecnější pravidlo.

Můžeme uvážit, že důležitá města vznikala nejdříve. Potom by dávalo smysl, že by starší města odpovídala městům s velkou mírou důležitosti (naší „hlavnosti“), a proto (za předpokladu, že se název těchto měst v průběhu času příliš výrazně nezměnil) obsahují ve větší míře samohlásky [a], [i] a [u], pokud tyto jsou starší než [o] a [e].

Platí-li všechny předpoklady uvedené v předchozím odstavci, potom existuje možnost, že hlavní města obsahují ve větší míře samohlásky [a], [i] a [u] z toho důvodu, že hlavní města jsou obecně starší než ta periferní, a proto jejich názvy obsahují hlásky, které se vyskytovaly již v nejstarších podobách jazyků.

1.2.4. Frekvence a pořadí znaků

Zůstaňme u písmen, která jsme odvodili výše a která by mohla souviset s „hlavností“ města a podívejme se na vlastnosti, které byly dosud částečně nebo zcela opomenuty. Při vytváření BoW na znacích jsme ztratili například informaci o pořadí jednotlivých znaků.

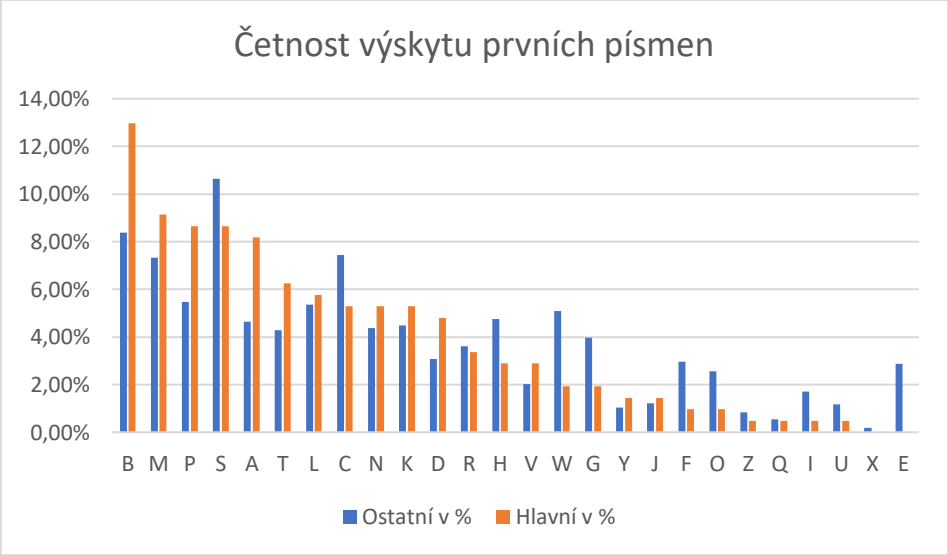
---

[48] Viz např. IPA (International Phonetic Association, 2020). *The international phonetic alphabet*. Dostupné z: https://www.internationalphoneticassociation.org/IPAcharts/IPA_chart_orig/pdfs/IPA_Kiel_2020_full.pdf [4.7.2021].

[49] Srov. Whitney, W. D. (1879). *A Sanskrit grammar: including both the classical language, and the older dialects (of Veda and Brahmana)*. Lipsko: Breitkopf a Härtel. Dostupné z: https://archive.org/details/sanskritgrammari00whituoft/page/2/mode/2up?view=theater [6.7.2021].

Nejjednodušším způsobem, jak pracovat s pořadím znaků, je podívat se na první písmena, konkrétně na jejich četnosti ve jménech hlavních a ostatních měst, a ta následně porovnávat. Vzhledem k velkému rozdílu mezi velikostmi datasetů by nebylo vhodné porovnávat absolutní četnosti, zatímco vybrat náhodný vzorek by bylo problematické z důvodu reprezentativnosti. Můžeme však například, jestliže zapíšeme frekvence jednotlivých prvních písmen hlavních a periferních měst do vektorů, změřit kosinovou podobnost obou distribucí. Její hodnota 0,926 naznačuje, že způsob užití jednotlivých písmen na začátku jmen bude podobný, nikoli však totožný.

Dále se nabízí například možnost vyjádřit v procentech zastoupení jednotlivých znaků ku celkovému počtu prvních písmen. Tím vyřešíme problém nepoměru mezi velikostmi datasetů a zároveň budeme měřit data na mnohonásobně větším počtu vzorků, než kdybychom náhodně vybrali data z jednoho konkrétního pokusu. Získáváme tak následující tabulku (*Tab. 10*) relativní četnosti výskytu, která je pro přehlednost znázorněna v grafu (*Graf 4*).



Graf 4: Porovnání četnosti výskytu různých písmen na začátku názvu hlavních a ostatních měst.

Přehled četností prvních písmen

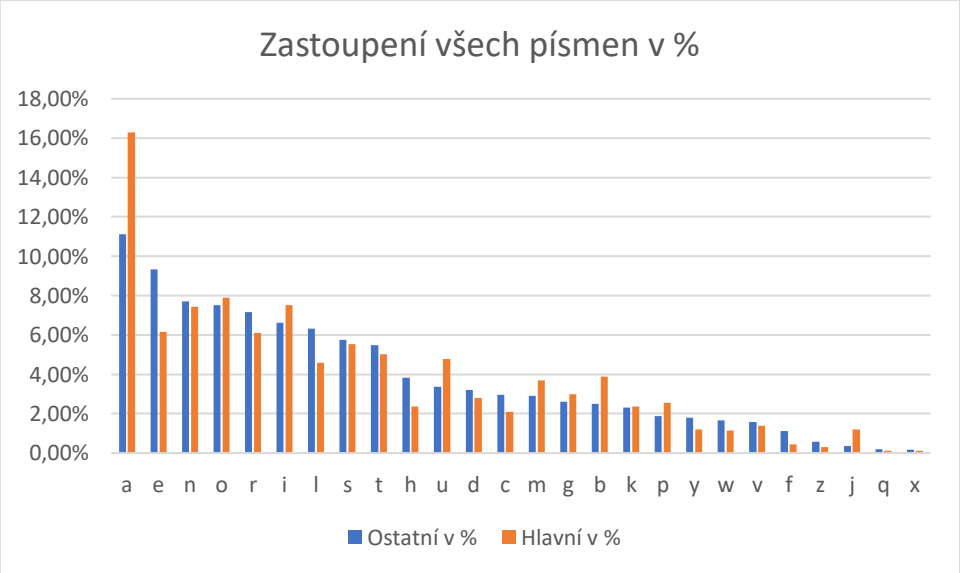
První písmeno	Absolutní četnost – ostatní města	Absolutní četnost – hlavní města	Relativní četnost – ostatní města	Relativní četnost – hlavní města
B	1588	27	8,39 %	12,98 %
M	1387	19	7,33 %	9,13 %
P	1035	18	5,47 %	8,65 %
S	2016	18	10,65 %	8,65 %
A	880	17	4,65 %	8,17 %
T	813	13	4,29 %	6,25 %
L	1014	12	5,36 %	5,77 %
C	1409	11	7,44 %	5,29 %
N	830	11	4,38 %	5,29 %
K	849	11	4,48 %	5,29 %
D	584	10	3,08 %	4,81 %
R	684	7	3,61 %	3,37 %
H	899	6	4,75 %	2,88 %
V	381	6	2,01 %	2,88 %
W	964	4	5,09 %	1,92 %
G	750	4	3,96 %	1,92 %
Y	198	3	1,05 %	1,44 %
J	228	3	1,20 %	1,44 %
F	561	2	2,96 %	0,96 %
O	483	2	2,55 %	0,96 %
Z	158	1	0,83 %	0,48 %
Q	103	1	0,54 %	0,48 %
I	323	1	1,71 %	0,48 %
U	221	1	1,17 %	0,48 %
X	34	0	0,18 %	0,00 %
E	543	0	2,87 %	0,00 %

Tab. 10: Četnosti výskytu různých písmen na začátku názvu hlavních a ostatních měst.

V tabulce (*Tab. 10*) a grafu (*Graf 4*) vidíme, že u některých písmen je distribuce podobná, zatímco u jiných písmen jsou patrné rozdíly. Z grafu (*Graf 4*) je zřetelné např. to, že na prvním místě se v názvech hlavních měst mnohem častěji objevují písmena *B* nebo *A*, než tomu je u ostatních měst. Pro zjištění, zda je za těmito distribucemi skryto nějaké pravidlo, je třeba získat

informaci o tom, jak je pravděpodobné vidět takovýto nebo ještě větší rozdíl mezi hlavními a periferními městy pouhou náhodou. To zjistíme pomocí chí-kvadrátu se simulovanou p-hodnotou pomocí metody Monte Carlo. Chí-kvadrát nabývá hodnoty 38,42 a p-hodnota = 0,08838, což znamená, že pozorovaný nebo extrémnější rozdíl mezi skupinou hlavních a periferních měst by se stal s pravděpodobností 0,08838 (tedy i v rámci 95 % případů) i pouze vlivem náhody. Nelze tudíž zamítnout možnost, že skupiny hlavních a periferních měst jsou z hlediska distribuce počátečních písmen identické. V ohledu využití prvního písmene ve jméně se tedy jména hlavních a periferních měst chovají stejně.

Nyní se dostáváme k frekvenci písmen, kterou jsme částečně využili již při vytváření frekvenčního BoW. Dále porovnáváme distribuce všech písmen v názvech. Z podobných důvodů jako u distribuce prvních písmen budeme zkoumat relativní četnosti. Získáváme následující absolutní a relativní četnosti (*Tab. 11*). Data pro lepší přehlednost opět graficky znázorníme (*Graf 5*).



Graf 5: Porovnání četnosti výskytu různých písmen v názvech hlavních a ostatních měst.

Kromě několika znaků se četnosti jeví dosti podobné, což potvrzuje hodnota kosinové podobnosti 0,962. Ověřit podobnost obou četností chí-kvadrátem v tomto případě není možné, jelikož jsou porušeny předpoklady pro použití tohoto testu. Jednotlivá pozorování totiž nejsou nezávislá: výskyt písmen v rámci slova je ovlivněn tím, jaká písmena jim předcházela.

Přehled distribuce jednotlivých písmen

Písmeno	Absolutní četnost – ostatní města	Absolutní četnost – hlavní města	Relativní četnost – ostatní města	Relativní četnost – hlavní města
A	18775	256	11,11 %	16,28 %
E	15771	97	9,33 %	6,17 %
N	13023	117	7,71 %	7,44 %
O	12716	124	7,53 %	7,89 %
R	12095	96	7,16 %	6,11 %
I	11177	118	6,61 %	7,51 %
L	10687	72	6,32 %	4,58 %
S	9715	87	5,75 %	5,53 %
T	9253	79	5,48 %	5,03 %
H	6452	37	3,82 %	2,35 %
U	5669	75	3,35 %	4,77 %
D	5404	44	3,20 %	2,80 %
C	4989	33	2,95 %	2,10 %
M	4922	58	2,91 %	3,69 %
G	4426	47	2,62 %	2,99 %
B	4225	61	2,50 %	3,88 %
K	3919	37	2,32 %	2,35 %
P	3182	40	1,88 %	2,54 %
Y	3014	19	1,78 %	1,21 %
W	2808	18	1,66 %	1,15 %
V	2691	22	1,59 %	1,40 %
F	1887	7	1,12 %	0,45 %
Z	973	5	0,58 %	0,32 %
J	604	19	0,36 %	1,21 %
Q	323	2	0,19 %	0,13 %
X	273	2	0,16 %	0,13 %

Tab. 11: Četnosti výskytu různých písmen v názvech hlavních a ostatních měst.

Přesto nás může zaujmout např. velký rozdíl mezi výskytem prvních dvou písmen, konkrétně *a* a *e*, který je z grafu vizuálně patrný, nicméně nemůžeme vyloučit, že tento rozdíl je způsobený vlivem pouhé náhody. To, co musíme opět ověřit, je vliv vzorkování či náhody, proto opět natrénujeme SVM model s lineárním jádrem na náhodných vzorcích hlavních měst a

periferních měst. Datasety opět zkrátíme tak, aby měly stejnou velikost a rozdělíme je na trénovací a testovací data. Při každém z 1 000 opakování je dataset periferních měst náhodně obměňován, stejně jako rozdělení dat na trénovací a testovací. Každé jméno následně reprezentujeme jako vektor o dvou prvcích: počet písmen *a* a *e*. Z principu nepředpokládáme, že by model dokázal úspěšně predikovat správnou třídu a očekáváme úspěšnost shodnou s náhodným modelem, tj. s průměrem 0,5. Jakmile bude konfidenční interval získaný 1 000 opakováními zahrnovat tuto hodnotu, nemůžeme zamítnout možnost, že mezi hlavními a periferními městy z hlediska počtu dvou zkoumaných vokálů neexistuje žádný rozdíl daný něčím jiným než jen náhodou.

Model natrénujeme a měříme úspěšnost jeho predikcí na trénovacím i testovacím datasetu. Vše opakujeme 1 000×. Následuje tabulkový přehled výsledků (*Tab. 12* a *13*).

Přehled výsledků pro trénovací dataset s využitím frekvencemí písmen* a *a* e *jako vlastností

Evaluační metrika	**Konfidenční interval**	**Průměr**	**Medián**
Přesnost (*Accuracy*)	< 0,54; 0,641 >	0,587	0,587
Preciznost (*Average precision*)	< 0,521; 0,589 >	0,55	0,548
Úplnost (*Recall*)	< 0,572; 0,928 >	0,748	0,729
F1	< 0,582; 0,69 >	0,641	0,643

Tab. 12: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím lineárního kernelu a písmen a *a* e *jako vstupních vlastností.*

Přehled výsledků pro testovací dataset s využitím frekvencemí písmen* a *a* e *jako vlastností

Evaluační metrika	**Konfidenční interval**	**Průměr**	**Medián**
Přesnost (*Accuracy*)	< 0,514; 0,658 >	0,581	0,579
Preciznost (*Average precision*)	< 0,507; 0,6 >	0,547	0,544
Úplnost (*Recall*)	< 0,529; 0,943 >	0,743	0,729
F1	< 0,541; 0,705 >	0,635	0,642

Tab. 13: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro testovací dataset s využitím lineárního kernelu a písmen a *a* e *jako vstupních vlastností.*

Získáváme překvapivý výsledek, a to, že model v rámci 95 % případů určí s úspěšností vyšší než 50 %, zda se jedná o hlavní nebo periferní město pouze na základě počtu výskytů písmen *a* a *e*. Úspěšnost platí jak pro trénovací, tak i pro testovací dataset, to znamená, že model dosahuje vyšší úspěšnosti než náhodný model i pro jména měst, která ještě neviděl. Výsledky pro trénovací i testovací dataset mají obdobně relativně nízké hodnoty, ale (z hlediska průměrů či mediánu z *Tab. 12* a *13*) stále vyšší než 0,5, a tedy zřejmě existuje ve využití obou písmen jistý latentní vztah k hlavnosti města.

Díky využití lineárního jádra můžeme opět získat koeficienty obou dvou písmen. Z nich spočítáme průměr a medián, které jsou uvedeny v následující tabulce (*Tab. 14*).

***Průměr a medián predikční váhy písmen* a *a* e**

Písmeno	průměr	Medián
A	0,065	$5{,}051\times10^{-15}$
E	–0,97	–1

Tab 14: Průměr a medián váhy písmen a *a* e.

Zjišťujeme, že na základě písmene *a* model v průměrném případě predikoval hlavnost města, zatímco podle písmene *e* v průměrném případě predikoval třídu ostatních měst, což je v souladu se znázorněním četností ve výše uvedeném grafu (*Graf 5*). Zároveň lze z predikční úspěšnosti modelu usoudit, že existuje vztah mezi „hlavností“ a písmenem *a* a „nehlavností“ a písmenem *e*, což je v rozporu s očekáváním jazykové arbitrárnosti. V návaznosti na úvodní diskutovanou problematiku fonémů lze konstatovat protiklad mezi otevřenou hláskou [a], která souvisí spíše s hlavností města, a středovou hláskou [e], která souvisí spíše s periferností města.

1.2.5. Limity uvedeného pokusu

Stejně jako v předchozím pokusu, ve kterém jsme zjišťovali korelace mezi různými kvantifikovatelnými vlastnostmi měst a jejich názvů, i zde má pokus zabývající se užitím písmen uvnitř názvů měst své limity a místa, ve kterých mohlo snadno dojít k chybám, které mohly zapříčinit, že výsledky by mohly být zkreslené, nebo odpovídat jen konkrétním řešeným problémům, nikoli však obecným trendům.

Prvním úskalím by mohl být geografický fakt, že v některých státech hlavní město neodpovídá nejdůležitějšímu městu, takže město, které se ocitlo ve třídě ostatních měst, může mít ve skutečnosti větší míru „hlavnosti“ než oficiální hlavní město. Otázkou zůstává, jakým

způsobem by se takový problém mohl projevit. Pokud v našich datasetech existují města s jinou mírou „hlavnosti“, než tou, která je typická pro jejich skupinu (1/0, hlavní nebo periferní), mohly být výsledky tímto faktem ovlivněny několika způsoby v závislosti na dalších faktorech.

První možností je, že se tímto míra jistoty v určení, zda jde o hlavní, nebo periferní město, snížila. Ta by nejspíš nastala v případě, že opravdu existuje souvislost mezi mírou „hlavnosti“ města a jeho názvem, jak naznačují výsledky pokusů v této kapitole. Čím lépe je totiž třída objektů reprezentována, tím snazší je zařadit do ní objekt, který se bude ostatním podobat.

Druhou možností potom je, že jsou naše výsledky v důsledku chyby lepší, než by měly ve skutečnosti být. To by ale vzhledem k statisticky významné úspěšnosti modelu, která je vyšší než 50 %, znamenalo, že se model orientoval podle nějaké jiné vlastnosti, kterou by měla jména měst společná, ať už by byla způsobená na základě intralingvistických nebo extralingvistických faktorů nebo náhody.

Třetí možností je, že je tento nedostatek dostatečně vyvážen velikostí datasetu periferních měst. Vzhledem k jeho rozsáhlosti by v něm měla převážit města, u nichž problém s reálnou „hlavností“ nepozorujeme, a výsledky by tedy nebyly nijak zásadně ovlivněny.

Dalším problémem, který může zapříčinit, že by výsledky nebyly reprezentativní, může být příliš rozdílná velikost datasetů. Při trénování jsme zkracovali náhodně vybraný vzorek ostatních měst na délku datasetu měst hlavních a při porovnávání četností prvních a všech písmen jsme zvolili relativní četnosti, což tento problém pro účely pokusů vyřešilo. Avšak je možné, že se v datasetu hlavních měst nemohly projevit všechny tendence, které by se mohly objevit, kdyby jich bylo více, například větší rozmanitost písmen. Kdyby tedy bylo na Zemi více hlavních měst, je možné, že by se rozdíly setřely.

Dále pracujeme pouze s psanou formou názvů. Je tedy možné, že některá písmena mají zcela jinou fonetickou podobu. Taktéž pracujeme s názvy v anglické formě pravopisu, takže názvy v některých případech nemusí vůbec odpovídat tomu, jak jsou psány místními obyvateli, a jejich přepisem do latinky, případně překladem do angličtiny, může dojít ke ztrátě informací. Zároveň je tedy možné, že pravidelnost v označení jmen hlavních a periferních měst může být vytvořena pouze nějakými pravidly angličtiny.

1.2.6. Shrnutí druhé kapitoly a diskuze: Jsou výše zmíněná zjištění opravdu projevy nearbitrárnosti?

V předchozí kapitole jsme odhalili řadu výsledků, které byly v rozporu s očekáváním na základě de Saussurovy definice arbitrárnosti znaku. Je však třeba alespoň nastínit různé jiné možnosti, které mohly tato zjištění zapříčinit nebo které je třeba vzít v úvahu, protože s problematikou souvisejí.

V této kapitole jsme zjistili, že natrénovaný SVM model dokáže se statisticky prokazatelnou úspěšností vyšší než 50 % určit, zda se podle zadaného názvu města jedná o hlavní nebo periferní město, a to jak na základě binárního BoW nebo frekvenčního BoW, tak i na základě pouhého počtu výskytu písmen *a* a *e*. Pomocí koeficientů modelu bylo také zjištěno, na základě kterých písmen se model v průměrném případě rozhodoval pro tu či onu třídu. Odhalili jsme například, že hlavní města častěji obsahují samohlásky, které tvoří vrcholy tzv. vokalického trojúhelníku ([a], [i], [u]) a že se jedná o samohlásky, které existovaly už v jednom z nejstarších indoevropských jazyků, sanskrtu.

Položme si nyní otázku, zda tato zjištění opravdu souvisí s arbitrárností. Zaprvé je třeba uvážit, že názvy měst vznikly ve dvou pomyslných krocích, nikoli v jednom, jak je tomu u běžné slovní zásoby daného jazyka. Města jsou totiž obvykle pojmenovávána ve vztahu k okolí či jiné skutečnosti – jmenujme např. Vysoké Mýto (ať už je název odvozený od celního poplatku, nebo od mýtiny[50]), Havířov, Háj ve Slezsku nebo Budišov nad Budišovkou. Ani jedno z těchto slov není vybrané náhodně, nýbrž odkazuje na historickou skutečnost (Vysoké Mýto), na charakter města (Havířov), na jeho umístění v rámci republiky (Háj ve Slezsku), nebo na vodní tok, v jehož blízkosti se obec nachází (Budišov nad Budišovkou). Jedná se tedy o slova, která jsou v rámci jazykové struktury nenáhodná a můžeme mluvit o závislosti jména města jak ve vztahu ke skutečnosti, podle které je pojmenováno (např. Háj ve Slezsku se nachází ve Slezsku), tak i k ostatním slovům daného jazyka (např. Havířov souvisí se slovem *havíř*).

Dále je třeba uvážit, že názvy měst (i toponym obecně) se v čase mění, a to z rozmanitých důvodů. Pozorujeme například, jak se jména proměňují v souladu se Zipfovým principem jazykové ekonomie, tedy že často používaná slova mají tendenci být co nejkratší. Tak se postupem času z *The Czech Republic* stává *Czechia*, obyvatelé *Českých Budějovic* své město

[50] Srov. Profous, A. (1951). *Místní jména v Čechách: jejich vznik, pův. význam a změny*. 3. díl: M-Ř. 1. vyd. Praha: Nakladatelství Československé akademie věd, s.168.

nenazvou jinak než *Budějice* a z *Lutecia des Parisii* se v průběhu času stala jen *Paříž (Paris)*. Spolu se zkracováním názvu se může vytrácet arbitrárnost/ikonicita v názvu, která v něm původně mohla být. Záleží tedy, jestli se budeme dívat na slova z čistě synchronního pohledu, nebo zda budeme sledovat celý historický vývoj slova, pokud budeme chtít rozhodnout, zda názvy jsou, nebo nejsou arbitrární.

Kromě příčin intralingvistických byla jména v průběhu historie častokrát měněna na základě různých vnějších vlivů, ať už se jednalo o důvody politické, historické a další, např. přejmenování dobytých měst Alexandrem Velikým na Alexandrie, přejmenování Leningradu/Petrohradu, nebo, zůstaneme-li prostorově i časově v bližším okolí, změny názvu Gottwaldova/Zlína. Důležité je, že se tímto způsobem proměňovaly především názvy větších měst, protože pojmenování malé vesnice nemělo z politického hlediska takový význam.

Speciálním případem je také přejmenovávání měst v dobách kolonialismu, které se týkalo především afrických a amerických, ale i jiných států[51]. Toto přejmenovávání mělo politický rozměr a státy se následně snažily vrátit k názvu původnímu, nebo nově vytvořit takový, který by vycházel z vlastního jazyka. Název hlavního města je totiž důležitý pro vyjádření národní, jazykové i kulturní identity obyvatel a jejich nezávislosti.[52] I přesto zůstávají města, jejichž název odráží původní koloniální jazyk. Uvážíme-li, že moderních koloniálních mocností nebylo příliš mnoho – standardně se mezi ně řadí Spojené království, Francie, Španělsko, Portugalsko, Nizozemí, Belgie, Německo, Itálie, Dánsko, Švédsko, Rusko, Turecko, Japonsko a USA.[53] Vidíme, že naprostá většina těchto mocností hovoří některým z románských nebo germánských jazyků. Z tohoto pohledu je nutné zvážit, že je možné, že náš model odhadoval „hlavnost" města pouze na základě toho, zda se jméno města blíží pojmenování v některém z těchto jazyků. Tuto možnost by bylo třeba řešit podrobnou analýzou, při které by bylo zjištěno, kolik jmen hlavních měst vychází z některého z germánských či románských jazyků, jaké jsou jejich typické rysy apod. Tato problematika už však přesahuje rámec této diplomové práce a vyžadovala by samostatný výzkum.

[51] Např. v Asii – srov. proměny názvu města Jakarta, viz např. *New World Encyclopedia* (2018). Heslo *Jakarta*. Dostupné z: https://www.newworldencyclopedia.org/entry/Jakarta [31. 7. 2021].

[52] Srov. *Organiser* (2018). *How Renaming the Cities is a Part of Decolonisation and Historical Justice*. Dostupné z: https://www.organiser.org/Encyc/2018/10/20/How-Renaming-the-Cities-is-a-Part-of-Decolonisation-and-Historical-Justice.html [10. 7. 2021].

[53] Srov. Ferro, M. *Dějiny kolonizací. Od dobývání až po nezávislost 13. – 20. století*. NLN, Praha 2007.

Nastiňme ještě jednu myšlenku, která by mohla vysvětlovat podobnost ve využití konkrétních grafémů v názvech hlavních měst mimo Evropu a měst pojmenovaných koloniálními velmocemi. První města se zakládala tam, kde to bylo nejvýhodnější z hlediska dostupnosti vody, přístupnosti apod. V průběhu času se města rozšiřovala a pokud tomu nezabránila nějaká překážka, postupně získávala na významu i velikosti. Přirozeným vývojem se z některých z těchto měst mohla stát města hlavní. (Samozřejmě zde platí výjimky, jako třeba u hlavního města Brazílie, kde byla Brasília uměle vytvořena až v roce 1960.[54] Faktem však u této výjimky je, že na základě statistik je důležitější Rio de Janeiro, ať už z hlediska počtu obyvatel, hustoty zalidnění, výše vyprodukovaného HDP nebo počtu univerzit.[55]) Pokud platí, že hlavní města budou obecně spíše starší než menší města, pak se díky stáří pojmenovávala asi i jmény vycházejícími z jednoho či několika málo jazyků, z čehož plynou podobnosti, zatímco nehlavní města budou spíše mladší, a tedy se jejich jména utvářela pod vlivem i jiných jazyků.

Pro prozkoumání nastíněných myšlenek se můžeme podívat na to, která města model na základě frekvenčního BoW všech písmen klasifikoval správně jako TP (*True Positive*, tj. správně predikovaná hlavní města), TN (*True Negative*, tj. správně predikovaná periferní města), FP (*False Positive*, tj. periferní města predikovaná jako hlavní) a FN (*False Negative*, tj. hlavní města, která byla predikovaná jako periferní). Z testovacího datasetu vybereme náhodný vzorek tak, abychom jej mohli zanalyzovat. Získáváme následující tabulku (*Tab. 15).*

Vzorek měst a třída jejich zařazení

True Positive (TP)	***False Positive (FP)***	***True Negative (TN)***	***False Negative (FN)***
Abidjan	Urraween	Kostrzyn nad Odra	Wellington
San Jose	Nairn	Kostelec nad Orlici	Kingstown
Paris	Mohlin	Park City	Amsterdam
Islamabad	Canton	Walthamstow	Georgetown
Amman	Anju	Clydebank	Copenhagen
Asuncion	Bella Vista	Aspen Hill	Rangoon
Quito	Baures	Macherla	Bridgetown
Gaborone	Tampa	Gardnerville	Riyadh
Addis Ababa	Mount Julie	Fairlawn	Basseterre
Paramaribo	Nasrabad	Richmond	Porto-Novo
Funafuti	Maitland	Round Lake	Tegucigalpa

[54] Srov. *Portal do Governo Brasileiro. Brasília, História & Fotos* (2014). Dostupné z: https://cidades.ibge.gov.br/brasil/df/brasilia/historico [10. 7. 2021].

[55] Srov. *versus.com* (2021). *Brasilia vs Rio de Janeiro: What is the difference?* Dostupné z: https://versus.com/en/brasilia-vs-rio-de-janeiro [13. 5. 2021]

True Positive (TP)	*False Positive (FP)*	*True Negative (TN)*	*False Negative (FN)*
Dushanbe	Kasli	North Riverside	Pyongyang
Oslo	Faridabad	Shanor-Northvue	Yerevan
Abu Dhabi	Pattoki	Tracy	Windhoek
Dodoma	Holic	Alstonville	Montevideo
Kabul	Cary	East Highland Park	Castries
Saint Johns'	Pocking	Independence	Port of Spain
Baghdad	Odate	Jefferson Valley-York-town	San Salvador
Mogadishu	Dublin[56]	Neukirchen	Conakry
Madrid	Munuf	Issy-les-Moulineaux	Libreville
Cairo	Burslem	Enterprise	La Paz
Buenos Aires	Vallauris	Severn	Lilongwe
Santiago	Mahabad	Fort Dix	Mexico City
Beirut	Banos	Carpentersville	Tehran
Juba	Sarubetsu	Fort Valley	Podgorica
Vilnius	Waitakere	Bad Schmiedeberg	Stockholm
Cotonou	Kalyani	Watervliet	
Pretoria	Suwa	Saint-Charles-Borromee	
Berlin	Hamden	Veintiocho de Noviem-bre	
Accra	Tasnad	Garner	
Majuro	Bilton	Lototla	
Bangui	Sankt Augustin	Ste. Anne	
Malabo	Lubuklinggau	Freilassing	
Ouagadougou	Piscataway	Settat	
Sofia	Zhangjiakou	New Orleans	
Nur-Sultan			
Moroni			
Nairobi			
Luxembourg			
Cape Town			
Riga			
Apia			
Manama			
Yaounde			

Tab. 15: Vzorek měst z testovacího datasetu a třída jejich zařazení.

[56] Nejedná se o hlavní město Irska, nýbrž o město Dublin ve státě Georgie, Spojené státy americké.

Nyní se podívejme, zda najdeme nějaké společné vlastnosti mezi *TP* a *FP*, tj. na základě kterých rysů měl model tendenci predikovat město do třídy hlavních měst. Vidíme, že se zde často objevují španělsky znějící jména (např. *San José, Asuncion, Bella Vista*). Ani to však není jisté, protože např. město *Montevideo* bylo nesprávně zařazeno mezi města periferní.

Dále se zde několikrát opakují názvy se sufixem *-abad,* konkrétně jednou v TP (*Islamabad*) a třikrát ve FP (*Nasrabad, Faridabad, Mahabad*). Tento sufix pochází ze střední perštiny, znamená „postavené a obydlené místo" a jedná se o rozšířený sufix v oblastech střední, jižní a západní Asie.[57]

Pozorujeme také obecně v TP a FP častý výskyt *B* a *P* na začátku jména (konkrétně osm výskytů v TP a osm výskytů ve FP oproti dvěma výskytům v TN a šesti výskytům ve FN) , což odpovídá námi zjištěným četnostem jednotlivých písmen v názvech hlavních a periferních měst (viz *Tab. 11* výše). Zajímavé je také správné zařazení jmen hlavních měst pobaltských států do TP (*Vilnius, Riga*). Tato dvě města mají společné to, že obsahují samohlásky, které tvoří vrcholy diskutovaného vokalického trojúhelníku ([a], [i], [u]).

Budeme-li hledat společné rysy pro města správně označená za periferní (*TN*) a chybně označená za periferní (*FN*), pozorujeme, že se často jedná o města s anglickým názvem (např. *Wellington, Kingstown, Park City, Bridgetown, Georgetown*). Najdeme zde také města s dominantní převahou písmen *e* a *o* (např. *New Orleans, Stockholm, Montevideo*). I z tohoto trendu však najdeme výjimky, např. *Amsterdam* chybně zařazený mezi periferní města. Častý výskyt písmene *y* spíše než jména měst odráží anglický pravopis (srov. *Pyongyang, Yerevan*).

Z geografického hlediska (viz *Tab. 16*) model správně určil jako hlavní (*TP*) města v oblasti Afriky (17), Jižní Ameriky (5), Evropy (8), Arabského poloostrova (5), Střední Asie (4), Austrálie a Tichomoří (3) a Severní Ameriky (2). Nesprávně jako hlavní (*FP*) města pak určil města v Evropě (11), USA (7), Střední Asii (6), Východní a Jihovýchodní Asii (6), Austrálii a Tichomoří (3) a v Jižní Americe (1).

Naopak správně jako periferní (*TN*) určil města v rámci USA (19) a Evropy (10), dále pak ve Střední a Jižní Americe (2), Africe (1), Střední Asii (1) a Austrálii (1). Nesprávně jako periferní (*FN*) potom města v oblasti Střední a Jižní Ameriky (10), v Africe (5), Evropě (5), na Arabském

[57] Viz Bulliet, R. (2011). *Cotton, climate, and camels in early Islamic Iran: a moment in world history.* Columbia University Press.

poloostrově a v Západní Asii (3), ve Východní a Jihovýchodní Asii (2) a Austrálii a Tichomoří (1).

Počet prvků jednotlivých tříd pro kontinenty a oblasti

Oblast	Počet TP	Počet FP	Počet TN	Počet FN
Afrika	17	0	1	5
Střední a Jižní Amerika	5	1	2	10
Severní Amerika	2	7	19	0
Austrálie a Tichomoří	3	3	1	1
Evropa	8	11	10	5
Střední Asie	4	6	1	0
Arabský poloostrov a Západní Asie	5	0	0	3
Východní a Jihovýchodní Asie	0	6	0	2

Tab. 16: Třídy zařazení podle modelu pro jednotlivé kontinenty a oblasti.

Nyní se vraťme k myšlence, že mnoho jmen hlavních měst by mohlo být díky kolonizaci v některém z románských (především španělština, francouzština, portugalština) nebo germánských (především angličtina) jazyků. Ze zařazení měst do třídy hlavních nebo periferních měst nezískáváme jednoznačnou odpověď. Model sice zařadil do třídy hlavních měst ($TP + FP$) města ze španělsky mluvících států, zejména z Jižní Ameriky (6), zároveň však 12 měst Střední a Jižní Ameriky zařadil nesprávně jako periferní (FN). V případě angličtiny model správně predikoval jako hlavní města v anglofonních oblastech (Afrika, Evropa, USA, Indie, Austrálie). Toto geografické rozdělení nám však v žádném případě nedává vyčerpávající vysvětlení. Anglické názvy měst se objevují i jinde ve světě než ve zmíněných oblastech, což je dáno historicko-politickým kontextem (především výše zmíněnou kolonizací).

Tak například hlavní město Guyany *Georgetown* (které bylo nesprávně zařazeno mezi periferní) se sice nachází v Jižní Americe, kde je dominantním jazykem španělština a portugalština, avšak toto město bylo založeno Brity v roce 1781 a pojmenováno po králi Jiřím III. (George III).[58] Budeme-li brát v úvahu geografické oblasti, ve kterých se nachází města, která model označil za hlavní/periferní, budou takováto města představovat problém

[58] Srov. *The Editors of Encyclopaedia Britannica* (2015). *Georgetown, national capital, Guyana*. Dostupné z: https://www.britannica.com/place/Georgetown-Guyana [12. 7. 2021].

v interpretaci na základě zobecnění geografických území a dominantího jazyka, kterým se v dané oblasti hovoří.

Můžeme také srovnat nejčastější písmena ve španělštině a angličtině s písmeny, které model považuje za ukazatele hlavnosti města (*j, a, b, p, i, u*). Mezi nejfrekventovanější písmena španělštiny patří *e, a, o*,[59] zatímco v angličtině to jsou *e, t, a*.[60] Ani v jednom případě se kromě písmene *a* neshodují frekventované grafémy daných jazyků s písmeny, na základě kterých model predikuje „hlavnost" města. Myšlenku, že názvy hlavních měst by byly tvořeny z velké části názvy v románských či germánských jazycích nelze tedy nijak přesvědčivě podložit (navíc s přihlédnutím k tomu, že velká část anglických jmen spadla do kategorie *FN*). Naopak nelze zavrhnout myšlenku, že názvy hlavních měst do značné míry ovlivnila kolonizace.

Pro myšlenku, že jména hlavních měst jsou starší, protože starší jsou i samotná města, hovoří například to, že model ve vybraném vzorku zařadil mezi TP a FP celkem 4 města obsahující sufix *-abad* pocházející ze střední perštiny. Tuto myšlenku nelze zavrhnout, pro její další zkoumání by však bylo třeba historiografických faktů týkajících se daných měst tak, aby bylo možné např. zjistit stáří hlavních měst ve srovnání se stářím jiných měst ve stejné územní oblasti, a také prozkoumat, jakými jazyky se v době jejich vzniku na daném území hovořilo a typické rysy těchto jazyků včetně využití jednotlivých grafémů/fonému a zjistit tak, jestli u nich nalezneme nějaké rozdíly.

Závěrem tohoto zamyšlení se vraťme k původní otázce, zda tato zjištění nějak souvisí s arbitrárností jmen měst, a zda se lze vůbec dopátrat odpovědi na otázku, jestli jsou jména měst arbitrární.

Je možné, že jména měst původně arbitrární byla, postupem času však převážily trendy způsobené jazykovou ekonomizací. Stejně tak je možné, že jména měst byla od počátku ikonická, avšak v průběhu času se ze stejných důvodů tato transparentnost ztratila.

Jisté je, že tuto otázku nelze vyřešit pouze statistikou nebo hledáním společných rysů hlavních měst, nýbrž je zapotřebí kombinace synchronního i diachronního přístupu, v rámci kterých

[59] Srov. *Stefan Trost Media* (2007–2021). *Alphabet and Character Frequency: Spanish (Español)*. Dostupné z: https://www.sttmedia.com/characterfrequency-spanish [12. 7. 2021].

[60] Srov. *Math Explorers' Club* (2019). *English Letter Frequency (based on a sample of 40,000 words)*. Dostupné z: http://pi.math.cornell.edu/~mec/2003-2004/cryptography/subs/frequencies.html [12. 7. 2021].

kromě měření podobností v současném jazyce bude zohledněn jazykový vývoj, a dalších historiografických metod a celé řady dalších poznatků z různých oborů.

2. Ikonicita ve jménech vodních toků

Podobně, jako jsme se v předchozí kapitole věnovali jménům měst, budeme se v této kapitole zabývat jmény vodních toků. Klademe si otázku, zda jsou jména vodních toků arbitrární, nebo zda najdeme souvislost mezi některými kvantifikovatelnými vlastnostmi jmen a kvantifikovatelnými vlastnostmi samotných toků. To budeme zjišťovat pomocí měření Spearmanova korelačního koeficientu a p-hodnoty, pomocí které zjistíme, zda případná korelace nevznikla pravděpodobně pouze vlivem náhody.

Jména vodních toků mají bohatou etymologii, která je popsána v řadě publikací a studií.[61] Zde je opět potřeba terminologicky odlišit nearbitrárnost od motivovanosti. Souvislostí mezi jménem vodního toku a dalšími slovy jazyka rozumíme *motivovanost (*např. *Mlýnský potok* – malý vodní tok, který tekl v okolí mlýna), ta je však v naší práci pouze na okraji zájmu, protože s vlastní arbitrárností jazykového znaku, tedy vnitřním vztahem mezi označujícím a označovaným (např. řeka *Hučava*, jejíž název imituje hukot vody), souvisí jen minimálně.

2.1. Korelace mezi délkou toku a délkou názvu

V minulé kapitole jsme nalezli signifikantně pozitivní korelaci mezi délkou jména obce a počtem jejích obyvatel. Abychom zjistili, zda se nejednalo o ojedinělý případ signifikantní korelace mezi jménem objektu a jeho kvantifikovatelnými vlastnostmi, budeme podobnými metodami zjišťovat, zda existuje vztah mezi délkou toku a délkou jeho názvu. Toho docílíme měřením Spearmanova korelačního koeficientu. Na základě de Saussurovy definice arbitrárnosti znaku však žádnou statisticky významnou korelaci neočekáváme, protože by neměl existovat žádný vnitřní vztah mezi *signifiant* (jménem řeky) a *signifié* (řekou samotnou, tedy ani žádnou její charakteristickou vlastností, např. délkou).

K tomuto pokusu použijeme dataset všech vodních toků, které se vyskytují na území dnešní Francie.[62] Francouzská jména jsme vybrali z důvodu, že Francie je země o relativně velké

[61] Viz např. McCafferty, M. (2004). Correction: Etymology of Missouri. *American Speech 79*(1), 32. https://www.muse.jhu.edu/article/54836 nebo Room, A. (2006). *Placenames of the World: Origins and Meanings of the Names for 6,600 Countries, Cities, Territories, Natural Features, and Historic Sites.* McFarland, Incorporated.

[62] Data jsou získána z webu *Sandre* (*Service d'administration nationale des données et des référentiels sur l'eau, 2021*). Dostupné z: http://www.sandre.eaufrance.fr/Rechercher-une-donnee-d-un-jeu [13. 7. 2021].

rozloze a obsahuje dostatečný počet vodních toků, aby s nimi bylo možné pracovat pomocí kvantitativních metod. Zároveň jsme chtěli pokus provést i na jiném jazyce, než byla čeština (viz 1.1. – hledání korelací mezi velikostí obce a délkou jejího názvu, datasetem byly obce ČR) nebo angličtina (viz 1.2. – trénování SVM modelu s cílem rozlišit hlavní a periferní města, jména měst byla uvedena v angličtině), a to z důvodu, aby zjištěné výsledky nebyly způsobeny pouze charakteristickými rysy českého, resp. anglického jazyka.

Dataset obsahuje celkem 1 102 vodních toků a měl by tak zahrnovat všechny důležité francouzské toky. Ve francouzštině se rozlišuje několik typů řek, které nelze snadno přeložit. *Fleuve* je řeka, která ústí do moře, *rivière* je přírodní vodní tok střední důležitosti, který se vlévá do jiného vodní toku, *ruisseau* (potok) je malý vodní tok, jehož šířka dosahuje 1–5 metrů, *ru* je malý vodní tok, jehož šířka dosahuje méně než 1 metru.[63] Náš dataset obsahuje všechny *fleuves* a *rivières* a kromě toho jsou v něm zastoupeny také *ruisseaux* a *rus*, ty však nejsou vzhledem k jejich velkému počtu a pouze lokálnímu významu zastoupeny kompletně. Jinými slovy, dataset obsahuje všechny důležité vodní toky i zástupce drobných vodních toků.

Délka toku je uváděna v kilometrech. Důležité je, že zahrnuje pouze délku toku na území Francie, nikoli jeho celkovou délku (ukázka dat viz *Tab. 17*).

Ukázka datasetu – vodní toky a jejich délky

Vodní tok	Délka (v km)	Délka názvu (počet písmen)
Loire	1 006	5
Seine	775	5
Rhône	545	5
Garonne	523	7
Marne	514	5
Meuse	486	5
Lot	485	3

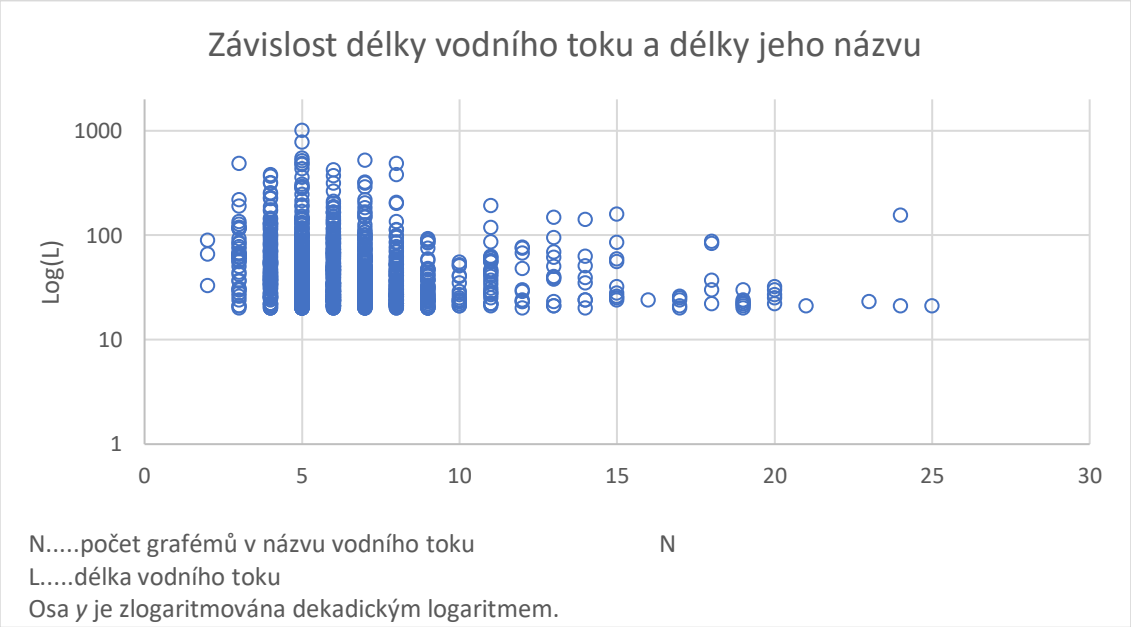
Tab. 17: Ukázka datasetu jmen vodních toků, délek vodních toků (v km) a délek názvů (počet písmen).

Pomocí počítačového programu spočítáme délky názvů všech vodních toků. Tyto délky odpovídají počtům grafémů (včetně mezery), např. řece *Gave de Pau* tak bude přiřazena délka názvu 11. Pro náhled na data vytvoříme bodový graf (*Graf 6*). Ze získaných dat spočítáme hodnotu Spearmanova korelačního koeficientu této délky názvu a délky toku v km, který udává, jak silná

[63] Viz *Culture-générale.fr* (2012). *Nombre de cours d'eau en France*. Dostupné z: https://www.culture-generale.fr/geographie/8450-nombre-de-cours-deau-en-france [13. 7. 2021].

závislost existuje mezi dvěma proměnnými. Poté spočítáme p-hodnotu, která vyjadřuje pravděpodobnost, s jakou by stejná nebo ještě extrémnější korelace mohla vzniknout pouze vlivem náhody. Hladinu alfa zvolíme opět na typickou hodnotu 0,05.
Získáváme hodnotu Spearmanova koeficientu –0,237 a p-hodnotu $p = 1{,}484 \times 10^{-15}$, která nám říká, že takovouto nebo extrémnější korelaci uvidíme pouze vlivem náhody v 0,0000000001484 % případů. Mezi délkou vodního toku a délkou jeho názvu tedy existuje signifikantní slabá záporná korelace.

Z grafu je vizuálně patrné, že nejdelší vodní toky mají v názvu pět grafémů (srov. *Seine, Loire, Rhône*), zatímco nejčastější délka názvu vodního toku (tedy medián) je 6 grafémů. V grafu je zřejmá obecně klesající tendence, přičemž v levé části grafu od vrcholu 5 grafémů vidíme lokální tendenci přímé úměry: čím delší je jméno vodního toku, tím delší je i samotný tok. Naopak napravo od vrcholu 5 grafémů vidíme lokální vztah nepřímé úměry: čím delší je název vodního toku, tím kratší je samotný vodní tok. Celkový trend je také klesající, tedy čím delší má tok název, tím je kratší, což odpovídá zákonu o jazykové ekonomii.



Graf 6: Závislost mezi délkou vodního toku (v km) a délkou jeho názvu (v počtech grafémů).

Na rozdíl od korelace mezi velikostí obce a délkou jejího názvu, která byla signifikantně pozitivní, a tedy platilo, že čím více má obec obyvatel, tím je delší její název, je korelace mezi délkou vodního toku a délkou jeho názvu signifikantně negativní, a platí tedy čím delší je jméno vodního toku, tím kratší tento vodní tok je. Tuto korelaci je možné odůvodnit dříve popsaným Zipfovým principem jazykové ekonomie, uvážíme-li, že čím je řeka delší, tím je významnější

a tím více se o ní mluví, a tedy se název řeky pravděpodobně jen tímto vlivem začne postupem času zkracovat. Na základě tohoto pokusu tedy nelze vyvrátit, že by neexistoval vztah mezi délkou jména vodního toku v grafémech a jejich fyzickou délkou.

2.2. Rozpoznávání jmen měst a jmen vodních toků

V následujícím pokusu bude cílem zjistit, zda SVM model dokáže správně rozlišit jména vodních toků od jmen měst. Budeme pracovat s datasety měst a vodních toků na území dnešní Francie, které jsme vybrali z důvodu popsaného výše (viz 2.1), tedy kvůli dostatečnému počtu měst i vodních toků na území Francie a z důvodu rozšíření našich pokusů i mimo rámec češtiny a angličtiny.

Datasety obsahují 1 102 francouzských řek (důkladnější popis datasetu, zdroj a terminologie viz kapitola 2.1.) a 35 357 francouzských obcí.[64] Během pokusu byla data načtena, ztokenizována, konvertována na malá písmena a byly odstraněny všechny nealfanumerické znaky. Vzhledem k rozdílné velikosti obou datasetů byly položky delšího z nich (tedy jména měst) náhodně zamíchány a zkráceny na velikost kratšího z nich (tedy na velikost datasetu vodních toků). Data byla opět rozdělena na trénovací a testovací v poměru 2/3, přičemž prvky do obou datasetů byly pokaždé vybírány náhodně. K natrénování modelu tak bylo použito 1 468 jmen (734 jmen měst a stejné množství jmen vodních toků) a zbylých 736 jmen (368 jmen měst a 368 jmen vodních toků) bylo využito k testování úspěšnosti predikce modelu na dosud neviděných datech.

Jména byla reprezentována pomocí frekvenčního BoW jednotlivých grafémů stejně, jako v kapitole 1.2.2., a následně bude pro srovnání vyzkoušena i možnost s reprezentací pomocí binárního BoW, který namísto frekvencí obsahuje pouze 1 nebo 0 podle toho, zda se v názvu daný grafém vyskytuje, nebo nevyskytuje. Využijeme SVM model s lineárním jádrem (tedy základní možnost bez jaderné transformace, u které existuje možnost zjistit koeficienty jednotlivých vlastností).

Po natrénování modelu tímto způsobem měříme úspěšnost jeho predikce pro trénovací dataset (tj. pro již viděná data) i pro testovací dataset (tj. dosud neviděná data). Tento postup provádíme

[64] Dataset dostupný z: Insee (*Institut national de la statistique et des études économiques*, 2018). *Liste des communes existantes au 1er janvier 2018.* Dostupné z: https://www.insee.fr/fr/information/3363419#titre-bloc-7 [14. 7. 2021].

1 000× a výsledky shrnujeme pomocí průměru, mediánu a 95% konfidenčního intervalu (viz *Tab. 18* a *19*).

Přehled výsledků pro trénovací dataset s využitím frekvenčního BoW

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	< 0,809; 0,843 >	0,826	0,826
Preciznost (*Average precision*)	< 0,74; 0,781 >	0,76	0,76
Úplnost (*Recall*)	< 0,853; 0,89 >	0,872	0,872
F1	< 0,818; 0,85>	0,833	0,833

Tab. 18: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím frekvenčního BoW.

Přehled výsledků pro testovací dataset s využitím frekvenčního BoW

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	<0,787; 0,84 >	0,813	0,814
Preciznost (*Average precision*)	< 0,716; 0,778 >	0,746	0,746
Úplnost (*Recall*)	< 0,823; 0,897 >	0,862	0,861
F1	< 0,796; 0,846>	0,822	0,822

Tab. 19: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím frekvenčního BoW.

Úspěšnost predikce modelu je pro trénovací i testovací data vyšší než 0,5. Model tedy dokáže s úspěšností vyšší, než je úspěšnost náhodného modelu, správně rozlišit jména měst od jmen vodních toků, a to jak pro jména, na kterých byl model trénován, tak i pro zatím neviděná jména.

Dále vyzkoušíme variantu, ve které nebudeme jména řek reprezentovat frekvenčním BoW, který obsahuje počet výskytů jednotlivých grafémů ve jméně, nýbrž binárním BoW, tedy vektorem 1 a 0 podle toho, zda se v názvu vyskytuje grafém z globálního slovníku či nikoli. V tomto případě tak ztrácíme nejen informaci o pořadí znaků, ale také o jejich frekvenci.

Z tabulek (*Tab. 20* a *21*) vidíme, že úspěšnost predikce modelu, ve kterém byly názvy reprezentovány vektorem 1 a 0, je pro trénovací i testovací data výrazně vyšší než 0,5. Model tedy opět dokáže s úspěšností vyšší, než je úspěšnost náhodného modelu, správně rozlišit jména měst

od jmen vodních toků pro trénovací i testovací data. Výsledky predikce při využití reprezentace frekvenčním BoW a binárním BoW se jeví podobné.

Přehled výsledků pro trénovací dataset s využitím binárního BoW

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	<0,798; 0,831 >	0,815	0,815
Preciznost (*Average precision*)	< 0,73; 0,768 >	0,75	0,75
Úplnost (*Recall*)	< 0,831; 0,875 >	0,853	0,853
F1	< 0,806; 0,837>	0,822	0,822

Tab. 20: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím binárního BoW.

Přehled výsledků pro testovací dataset s využitím binárního BoW

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	<0,773; 0,829 >	0,802	0,802
Preciznost (*Average precision*)	< 0,705; 0,766 >	0,735	0,735
Úplnost (*Recall*)	< 0,802; 0,88 >	0,84	0,84
F1	< 0,782; 0,834>	0,809	0,809

Tab. 21: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím binárního BoW.

Lze namítnout, že model úspěšně rozlišuje jméno řeky od jména města jen proto, že jména řek obsahují typická slova, jako *řeka*, *potok*, *voda*, která se často opakují. Po zanalyzování datasetu zjišťujeme, že názvy vodních toků obsahují slova *eau* (fr. *voda*), např. *Eau Bourde, Eau Mère*, dále *rivière* (fr. *řeka*), např. *Rivière d'Étel, Rivière de Penerf*, slovo *ruisseau* (fr. *potok*), např. *Ruisseau de Neuffonds, Ruisseau d'Escalmels,* a *ru* (malý potok, viz výše uvedená terminologie), např. *Ru du Cuivre, Ru de Bréon*. Tyto názvy navíc obsahují ve větší míře předložku *de* (tato předložka vyjadřuje vazbu na 2. pád).

Abychom zjistili, zda tato slova v názvech mají vliv na rozhodování modelu, zda je dané jméno jménem města nebo vodního toku, všechny názvy obsahující výše zmíněná slova z datasetu odstraníme. Tímto postupem nám zbyde 1 079 jmen řek, která neobsahují žádné z uvedených

slov v názvu. Model natrénujeme stejným způsobem a evaluujeme jako výše a porovnáme úspěšnost predikce (*Tab. 22* a *23*).

Přehled výsledků pro trénovací dataset s využitím frekvenčního BoW

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	<0,64; 0,688 >	0,664	0,663
Preciznost (*Average precision*)	< 0,586; 0,622 >	0,604	0,603
Úplnost (*Recall*)	< 0,73; 0,847 >	0,785	0,782
F1	< 0,674; 0,727 >	0,7	0,7

Tab. 22: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím frekvenčního BoW po odstranění často se vyskytujících slov ve jménech řek.

Přehled výsledků pro testovací dataset s využitím frekvenčního BoW

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	<0,626; 0,69 >	0,66	0,66
Preciznost (*Average precision*)	< 0,576; 0,624 >	0,6	0,6
Úplnost (*Recall*)	< 0,708; 0,852 >	0,78	0,778
F1	< 0,661; 0,728>	0,696	0,696

Tab. 23: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím frekvenční BoW po odstranění často se vyskytujících slov ve jménech řek.

Z tabulek je patrné, že po odstranění často se opakujících slov ve jménech řek sice úspěšnost predikce mírně poklesla, avšak stále je statisticky prokazatelně vyšší, než by byla úspěšnost náhodného modelu, neboť hodnoty jednotlivých evaluačních metrik jsou stále vyšší než 0,5.

Další námitkou by mohlo být často opakované slovo *ville* (fr. *město*) v názvech obcí. Dataset měst je sice dostatečně velký a převáží v něm města bez tohoto slova v názvu, ale i přesto je třeba ověřit, zda za úspěšností modelu nestojí jen tento fakt (nebo ji alespoň do velké míry nezlepšuje). Odstraníme tedy všechna města obsahující slovo *ville* a natrénujeme a evaluujeme model stejným způsobem (viz *Tab. 24* a *25*).

I po odstranění všech jmen obsahujících charakteristická slova z datasetů je stále úspěšnost modelu prokazatelně vyšší než úspěšnost náhodného modelu. Lze tedy předpokládat, že jména

řek nesou informaci o tom, že označují vodní tok na úrovni jednotlivých grafémů jmen vodních toků.

Přehled výsledků pro trénovací dataset s využitím frekvenčního BoW

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	<0,629; 0,674 >	0,651	0,651
Preciznost (*Average precision*)	< 0,578; 0,612 >	0,594	0,594
Úplnost (*Recall*)	< 0,729; 0,825 >	0,78	0,783
F1	< 0,668; 0,713 >	0,691	0,691

Tab. 24: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím frekvenčního BoW po odstranění často se vyskytujících slov ve jménech řek i měst.

Přehled výsledků pro testovací dataset s využitím frekvenčního BoW

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	<0,615; 0,678 >	0,646	0,646
Preciznost (*Average precision*)	< 0,568; 0,616 >	0,591	0,59
Úplnost (*Recall*)	< 0,708; 0,833 >	0,777	0,778
F1	< 0,653; 0,715>	0,687	0,688

Tab. 25: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím frekvenční BoW po odstranění často se vyskytujících slov ve jménech řek i měst.

2.2.1. Analýza arbitrárnosti jednotlivých grafémů vodních toků

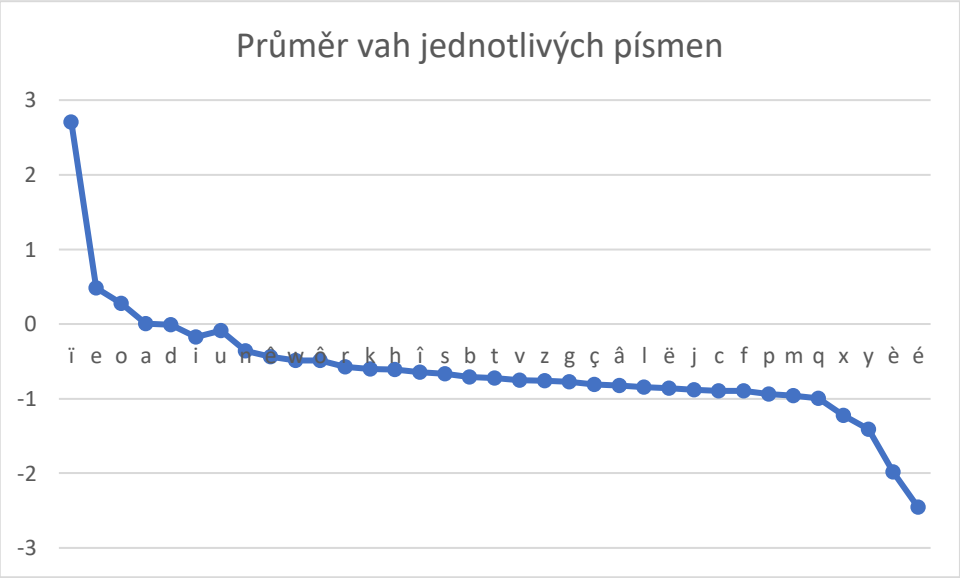
Nyní se budeme zabývat tím, jakou úlohu hrají jednotlivé grafémy při rozhodování modelu. Vzhledem k tomu, že jsme pro rozlišování mezi jmény měst a jmény vodních toků trénovali SVM model s lineárním jádrem, existuje opět možnost získat koeficienty (váhy) jednotlivých grafémů, abychom tak zjistili, zda se na jejich základě model rozhodoval spíše pro třídu měst, nebo pro třídu vodních toků (přehled viz *Tab. 26* a grafické znázornění viz *Graf 9*).

Platí, že kladné hodnoty mají grafémy, které dodávají informaci o třídě vodních toků, zatímco záporné hodnoty mají znaky, které dodávají informaci o třídě měst s konkrétní váhou. Čím dále je hodnota od nuly, tím vyšší má váhu pro predikci pro první či druhou třídu.

Přehled vah jednotlivých písmen

Znak	Průměr	Medián	Znak	Průměr	Medián
ï	2,706	2,706	**V**	–0,751	–0,751
e	0,48	0,478	**Z**	–0,757	–0,771
o	0,275	0,274	**G**	–0,774	–0,774
a	0,004	0,005	**Ç**	–0,813	–1
d	–0,013	–0,01	**Â**	–0,827	–1
i	–0,175	–0,174	**l**	–0,848	–0,848
u	–0,087	–0,089	**ë**	–0,863	–1
n	–0,359	–0,358	**j**	–0,881	–0,889
ê	–0,438	0.0	**c**	–0,894	–0,891
w	–0,487	–0,489	**f**	–0,897	–0,885
ô	–0,491	–0,324	**p**	–0,94	–0,93
r	–0,578	–0,573	**m**	–0,96	–0,956
k	–0,601	–0,694	**q**	–0,996	–1
h	–0,61	–0,608	**x**	–1,226	–1,217
î	–0,644	–0,816	**y**	–1,411	–1,413
s	–0,665	–0,66	**è**	–1,982	-2
b	–0,707	–0,703	**é**	–2,457	–2,543
t	–0,722	–0,72			

Tab. 26: Průměry a mediány váhy jednotlivých písmen. Řazeno podle průměru sestupně.



Graf 9: Aritmetický průměr koeficientů jednotlivých písmen.

Na základě tabulky lze říci, že podle grafémů *ï, e, o, a* se model rozhoduje spíše pro třídu řek, zatímco podle ostatních grafémů se rozhoduje spíše pro třídu měst. Je to poněkud zvláštní výsledek, protože v názvech řek se podle koeficientů vyskytují ve větší míře především samohlásky. Přitom grafém *ï* nemůžeme zohledňovat, protože se v datasetu vodních toků vyskytuje pouze jednou a v datasetu francouzských měst ani jednou.

Nabízí se otázka, zda by pro úspěšnou predikci nebyl postačující počet samohlásek, které jméno obsahuje. Tuto možnost vyzkoušíme. Každé jméno reprezentujeme číselným vektorem, který bude obsahovat počet samohlásek *a, e, i, o, u*, natrénujeme model a necháme jej predikovat pro trénovací i testovací dataset, přičemž měříme úspěšnost jeho predikce (*Tab. 27* a *28*).

Přehled výsledků pro trénovací dataset s využitím frekvencí samohlásek

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	<0,636; 0,676 >	0,656	0,656
Preciznost (*Average precision*)	< 0,582; 0,613 >	0,598	0,598
Úplnost (*Recall*)	< 0,736; 0,827 >	0,778	0,774
F1	< 0,675; 0,716 >	0,694	0,692

Tab. 27: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím frekvencí samohlásek a, e, i, o, u.

Přehled výsledků pro testovací dataset s využitím frekvencí samohlásek

Evaluační metrika	Konfidenční interval	Průměr	Medián
Přesnost (*Accuracy*)	<0,62; 0,683 >	0,652	0,652
Preciznost (*Average precision*)	< 0,571; 0,619 >	0,595	0,595
Úplnost (*Recall*)	< 0,712; 0,842 >	0,775	0,771
F1	< 0,657; 0,723 >	0,69	0,69

Tab. 28: Konfidenční intervaly, průměry a mediány jednotlivých evaluačních metrik pro trénovací dataset s využitím frekvencí samohlásek a, e, i, o, u.

Z výsledků je zřejmé, že modelu stačí znát pouze frekvence výskytu samohlásek *a, e, i, o, u* v daném jméně, aby dokázal predikovat, zda se jedná o město, nebo o vodní tok, s úspěšností predikce, která je vyšší než úspěšnost náhodného modelu.

Nicméně je třeba uvážit, že francouzština je jazyk, ve kterém se do velké míry neshoduje psaná forma slova s jeho výslovností, a to právě obzvláště u samohlásek (např. *eau*, fr. voda, vyslovíme jako [o] – tři grafémy tak odpovídají jednomu fonému). To, že jména řek obsahují oproti jménům měst charakteristické znaky, tedy do velké míry platí pouze v psané podobě.

2.2.2. Shrnutí a diskuze

V této kapitole jsme měřili korelace mezi délkou vodního toku a délkou jeho názvu, abychom zjistili, zda se jména vodních toků chovají arbitrárně. Pokus byl proveden na příkladě francouzských vodních toků. Zjistili jsme, že mezi názvem a délkou toku existuje signifikantní slabá záporná korelace. Jedním z možných vysvětlení, které zde považujeme za nejpravděpodobnější, je Zipfův princip jazykové ekonomie. Zjištění tak nemusí s arbitrárností nutně souviset.

Dále jsme zkoušeli natrénovat SVM model tak, aby dokázal rozlišovat mezi jmény vodních toků a jmény měst. Bylo zjištěno, že model dokáže predikovat s prokazatelně lepší úspěšností, než je úspěšnost náhodného modelu, mezi jmény vodních toků a jmény měst, když jsou reprezentována frekvenčním BoW, binárním BoW, nebo dokonce jen počty samohlásek *a, e, i, o, u* v názvu. Model dokáže úspěšně predikovat, i když z datasetů odstraníme slova, která se v názvech typicky často opakují – např. *ville* pro města nebo *ruisseau* pro vodní toky.

Zamysleme se ale ještě dále nad vztahem měst a vodních toků, protože jejich názvy spolu souvisí a mohly by se vzájemně ovlivňovat. Města byla zakládána podél toků řek, protože lidé k životu potřebují vodu. Města jsou také často pojmenována podle toku, v jehož blízkosti se nachází (např. *Lipník nad Bečvou, Ústí nad Labem*, z datasetu francouzských měst např. *Cormoranche-sur-Saône, Gannay-sur-Loire*). Názvy měst také často obsahují obecná slova označující vodní toky (např. *Červená Voda, Rio de Janeiro*).

Může vyvstat otázka, zda byla nejprve pojmenována řeka, nebo město, které u ní leží. Odpověď bude zřejmě pro konkrétní města různá. Například město Amsterdam bylo pojmenováno podle řeky Amstel,[65] avšak např. Budišov nad Budišovkou byl pojmenován podle klášterního opata nebo od bud horníků, kteří ve městě sídlili, a řeka až následně.[66] Některé obce v názvu řeku původně neměly, až postupem času vznikla potřeba upřesnit, o kterou Lhotu nebo Město se

[65] Srov. *Online ethymology dictionary* (2017). Heslo *Amsterdam*. Dostupné z: https://www.etymonline.com/word/amsterdam [15. 7. 2021].

[66] Srov. MěÚ Budišov nad Budišovkou (2021). *Historie*. Dostupné z: https://www.budisov.eu/mesto/historie-1/ [15. 7. 2021].

jedná. V každém případě jména řek mají se jmény měst velmi provázaný vztah. O to neočekávanější je, že model mezi nimi dokáže správně rozlišovat.

Položme si otázku, zda představuje schopnost modelu rozlišit mezi vodním tokem a městem podle jména protipříklad arbitrárnosti jazykového znaku. Jisté je, že pokud mezi oběma třídami dokáže model rozlišit, potom musí existovat vnitřní vztah mezi *signifiant* a *signifié*. Nelze však vyloučit, že tento vztah může být motivován nějakým externím faktorem, který by způsoboval, že např. jména řek obsahují více vokálů než jména měst. Hledání tohoto faktoru už by však vyžadovalo hloubkovou kvalitativní analýzu, hledání kontextu pojmenovávání konkrétních vodních toků i měst v průběhu historie až do současnosti. Takovýto výzkum už by však dalece přesahoval rámec této diplomové práce.

Závěr

Cílem této práce bylo kvantitativními metodami ověřit arbitrárnost jazykového znaku, jak ji definoval Ferdinand de Saussure. Na základě jeho definice jsme předpokládali, že slova jsou arbitrární, tedy že neexistuje žádný vnitřní vztah mezi označujícím a označovaným.

Za účelem nahlédnutí arbitrárnosti jsme kvantifikovali různé vlastnosti objektů a jejich názvů. Konkrétně jsme pracovali s hydronymy a toponymy. Pomocí statistických metod jsme testovali rozdíly mezi skupinami a hledali jsme korelace mezi kvantifikovatelnými vlastnostmi měst a vodních toků a kvantifikovatelnými vlastnostmi jejich názvů. Pomocí metod strojového učení jsme testovali přítomnost vzorů v rámci jednotlivých tříd, které se model snažil nalézt za účelem správného zařazení objektu do třídy. Úspěšnost predikce byla evaluována na trénovacích i testovacích datech.

V první kapitole jsme se zabývali arbitrárností jmen měst. Zjišťovali jsme korelace mezi počtem obyvatel obce a délkou jejího názvu měřenou v počtech grafémů a v počtech slov. Tímto postupem byly odhaleny signifikantně pozitivní korelace mezi počtem obyvatel obce a délkou jejího názvu měřenou jak v počtech grafémů, tak v počtech slov. Diskutovali jsme možné příčiny těchto korelací. Signifikantně pozitivní korelace by mohly například souviset s potřebou upřesňovat jména měst, která jsou středně důležitá, což implikuje automaticky delší název. I přes toto behaviorální vysvětlení zde však pozorujeme trend založený na ikonicitě – malé obce mají nejkratší názvy.

Ve druhé části první kapitoly jsme trénovali SVM model a zjišťovali, zda dokáže rozlišovat mezi hlavními a periferními městy. Jména měst byla reprezentována binárním a frekvenčním BoW. Na základě definice arbitrárnosti jazykového znaku byla očekávaná průměrná úspěšnost 0,5. Po natrénování modelu jsme evaluovali úspěšnost jeho predikce na trénovacích i testovacích datech. Ve všech případech dokázal model správně určit, zda se jedná o město ze třídy hlavních, nebo periferních měst, se statisticky prokazatelnou úspěšností vyšší, než je 0,5, tedy než úspěšnost náhodného modelu. Úspěšnost modelu však lze vysvětlit i jinými způsoby, než je nearbitrárnost. V textu práce jsme se zabývali myšlenkami, že za úspěšností modelu může stát např. doba vzniku hlavních měst, která budou až na výjimky obecně starší než periferní města, nebo jazyk, který byl v době pojmenovávání měst dominantní.

Ve druhé kapitole jsme se zabývali arbitrárností jmen vodních toků. Byla zjišťována korelace mezi délkou vodního toku a délkou jeho názvu a bylo odhaleno, že mezi nimi existuje signifikantní slabá záporná korelace. Tuto korelaci lze vysvětlit například Zipfovým principem jazykové ekonomie, podle kterého má být mezi frekvencí slova a jeho délkou vztah nepřímé úměry.

Dále jsme trénovali SVM model za účelem rozlišování mezi jmény měst a jmény vodních toků. Úspěšnost predikce modelu byla opět statisticky prokazatelně výrazně vyšší, než by byla úspěšnost predikce náhodného modelu.

Přítomnost signifikantních korelací a zjištěná úspěšnost modelu strojového učení je v rozporu s naším očekáváním arbitrárnosti jazykového znaku. Lze konstatovat, že jasně existují tendence, které grafémy budou obsaženy ve jménech měst, hlavních měst či vodních toků. Na základě uvedených pokusů však nelze jazykovou arbitrárnost vyvrátit (či potvrdit), protože, jak se v průběhu práce ukázalo, existuje příliš mnoho dalších faktorů, které musí být zváženy, a pro další interpretaci zjištěných výsledků by bylo dále zapotřebí rozsáhlého výzkumu, který by kombinoval synchronní i diachronní metody a zohledňoval a popisoval jazykový vývoj, stejně jako historický a politický kontext, který jména měst ovlivnil.

Závěrem můžeme říci, že jazykovou arbitrárnost na základě kvantitativních metod nemůžeme vyvrátit, ani jednoznačně potvrdit. V souladu s výsledky pokusů můžeme pouze nastínit možnost, že by v jazyce nemuselo být vše arbitrární. Obecně se ukázalo, že problematika arbitrárnosti jmen je mnohem širší a komplikovanější, než se zpočátku mohlo zdát, a že ji nelze obsáhnout pouze pomocí statistiky a dalších kvantitativních metod.

Bibliografie

Albrecht, J., Ramachandran, S. a Winkler, C. (2020). *Blueprints for Text Analytics Using Python*. O´Reilly Media.

Bentz, C. a Ferrer-i-Cancho, R. (2016). *Zipf's law of abbreviation as a language universal*, http://dx.doi.org/10.15496/publikation-10057.

Best, D. a Roberts, D. (1975). Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *24*(3), 377-379, https://doi.org/10.2307/2347111.

Brown, R. W., Black, A. H., & Horowitz, A. E. (1955). Phonetic symbolism in natural languages. *The Journal of Abnormal and Social Psychology*, *50*(3), https://doi.org/10.1037/h0046820.

Brownlee, J. (2016). *Overfitting and Underfitting With Machine Learning Algorithms*. Machine Learning Algorithms. Dostupné z: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/ [1. 7. 2021].

Bulliet, R. (2011). *Cotton, climate, and camels in early Islamic Iran: a moment in world history.* Columbia University Press.

Culture-générale.fr (2012). *Nombre de cours d'eau en France*. Dostupné z: https://www.culture-generale.fr/geographie/8450-nombre-de-cours-deau-en-france [13. 7. 2021].

Černý, J. (1996). *Dějiny lingvistiky*. Votobeia.

Ferro, M. *Dějiny kolonizací. Od dobývání až po nezávislost 13. – 20. století*. NLN, Praha 2007.

Grefenstette, G. a Tapanainen, P. (1997). What is a word, What is a sentence? Problems of Tokenization. Dostupné z: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.8947&rep=rep1&type=pdf [3. 8. 2021].

Hastie, T., Tibshirani R. a Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2. vydání, Springer. Dostupné z: https://web.stanford.edu/~hastie/ElemStatLearn/ [8. 4. 2021].

Hollander, M. a Wolfe, D. A. (1973). *Nonparametric Statistical Methods.* New York: John Wiley & Sons.

Internetová jazyková příručka [online] (2008–2021). Praha: Ústav pro jazyk český AV ČR. https://prirucka.ujc.cas.cz/ [1. 8. 2021].

IPA (International Phonetic Association, 2020). *The international phonetic alphabet*. Dostupné z: https://www.internationalphoneticassociation.org/IPAcharts/IPA_chart_orig/pdfs/IPA_Kiel_2020_full.pdf [4.7.2021].

Math Explorers' Club (2019). *English Letter Frequency (based on a sample of 40,000 words)*. Dostupné z: http://pi.math.cornell.edu/~mec/2003-2004/cryptography/subs/frequencies.html [12. 7. 2021].

McCafferty, M. (2004). Correction: Etymology of Missouri. *American Speech 79*(1), 32. https://www.muse.jhu.edu/article/54836.

MěÚ Budišov nad Budišovkou (2021). *Historie*. Dostupné z: https://www.budisov.eu/mesto/historie-1/ [15. 7. 2021].

New World Encyclopedia (2018). Heslo *Jakarta*. Dostupné z: https://www.newworldencyclopedia.org/entry/Jakarta [31. 7. 2021].

Online ethymology dictionary (2017). Heslo *Amsterdam*. Dostupné z: https://www.etymonline.com/word/amsterdam [15. 7. 2021].

Organiser (2018). *How Renaming the Cities is a Part of Decolonisation and Historical Justice*. Dostupné z: https://www.organiser.org/Encyc/2018/10/20/How-Renaming-the-Cities-is-a-Part-of-Decolonisation-and-Historical-Justice.html. [10. 7. 2021].

Peirce, C. S. (1894). *The Art of Reasoning. Kapitola II. What is a sign?* MS [R] 404; MS [R] 1009.

Popescu, I. (2009). Word Frequency Studies. Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110218534.

Portal do Governo Brasileiro. Brasília, História & Fotos (2014). Dostupné z: https://cidades.ibge.gov.br/brasil/df/brasilia/historico [10. 7. 2021].

Powers, D. M. (2011): *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. Journal of Machine Learning Technologies, 2(1), s. 37–63.

Profous, A. (1951). *Místní jména v Čechách: jejich vznik, pův. význam a změny*. 3. díl: M-Ř. 1. vyd. Praha: Nakladatelství Československé akademie věd.

Radding, L. a Western, J. (2010). What's In the Name? Linguistics, Geography, and Toponyms. *Geographical Review*, 100: 394–412, https://doi.org/10.1111/j.1931-0846.2010.00043.x.

Room, A. (2006). *Placenames of the World: Origins and Meanings of the Names for 6,600 Countries, Cities, Territories, Natural Features, and Historic Sites.* McFarland, Incorporated.

Rozycki, W. (1997) Phonosymbolism and the Verb. *Journal of English Linguistics*, 25/3, s. 202–206, https://doi.org/10.1177/007542429702500303.

Rutkiewicz-Hanczewska, M. (2012). Iconicity in urban place naming (with examples of names from places in Poland). *Semiotica*, *2012*(189), 49-64, https://doi.org/10.1515/semi.2011.077.

Saussure, F., Bally, C., Sechehaye, A. (1996). *Kurs obecné lingvistiky*. Přeložil Čermák, F. Praha: Academia.

Schwanzer, V. (1978). Slovo a text. *Slovo a slovesnost, 39 (1978), 3-4, s. 259-261*. Dostupné z: http://sas.ujc.cas.cz/archiv.php?art=2548 [3. 4. 2021].

Sigurd, B., Eeg-Olofsson, M. a Van Weijer, J. (2004), Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica*, 58: 37-52. https://doi.org/10.1111/j.0039-3193.2004.00109.x.

Stefan Trost Media (2007–2021). *Alphabet and Character Frequency: Spanish (Español)*. Dostupné z: https://www.sttmedia.com/characterfrequency-spanish [12. 7. 2021].

Steinwart, I. a Christmann, A. (2008). Support Vector Machines, Springer-Verlag, New York.

The Editors of Encyclopaedia Britannica (2015). *Georgetown, national capital, Guyana*. Dostupné z: https://www.britannica.com/place/Georgetown-Guyana [12. 7. 2021].

Ústav pro jazyk český ČSAV (2011). *Slovník spisovného jazyka českého*. Dostupné z: https://ssjc.ujc.cas.cz/search.php?db=ssjc [2. 8. 2021].

Vapnik, V. (1963). *Pattern Recognition Using Generalized Portrait Method*. Automation and Remote Control, 774–780.

versus.com (2021). *Brasilia vs Rio de Janeiro: What is the difference?* Dostupné z: https://versus.com/en/brasilia-vs-rio-de-janeiro [13. 5. 2021]

Whitney, W. D. (1879). *A Sanskrit grammar: including both the classical language, and the older dialects (of Veda and Brahmana)*. Lipsko: Breitkopf a Härtel. Dostupné z: https://archive.org/details/sanskritgrammari00whituoft/page/2/mode/2up?view=theater [6.7.2021].

Zipf, G. K. (1936). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton-Mifflin, Boston.

Zipf, G. K. (1949). *Human behaviour and the principle of the least effort*. Addison-Wesley Press.

Zdroje dat

Collectivités locales. *Les chiffres-clés des collectivités locales* (2018). Dostupné z: https://www.collectivites-locales.gouv.fr/files/files/statistiques/brochures/chapitre_1_-_les_chiffres_cles_des_collectivites_locales_1.pdf [18. 2. 2021].

Cvrček, V. a Vondřička, P. (2011). SyD – Korpusový průzkum variant. FF UK. Praha 2011. Dostupné z: https://syd.korpus.cz/ [4. 4. 2021].

Česko-Slovenská filmová databáze (2001–2021). *Žebříčky – nejlepší filmy.* Dostupné z: https://www.csfd.cz/zebricky/filmy/nejlepsi/ [12. 3. 2021].

Česko-Slovenská filmová databáze (2001–2021). *Žebříčky – nejlepší seriály.* Dostupné z: https://www.csfd.cz/zebricky/serialy/nejlepsi/ [12. 3. 2021].

Český statistický úřad (2020). *Tab. 3 Počet obyvatel v obcích České republiky k 1. 1. 2020*. Dostupné z: https://www.czso.cz/csu/czso/pocet-obyvatel-v-obcich-k-112019 [2. 12. 2020].

Český statistický úřad (2018). *Stav a pohyb obyvatelstva v ČR – rok 2017*. Dostupné z: https://www.czso.cz/csu/czso/stav-a-pohyb-obyvatelstva-v-cr-rok-2017 [14. 5. 2020].

Insee (*Institut national de la statistique et des études économiques*, 2018). *Liste des communes existantes au 1er janvier 2018.* Dostupné z: https://www.insee.fr/fr/information/3363419#titre-bloc-7 [14. 7. 2021].

Pareto Software, SimpleMaps.com (2010–2021). *World Cities Database*. Dostupné z: https://simplemaps.com/data/world-cities [9. 4. 2021].

Sandre (*Service d'administration nationale des données et des référentiels sur l'eau*, *2021*). Dostupnéz: http://www.sandre.eaufrance.fr/Rechercher-une-donnee-d-un-jeu [13. 7. 2021].

Worldmeters.info (2021). *Countries in the world by population.* Dover, Delaware, USA. Dostupné z: https://www.worldometers.info/world-population/population-by-country/ [31. 3. 2021].

Přílohy

Přílohou této práce je CD, na kterém jsou k dispozici zdrojové kódy k pokusům uvedeným v této práci.

Abstrakt

Název práce: Jazyková arbitrárnost pohledem kvantitativních metod

Autor práce: Bc. Klára Hájková

Vedoucí práce: Mgr. Vladimír Matlach, Ph.D.

Počet stran a znaků: 65 stran (129 000 znaků – zaokrouhlená hodnota)

Počet příloh: 1 (CD)

Abstrakt:

Cílem této práce je nahlédnout pomocí kvantitativních metod jazykovou arbitrárnost definovanou Ferdinandem de Saussurem. Na příkladech toponym a hydronym jsou v jednotlivých pokusech měřeny korelace mezi různými kvantifikovatelnými vlastnostmi objektů a kvantifikovatelnými vlastnostmi jejich názvů za účelem zjištění, zda jsou nebo nejsou tato jména založena na ikonicitě. Pomocí metody podpůrných vektorů (SVM) jsou trénovány modely strojového učení, které mají za cíl rozlišit mezi různými toponymy, příp. toponymy a hydronymy, aby tímto bylo prokázáno, zda se ve slovech jednotlivých tříd vyskytují, nebo nevyskytují určité vzory, na jejichž základě by byl model schopný úspěšné predikce. Statistickými metrikami jsou vyhodnocovány výsledky pokusů a vše je podrobeno rozsáhlé diskuzi s ohledem na další empirické lingvistické zákony, grafické i fonetické proměny jmen v průběhu historie a jiné.

Klíčová slova: arbitrárnost, ikonicita, toponyma, hydronyma, strojové učení, kvantitativní metody

Abstract

Title: Linguistic arbitrariness from the point of view of quantitative methods

Author: Bc. Klára Hájková

Supervisor: Mgr. Vladimír Matlach, Ph.D.

Number of pages and characters: 65 pages (129 000 characters – rounded value)

Number of appendices: 1 (CD)

Abstract:

The aim of this thesis is to look into the linguistic arbitrariness defined by Ferdinand de Saussure using quantitative methods. Using the examples of toponyms and hydronyms, correlations between various quantifiable properties of objects and quantifiable properties of their names are measured in individual experiments to determine whether or not these names are based on iconicity. Using the support vector machine (SVM) method, machine learning models are trained to distinguish between different toponyms, or toponyms and hydronyms, in order to demonstrate whether or not certain patterns are present in the words of each class, based on which the model would be able to make successful predictions. Statistical metrics are used to evaluate the results of the experiments and everything is subjected to extensive discussion with respect to other empirical linguistic laws, graphical and phonetic changes of names throughout history, and others.

Keywords: arbitrariness, ikonicity, toponyms, hydronyms, machine learning, quantitative methods