Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Model distribúcie interpunkčných znamienok v rôznych textoch

Diplomová práca

2026
Bc. Oliver Laštík

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Model distribúcie interpunkčných znamienok v rôznych textoch

Diplomová práca

Študijný program:	Aplikovaná Informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra informatiky
Školiteľ:	doc. RNDr. Mária Markošová, PhD.

Bratislava, 2026
Bc. Oliver Laštík

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Oliver Laštík
Študijný program: aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Model distribúcie interpunkčných znamienok v rôznych textoch
Model of punctuation mark distributions in various texts

Anotácia: Študent naprogramuje, alebo použije už naprogramovanú aplikáciu na hľadanie distribúcie interpunkčných znamienok v textoch v rôznych jazykoch. Ak študent použije hotovú aplikáciu, musí k nej doprogramovať v šetko potrebné na anlýzu distribúcie a na tvorbu jej modelu. Študent zanalyzuje distribú cie z textov toho istého autora v rôznych časových obdobiach aby zistil, ako sa táto vlastnosť textu mení a nakoľko ostáva stabilná. Takisto zanalyzuje tieto distribúcie v textoch ľudí s afáziou. Navrhne matematický model tejto distribúcie.

Cieľ: = naprogramovať aplikáciu, alebo doprogramovať potrebné nástroje do aplikácie na analýzu distribúcií interpunkčných znamienok v textoch a vytvoriť matematický model opisujúci získané výsledky

Literatúra: Kulig and others. In narrative texts punctuation marks obey the same statistics as words, Information Sciences
Volume 375, 1 January 2017, Pages 98-113

Vedúci: doc. RNDr. Mária Markošová, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: doc. RNDr. Tatiana Jajcayová, PhD.

Dátum zadania: 23.02.2024

Dátum schválenia: 25.09.2024 prof. RNDr. Roman Ďurikovič, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie

Abstrakt

Táto diplomová práca sa zaoberá analýzou distribúcie interpunkčných znamienok v textoch v rôznych jazykoch, s cieľom vytvoriť matematický model tejto distribúcie. Hlavným cieľom je preskúmať, ako sa interpunkčné znamienka rozmiestňujú v textoch rôznych autorov a v rôznych časových obdobiach, a ako tieto vzory môžu byť použité na zlepšenie textovej analýzy v oblasti spracovania prirodzeného jazyka (NLP). Práca sa zameriava na analýzu textov autorov z rôznych historických období, ako aj na porovnanie distribúcií v textoch vytvorených osobami trpiacimi afáziou.

V rámci výskumu bude vyvinutý nástroj na analýzu textov, ktorý bude schopný identifikovať a spracovávať interpunkčné znamienka v rôznych jazykoch. Na základe týchto údajov bude následne vytvorený matematický model, ktorý bude schopný predikovať distribúciu interpunkčných znamienok v textoch. Model bude validovaný pomocou testovacích dát a porovnaný s existujúcimi prístupmi v literatúre. Výsledky tejto práce môžu prispieť k lepšiemu pochopeniu textovej štruktúry a poskytnúť nástroje na zlepšenie analýzy textov v oblasti automatického generovania textov, literárnej analýzy a spracovania prirodzeného jazyka.

Kľúčové slová:

Abstract

Keywords:

Obsah

Úvod

...

Kapitola 1

Súčasný stav riešenej problematiky

Skúmanie štatistických vlastností prirodzeného jazyka má dlhú tradíciu v kvantitatívnej lingvistike, informatike aj fyzike komplexných systémov. Klasické práce sa sústreďujú najmä na správanie slov: ich frekvenčné rozdelenia, dĺžky, n–gramy, jazykové modely a rôzne mierky „komplexity“ textu. Tieto prístupy sa využívajú pri automatickom rozpoznávaní autora, pri klasifikácii žánrov, v informačnom vyhľadávaní aj v moderných jazykových modeloch.

V takto zameraných prácach je interpunkcia často chápaná iba ako technický prostriedok na členenie textu, ktorý sa pri predspracovaní dokonca odstraňuje. V posledných rokoch sa však objavujú štúdie, ktoré ukazujú, že interpunkčné znamienka nesú štatisticky relevantnú informáciu a sú citlivé na jazyk, žáner a štýl autora [5]. Interpunkcia tak prestáva byť „šumom“ a stáva sa samostatným objektom kvantitatívnej analýzy. Na túto líniu výskumu nadväzuje aj táto práca, ktorá sa zameriava práve na distribúciu a prechodové vzťahy interpunkčných znamienok v textoch.

Táto kapitola stručne mapuje tri hlavné oblasti, o ktoré sa práca opiera: kvantitatívnu analýzu textov, sieťové modelovanie jazyka a informačno–teoretické miery vzdialenosti pravdepodobnostných rozdelení. Zároveň naznačuje, ako budú tieto východiská využité v ďalších častiach práce: pri konštrukcii vlastného korpusu, modelovaní interpunkcie Markovovskými reťazcami a pri porovnávaní jednotlivých jazykov.

1.1 Kvantitatívna analýza textov a interpunkcia

Kvantitatívne prístupy k jazyku sledujú napríklad rozdelenie frekvencií slov, vzťah medzi frekvenciou a poradím (Zipfov zákon), entropiu textu či vlastnosti jazykových n–gramov. Tieto metódy sa používajú pri stylometrii (identifikácia autora), pri klasifikácii žánrov, v strojovom preklade a pri budovaní jazykových modelov založených na pravdepodobnosti.

V mnohých aplikáciách sa interpunkcia buď úplne odstraňuje, alebo sa zjednodušuje

na jeden symbol. Práca [5] explicitne ukazuje, že takýto prístup môže byť škodlivý, ak nás zaujíma jemnejšia štruktúra textu. Autori demonštrujú, že interpunkčné znamienka v naratívnych textoch:

- podliehajú podobným škálovacím zákonom ako slová,
- majú stabilné frekvenčné vzory v rámci autora a žánru,
- možno ich využiť pri úlohách autorstva alebo klasifikácie textov.

Z pohľadu tejto diplomovej práce je dôležité, že:

- interpunkcia je voči prekladu stabilnejšia než konkrétne slová,
- jej použitie je ovplyvnené jazykovými normami (gramatikou) aj individuálnym štýlom,
- intenzita a typ interpunkcie súvisia s členitosťou viet, rytmom textu a prítomnosťou dialógov.

V ďalších kapitolách preto budeme interpunkciu analyzovať samostatne: najskôr na úrovni základných frekvencií (počty znamienok na jednotku textu) a následne aj na úrovni prechodových vzťahov, ktoré budeme modelovať Markovovskými reťazcami. Takýto prístup umožní porovnávať jazyky a knihy nielen podľa toho „koľko“ interpunkcie obsahujú, ale aj podľa toho, „ako“ sa jednotlivé znamienka v textoch striedajú.

1.2 Textové korpusy a Project Gutenberg

Pre kvantitatívnu lingvistiku sú kľúčové rozsiahle a dobre zdokumentované textové korpusy. Project Gutenberg predstavuje jeden z najväčších verejne dostupných archívov literárnych diel. Obsahuje tisíce kníh rôznych autorov, jazykov a žánrov voľne dostupných v textovom formáte, čo z neho robí prirodzený zdroj dát pre štatistickú analýzu.

Pri priamom použití týchto textov však vzniká niekoľko praktických problémov:

- texty obsahujú *boilerplate* — úvodné a záverečné technické sekcie, licenčné poznámky a iné časti, ktoré nepatria k samotnému dielu,
- metadáta (autor, jazyk, rok vydania, žáner) nie sú úplne jednotné a môžu byť čiastočne chybné alebo neúplné,
- jednotlivé vydania toho istého diela môžu mať odlišné technické úpravy, ktoré ovplyvňujú štatistiku interpunkcie.

Gerlach a kolegovia [3] preto navrhujú štandardizovaný Gutenberg korpus, v ktorom systematicky:

1. čistia texty od boilerplate sekcií,
2. zjednocujú a opravujú metadáta,
3. definujú presné kritériá, ktoré diela sú zahrnuté do konečného korpusu.

Ich prístup zdôrazňuje dve vlastnosti dôležité aj pre túto prácu: **transparentnosť výberu** (jasné kritériá inclusion/exclusion) a **reprodukovateľnosť** (výber aj čistenie prebieha pomocou skriptov, nie manuálne).

V tejto diplomovej práci je Project Gutenberg primárnym zdrojom textov pre tri germánske jazyky: nemčinu, angličtinu a holandčinu. Pri výbere diel sa inšpirujeme filozofiou štandardizovaného korpusu [3], no kritériá prispôsobujeme nášmu cieľu: porovnať distribúciu interpunkcie v literárnych textoch z historicky porovnateľného obdobia. V ďalšej kapitole (Metodika) preto podrobne opíšeme:

- výberové kritériá (jazyk, obdobie, žáner, maximálny počet diel na autora),
- spôsob čistenia textu a odstránenia boilerplate častí,
- rozdelenie textov na segmenty rovnakej dĺžky, aby bolo porovnanie jazykov férové.

1.3 Sieťové modelovanie jazyka

Jazyk možno modelovať aj pomocou abstraktných sietí, v ktorých vrcholy predstavujú jazykové jednotky (slová, morfémy, vety, koncepty) a hrany ich vzťahy (susednosť, syntaktické väzby, sémantickú podobnosť). Takto vzniknuté jazykové siete často vykazujú typické vlastnosti komplexných sietí: škálovacie rozdelenie stupňov, malú priemernú vzdialenosť, vysokú zhlukovitosť či komunitnú štruktúru [2, 6].

V literatúre existuje viacero prístupov k sietiam na texte:

- slovné siete založené na susednosti slov v texte (slová sú prepojené, ak sa vyskytujú vedľa seba),
- syntaktické siete prevzaté zo syntaktických stromov,
- sémantické siete, kde hrany reprezentujú asociácie alebo spoluvýskyt konceptov.

Sieťový pohľad sa ukázal užitočný aj pri skúmaní porúch jazyka a kognitívnych porúch. Práca [4] demonštruje rozdiely v štruktúre sietí získaných z textov nesúcich znaky afázie oproti typickým textom. Topológia jazykovej siete tak môže odrážať kognitívne procesy autora a mieru „porušenia“ jazykového systému.

V tejto práci nebudujeme plnohodnotnú veľkú slovnú sieť, ale aplikujeme podobnú myšlienku na interpunkciu. Prechodovú maticu interpunkčných znamienok možno chápať ako orientovanú sieť, v ktorej:

- uzly predstavujú typy interpunkčných znamienok (bodka, čiarka, bodkočiarka, dvojbodka, otáznik, výkričník, ...),
- hrany reprezentujú typické prechody medzi znamienkami (napr. výskyt otáznika po čiarke).

V ďalších častiach práce budeme túto „sieť interpunkcie“ kvantifikovať pomocou Markovovských modelov, pričom priemerná prechodová matica pre jazyk bude slúžiť ako jeho typický „interpunkčný profil“. Neskôr bude možné tieto siete rozšíriť o ďalšie sieťové ukazovatele, napríklad o rozdelenie stupňov, zhlukovitosť či komunitnú štruktúru znamienok.

1.4 Informačno–teoretické miery a Markovovské modely

Pri porovnávaní jazykových modelov je potrebné zvoliť vhodnú mieru vzdialenosti medzi pravdepodobnostnými rozdeleniami. Informačná teória ponúka prirodzené nástroje v podobe entropie, Kullbackovej–Leiblerovej divergencie a z nej odvodených symetrických mier, akou je napríklad Jensen–Shannonova divergencia [1]. Tieto miery majú jasnú interpretáciu „množstva informácie“ potrebnej na rozlíšenie dvoch rozdelení, čo ich robí vhodnými na porovnávanie jazykových profilov.

Markovovské modely prvého rádu sú štandardným nástrojom pri modelovaní sekvencií symbolov: predpokladajú, že pravdepodobnosť nasledujúceho symbolu závisí iba od aktuálneho stavu. V kontexte textu to znamená, že pravdepodobnosť ďalšieho interpunkčného znamienka závisí od predchádzajúceho znamienka (prípadne špeciálneho „stavu medzi vetami“). Takýto model je dostatočne jednoduchý na to, aby bol dobre interpretovateľný, a zároveň dostatočne bohatý na zachytenie typických vzorov nadväzovania znamienok.

V tejto práci budeme postupovať nasledovne:

- **Konštrukcia a predspracovanie korpusu**: z Project Gutenberg vyberieme literárne texty v nemčine, angličtine a holandčine, podľa jasne definovaných kritérií (jazyk, obdobie, žáner, maximálny počet diel na autora). Texty očistíme od boilerplate častí a rozdelíme ich na segmenty rovnakej dĺžky, aby bolo porovnanie jazykov férové a reprodukovateľné.

- **Základná distribúcia interpunkcie**: pre každý segment vypočítame frekvencie vybraných interpunkčných znamienok, normalizované na pevnú jednotku (napr. na 1000 slov), a tieto štatistiky následne agregujeme na úroveň knihy a jazyka. Získame tak základný obraz o tom, „koľko“ interpunkcie jednotlivé jazyky a knihy používajú a aká je variabilita medzi dielami.

- **Markovovské modely interpunkcie**: pre každú knihu skonštruujeme prechodovú maticu interpunkčných znamienok a z nich odvodíme priemerné jazykové Markovovské modely. Pomocou informačno–teoretických mier (najmä Jensen–Shannonovej divergencie) budeme:

  1. porovnávať priemerné modely jednotlivých jazykov medzi sebou,
  2. merať vzdialenosť jednotlivých kníh od jazykového priemeru a analyzovať stabilitu a odľahlé prípady.

- **Rank–frequency analýza slov a interpunkcie**: nadviažeme na výsledky [5] a pre vybrané texty vytvoríme rank–frequency krivky osobitne pre:

  1. slovné tokeny,
  2. kombinované slovné a interpunkčné tokeny.

  Na tieto krivky budeme fitovať modely typu Zipf a Zipf–Mandelbrot a porovnáme odhadnuté parametre medzi jazykmi a medzi dvoma režimami (iba slová vs. slová + interpunkcia). Cieľom je ukázať, do akej miery sa interpunkcia správa „ako bežné tokeny“ v zmysle rank–frequency zákonitostí.

- **Siete založené na susednosti tokenov**: z tokenizovaných textov skonštruujeme orientované siete, v ktorých uzly reprezentujú slová a interpunkčné znamienka a hrany reprezentujú ich susednosť v texte. Pre tieto siete vypočítame vybrané sieťové metriky (napr. rozdelenie stupňov, priemernú dĺžku najkratšej cesty, zhlukovitosť) a porovnáme ich medzi jazykmi. Zameriame sa aj na vzťah medzi frekvenciou tokenu a jeho stupňom v sieti, v duchu známych výsledkov o komplexných sieťach [2, 6].

- **Porovnanie s generatívnym modelom rastu sietí**: ako referenčný model použijeme jednoduchý generatívny model rastu sietí (napríklad Barabási–Albertov model preferenčného pripájania). Pre vybrané reálne siete vytvorené z textov vygenerujeme umelé siete s rovnakým počtom uzlov a približným počtom hrán a porovnáme vybrané štatistiky (rozdelenie stupňov, zhlukovitosť, priemernú dĺžku najkratšej cesty). Cieľom je zistiť, v čom sa reálne „jazykové“ siete podobajú jednoduchým modelom rastu a v čom sa od nich odlišujú.

Takto navrhnutý postup prepája frekvenčné štatistiky, Markovovské modely, sieťovú analýzu a generatívne modely sietí do jedného rámca. Umožní nielen kvantitatívne porovnať tri germánske jazyky na úrovni interpunkcie, ale aj diskutovať, nakoľko sú pozorované štruktúry špecifické pre jazyk a text a nakoľko sa dajú vysvetliť jednoduchými modelmi rastu komplexných sietí.

1.5 Zhrnutie kapitoly

Súčasný stav výskumu ukazuje, že:

- interpunkcia nesie štatisticky významnú informáciu a jej ignorovanie môže viesť k strate dôležitých štrukturálnych znakov textu [5],
- textové korpusy z Project Gutenberg je možné systematicky vyčistiť a spracovať pre kvantitatívnu analýzu tak, aby bol výber diel transparentný a reprodukovateľný [3],
- jazykové štruktúry možno úspešne modelovať ako siete a Markovovské reťazce [2, 6],
- informačná teória poskytuje prirodzené miery podobnosti pravdepodobnostných modelov [1],
- sieťový pohľad na jazyk je citlivý aj na jemné zmeny v štruktúre textu a môže byť využitý aj pri analýze jazykových porúch [4].

Na týchto východiskách je postavená metodika práce. V nasledujúcej kapitole podrobne opíšeme, ako konštruujeme vlastný korpus z Project Gutenberg, akými krokmi predspracovania prechádza a ako z neho získavame frekvenčné štatistiky a Markovovské modely interpunkcie, ktoré budú následne analyzované a porovnávané medzi jednotlivými jazykmi.

Kapitola 2

Cieľ práce

Kapitola 3

Metodika práce a metódy skúmania

Kapitola 4

Metodika práce a metódy skúmania

Cieľom tejto kapitoly je podrobne opísať, akým spôsobom sú pripravené dáta, aké veličiny meriame a ako z nich konštruujeme modely interpunkcie. Dôraz kladieme na reprodukovateľnosť: jednotlivé kroky sú realizované pomocou skriptov, verzovaných v repozitári, a ich výstupy sú systematicky ukladané do adresárovej štruktúry.

4.1 Prehľad spracovateľskej pipeline

Celý proces spracovania textov možno schematicky rozdeliť do týchto krokov:

1. výber vhodných diel z katalógu Project Gutenberg,
2. stiahnutie textov a ich čistenie od boilerplate častí,
3. rozdelenie očistených textov na okná s pevnou dĺžkou,
4. výpočet základných štatistík interpunkcie pre každé okno,
5. konštrukcia Markovovských prechodových matíc,
6. výpočet mier podobnosti medzi jazykmi a jednotlivými knihami.

Jednotlivé kroky sú implementované v jazyku Python, s využitím knižníc `pandas`, `numpy` a `matplotlib` na prácu s tabuľkami a vizualizáciu. Všetky medzivýsledky (filtrovaný katalóg, logy, štatistiky a matice) sú ukladané do samostatných súborov, aby bolo možné jednotlivé časti analýzy opätovne použiť bez nutnosti prechádzať celý proces od začiatku.

4.2 Výber textov z Project Gutenberg

Vstupom je CSV súbor s metadátami Project Gutenberg, ktorý obsahuje informácie o jazyku textu, autorovi, roku vydania, žánri a identifikátore knihy. Na základe týchto údajov sa vykoná filtrácia podľa nasledujúcich kritérií:

- jazyk je označený ako nemčina (`de`), angličtina (`en`) alebo holandčina (`nl`),
- rok vydania spadá do zvoleného historicky porovnateľného obdobia (19. až začiatok 20. storočia),
- žáner zodpovedá umeleckej próze (romány, poviedky),
- od jedného autora sú vybrané maximálne tri diela v danom jazyku.

Takto vznikne kandidátny zoznam diel, ktorý je uložený v súbore `accepted.csv`. Tento súbor slúži ako referenčný zoznam korpusu a je sprevádzaný log súborom, ktorý obsahuje informácie o tom, prečo boli niektoré položky z katalógu vyradené (nesprávny jazyk, chýbajúci text a pod.). Tým sa zabezpečuje transparentnosť výberu, v duchu odporúčaní [3].

4.3 Sťahovanie textov a čistenie boilerplate

Pre každé dielo zo zoznamu `accepted.csv` sa stiahne z Project Gutenberg jeho textová verzia. Následne sa na ňom vykoná čistiaca procedúra, inšpirovaná prístupom [3], prispôsobená konkrétnym formálnym znakom použitých verzií textov:

1. identifikácia a odstránenie úvodných a záverečných hlavičiek (poznámky projektu, licencie, technické informácie),
2. normalizácia kódovania a nahradenie netlačiteľných znakov,
3. zjednotenie typografických variantov interpunkčných znamienok (napr. rôzne uvozovky),
4. odstránenie pasáží, ktoré nepredstavujú samotný literárny text (poznámky pre sadzača, reprintové komentáre a pod.).

Výsledkom je „očistený“ text, ktorý obsahuje iba samotné dielo. Tieto texty sú uložené v adresári `data_core/clean/` a predstavujú východisko pre ďalšie kroky.

4.4 Rozdelenie na okná fixnej dĺžky

Knihy majú prirodzene veľmi rôzne dĺžky. Ak by sme ich porovnávali ako celky, výsledné štatistiky by boli ovplyvnené tým, či ide o krátke poviedky alebo rozsiahle romány. Preto pracujeme s jednotkami pevnej dĺžky: pre každý očistený text:

- text tokenizujeme na slová pomocou jednoduchej tokenizácie založenej na bielych znakoch a interpunkcii,
- vytvoríme po sebe idúce segmenty s dĺžkou 30 000 slov,
- posledný segment, ktorý nedosahuje zvolenú dĺžku, sa do analýzy nezaraďuje.

Každý segment reprezentuje „okno“ textu s porovnateľnou dĺžkou pre všetky knihy a jazyky. Tieto okná sú uložené v adresári `data_core/windows30k/`. Pri ďalšej analýze pracujeme práve s týmito oknami, nie s celými knihami naraz.

4.5 Definícia a počítanie interpunkčných znakov

Pre potreby analýzy je potrebné rozhodnúť, ktoré znaky budeme považovať za interpunkciu a ako budeme zaobchádzať s hraničnými prípadmi (dvojbodky, pomlčky, uvozovky rôzneho typu, viacnásobné výskyty). Vychádzame z nasledujúcej množiny typov:

- bodka, otáznik, výkričník,
- čiarka, bodkočiarka, dvojbodka,
- párové uvozovky (normalizované na jednotný tvar),
- zátvorky,
- pomlčka oddelená medzerami.

Počas spracovania každej knihy sa vytvára interná reprezentácia, v ktorej sú slová a interpunkčné znaky oddelené ako samostatné tokeny. Následne pre každé okno počítame:

- absolútny počet výskytov jednotlivých typov interpunkcie,
- počet slov v okne (bez interpunkcie),
- pre každé znamienko normalizovanú hodnotu „počet na 1000 slov“.

Výsledky sú zapísané do tabuľky `punct_stats_core.csv`, kde každý riadok zodpovedá jednému oknu, identifikovanému jazykom, Gutenberg ID a poradím segmentu v rámci diela.

4.6 Agregácia štatistík po jazykoch

Zo základných štatistík jednotlivých okien vytvárame agregované charakteristiky po jazykoch. Pre každý jazyk a každý typ interpunkcie počítame:

- priemernú normalizovanú frekvenciu (na 1000 slov),
- medián a vybrané percentily rozdelenia,
- odhad intervalov spoľahlivosti napríklad pomocou bootstrapu nad oknami (opätovné náhodné vzorkovanie okien s návratom).

Agregované údaje ukladáme do súboru `punct_summary_by_language.csv`. Na ich základe sa kreslia stĺpcové grafy a boxploty porovnávajúce používanie interpunkcie medzi jazykmi.

4.7 Konštrukcia Markovovských prechodových matíc

Ak chceme zachytiť *spôsob*, akým sú interpunkčné znaky v texte usporiadané, nestačí sledovať iba počty výskytov. Potrebujeme model, ktorý zohľadní aj ich suslednosť. Použijeme Markovovský reťazec prvého rádu, kde stavmi sú typy interpunkcie.

Nech S je množina typov znamienok. Pre každé okno prechádzame text a pri každom výskyte dvojice po sebe idúcich znamienok (i, j) zvyšujeme počítadlo C_{ij}. Po prechode celým oknom normalizujeme tieto počty tak, aby sme získali prechodovú maticu

$$P_{ij} = \frac{C_{ij} + \alpha}{\sum_{k \in S}(C_{ik} + \alpha)},$$

kde α je malý pseudopočet (napríklad $\alpha = 10^{-6}$), zabezpečujúci, že nevznikajú nulové pravdepodobnosti. Takto definovaná matica P predstavuje Markovovský model prechodov medzi interpunkčnými typmi v danom okne.

Pre každé okno ukladáme vektorovo rozvinutú podobu matice do súboru `punct_transitions_core.csv`. Z týchto matíc následne odvodzujeme jazykové priemery aj vzdialenosti medzi knihami.

4.8 Jensen–Shannonova divergencia ako miera vzdialenosti

Na porovnanie dvoch Markovovských modelov potrebujeme zmysluplnú mieru vzdialenosti medzi prechodovými maticami. Vychádzame z informačnej teórie [1] a používame Jensen–Shannonovu divergenciu (JSD), ktorá je symetrická a vždy konečná.

Pre dve diskrétne rozdelenia P a Q definujeme

$$\mathrm{JSD}(P, Q) = \frac{1}{2} D_{\mathrm{KL}}(P \,\|\, M) + \frac{1}{2} D_{\mathrm{KL}}(Q \,\|\, M)\,,$$

kde $M = \frac{1}{2}(P + Q)$ a D_{KL} je Kullbackova–Leiblerova divergencia. V našom prípade aplikujeme JSD na *riadky* prechodových matíc: pre každý typ znamienka i porovnávame rozdelenia následníkov $P_{i\cdot}$ a $Q_{i\cdot}$ a vypočítame $\mathrm{JSD}(P_{i\cdot}, Q_{i\cdot})$. Tieto riadkové hodnoty potom agregujeme priemerom (neváženým alebo váženým podľa výskytu daného typu znamienka).

Takto získavame:

- párové vzdialenosti medzi priemernými jazykovými prechodovými maticami,
- vzdialenosti jednotlivých kníh od priemerného modelu daného jazyka.

Výsledky agregujeme do tabuliek `language_distance_jsd.csv`, `rowwise_language_distance` a `per_book_distance_to_language_markov.csv`.

4.9 Odhad intervalov spoľahlivosti a stabilita jazykových modelov

Keďže pracujeme s konečným počtom kníh, odhadnuté jazykové prechodové matice sú zaťažené štatistickou neistotou. Preto pre vybrané veličiny odhadujeme aj intervaly spoľahlivosti. Prakticky postupujeme tak, že:

- pre každý jazyk opakovane náhodne vzorkujeme knihy s návratom (bootstrap),
- pre každú vzorku vypočítame príslušné metriky (napr. medián vzdialeností kníh od jazykového priemeru),
- z empirického rozdelenia bootstrapových hodnôt odčítame 2,5 % a 97,5 % percentil ako odhad 95 % intervalu spoľahlivosti.

Takto získame napríklad súhrnnú tabuľku pre každý jazyk s údajmi o mediáne, minimálnej a maximálnej vzdialenosti knihy od jazykového modelu a odhadovanom 95 % intervale spoľahlivosti mediánu.

4.10 Zhrnutie kapitoly

Metodika práce kombinuje:

- systematický výber a čistenie textov z Project Gutenberg [3],

- normalizáciu dĺžky textov pomocou okien s pevnou dĺžkou,
- jednoduché, ale dobre interpretovateľné štatistiky interpunkcie (počty na 1000 slov),
- Markovovský model prechodov medzi interpunkčnými typmi,
- informačno–teoretickú mieru podobnosti modelov [1].

V nasledujúcej kapitole si ukážeme, aké konkrétne výsledky tieto metódy prinášajú pri porovnaní troch germánskych jazykov.

Kapitola 5

Návrh softvérového diela

Kapitola 6

Výskum a výsledky práce

V tejto kapitole prezentujeme výsledky získané na korpuse troch germánskych jazykov: nemčiny, angličtiny a holandčiny. Najskôr charakterizujeme samotný korpus, potom sa zameriame na základné frekvenčné štatistiky interpunkcie a napokon na výsledky Markovovských modelov a informačno–teoretických mier vzdialenosti.

6.1 Charakteristika korpusu

Finálny korpus obsahuje:

- 80 kníh v nemčine (de),
- 80 kníh v angličtine (en),
- 80 kníh v holandčine (nl).

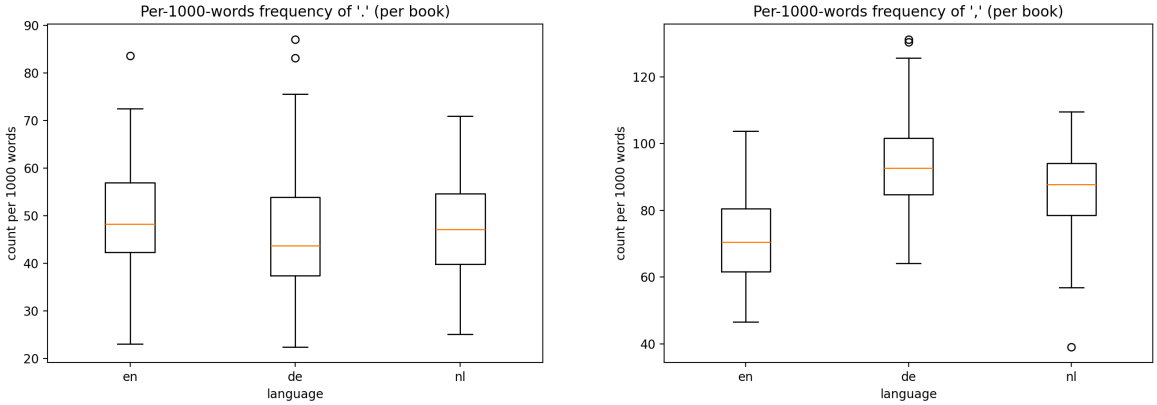
Pre každú knihu bol text vyčistený od boilerplate častí a následne rozrezaný na segmenty s dĺžkou 30 000 slov. Každý segment predstavuje samostatné „okno“ pre analýzu. Počet okien na knihu závisí od dĺžky diela, no v priemere ide o niekoľko segmentov, čo poskytuje dostatočnú vzorku na odhad jazykových štatistík.

To, že každé okno má rovnakú dĺžku, zabezpečuje férové porovnanie medzi jazykmi a knihami a znižuje vplyv extrémne dlhých alebo krátkych diel. Zároveň nám rozloženie na okná umožňuje odhadnúť variabilitu v rámci jednej knihy.

6.2 Základné štatistiky interpunkčných znamienok

Z tabuľky `punct_stats_core.csv` sme pre každý jazyk vypočítali normalizované frekvencie vybraných interpunkčných znamienok (na 1000 slov) na úrovni jednotlivých okien. V tejto podsekcii prezentujeme rozdelenia týchto frekvencií pomocou boxplotov a interpretujeme hlavné pozorované trendy.

Najčastejšie znaky z hľadiska štruktúry viet sú bodka a čiarka. Na obrázku 6.1 sú zobrazené boxploty normalizovaných frekvencií bodiek a čiarok (na 1000 slov) pre všetky tri jazyky.

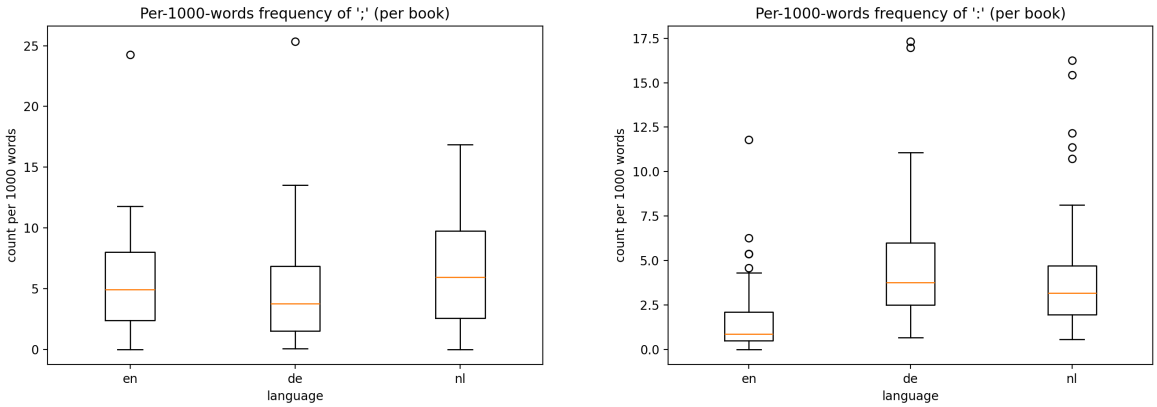


(a) Bodka (.) (b) Čiarka (,)

Obr. 6.1: Normalizovaná frekvencia bodiek a čiarok (na 1000 slov) pre angličtinu, nemčinu a holandčinu. Boxploty zobrazujú rozdelenie naprieč všetkými oknami v korpuse (medián, kvartily a extrémne hodnoty).

Z obrázka 6.1 vidno, že bodky aj čiarky patria medzi najčastejšie interpunkčné znamienka vo všetkých sledovaných jazykoch. Mediánové hodnoty sú relatívne stabilné, no čiarka vykazuje väčšiu variabilitu medzi jednotlivými knihami – niektoré texty používajú výrazne viac vložených viet a súvetí, čo sa prejavuje vyššími hodnotami v hornej časti boxplotov.

Ďalej sa zameriavame na menej časté, ale štýlovo zaujímavé znamienka: bodkočiarku a dvojbodku. Ich frekvencie sú zobrazené na obrázku 6.2.

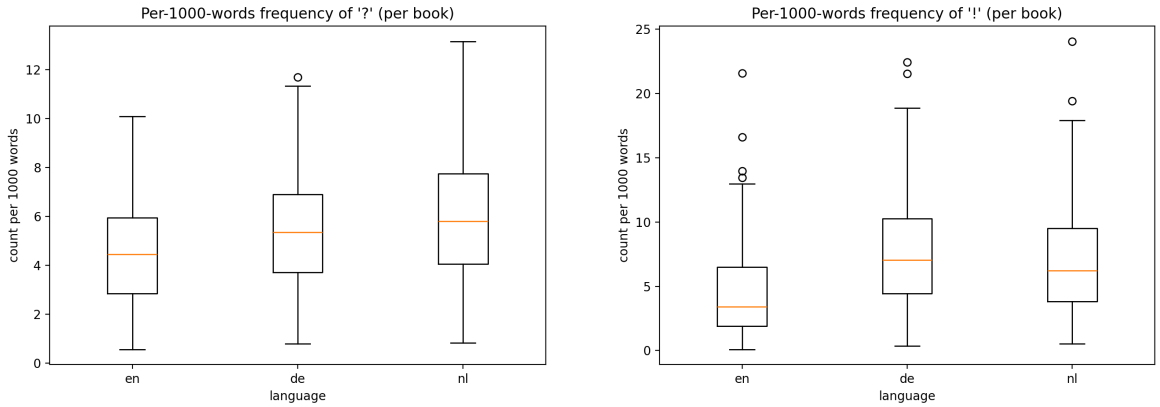


(a) Bodkočiarka (;) (b) Dvojbodka (:)

Obr. 6.2: Normalizovaná frekvencia bodkočiarok a dvojbodiek (na 1000 slov) naprieč anglickými, nemeckými a holandskými textami.

Obrázok 6.2 ukazuje, že bodkočiarka aj dvojbodka sú v porovnaní s bodkou a čiarkou používané podstatne menej často. Zároveň medzi knihami vidíme veľké rozdiely: v niektorých textoch sa bodkočiarka takmer nevyskytuje, inde má vyššiu frekvenciu a signalizuje preferenciu dlhších, syntakticky zložitejších viet. Dvojbodka sa častejšie objavuje v dielach, kde autor často uvádza priame reči, vysvetlenia alebo zoznamy.

Napokon skúmame aj frekvencie otáznikov a výkričníkov, ktoré súvisia skôr so sémantikou a náladou textu (otázky, emocionálne zafarbené pasáže). Ich rozdelenia sú na obrázku 6.3.



(a) Otáznik (?)

(b) Výkričník (!)

Obr. 6.3: Normalizovaná frekvencia otáznikov a výkričníkov (na 1000 slov) v angličtine, nemčine a holandčine.

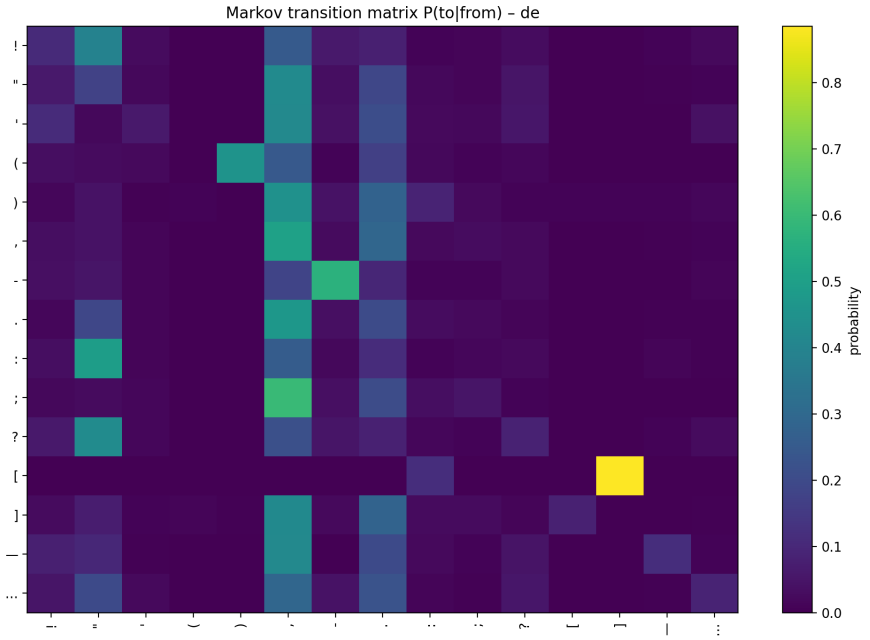
V oboch prípadoch sú mediány nízke a rozdelenia sú výrazne vychýlené smerom k nule. Väčšina textov používa otázniky aj výkričníky len sporadicky, zatiaľ čo niekoľko diel s vysokým podielom dialógov alebo expresívneho štýlu sa prejavuje ako body v hornej časti rozdelenia. Takéto diela môžeme vnímať ako štýlové outliery v rámci daného jazyka. Takéto základné štatistiky sú v zhode s pozorovaniami napr. v [5], kde interpunkcia vykazuje jazykovo špecifické, ale štruktúrne podobné vzory.

6.3 Markovovské modely interpunkcie pre jednotlivé jazyky

Základné frekvencie ukazujú, *koľko* interpunkčných znakov sa v texte vyskytuje, ale neodhaľujú, *ako* na seba jednotlivé znaky nadväzujú. Na zachytenie tejto informácie používame Markovovský model prvého rádu, v ktorom stavy zodpovedajú typom interpunkčných znakov a hrany predstavujú prechody medzi nimi.

Pre každú knihu a jazyk sme zostavili prechodovú maticu $P(\text{to} \mid \text{from})$, ktorá udáva pravdepodobnosť, že po znaku typu `from` nasleduje (v definovanom okolí) znak typu

`to`. Následne sme tieto matice spriemerovali za jednotlivé jazyky. Výsledné priemerné Markovovské matice zobrazujú obrázky 6.4– 6.6.

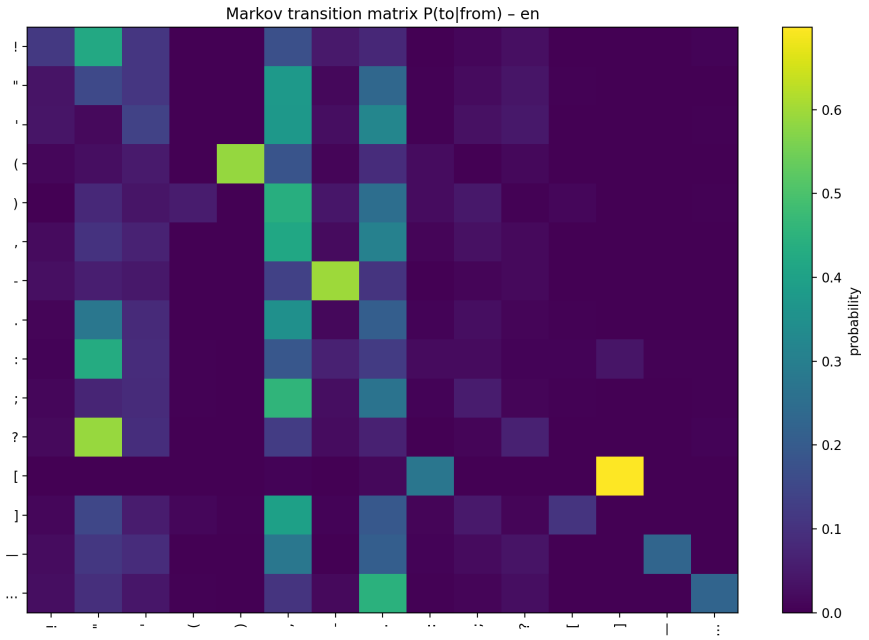


Obr. 6.4: Priemerná Markovovská prechodová matica interpunkčných znakov pre nemčinu (de). Riadky zodpovedajú východzím stavom (`from`), stĺpce cieľovým stavom (`to`); intenzita farby reprezentuje veľkosť prechodovej pravdepodobnosti.

Na obrázku 6.4 vidno výrazný diagonálny pás, ktorý zodpovedá prechodom „znak na seba“ alebo prechodom do stavu reprezentujúceho úsek bez interpunkcie. Silné prechody spojené s bodkou ilustrujú typické ukončenie vety a následné pokračovanie textu bez ďalšej interpunkcie v najbližších tokenoch.

Anglická prechodová matica (obr. 6.5) má podobnú globálnu štruktúru ako nemecká: dominujú prechody spojené s bodkou a čiarkou, výrazný je tiež prechod do stavu „bez interpunkcie“. Jemné rozdiely v intenzite jednotlivých políčok (napríklad pri bodkočiarke alebo dvojbodke) však naznačujú odlišné syntaktické zvyklosti a preferované konštrukcie v texte.

Holandská matica na obrázku 6.6 opäť vykazuje podobný základný vzor, no rozloženie pravdepodobností v jednotlivých riadkoch a stĺpcoch sa mierne líši. Aj keď sú rozdiely na úrovni jedného konkrétneho prechodu malé, v súhrne vedú k charakteristickej „podpísanej“ štruktúre pre každý jazyk. Tieto priemerné matice možno zároveň chápať ako orientované siete [2, 6], v ktorých hrany reprezentujú pravdepodobné prechody medzi znamienkami.



Obr. 6.5: Priemerná Markovovská prechodová matica interpunkčných znakov pre angličtinu (en).

6.4 Vzdialenosti medzi jazykmi

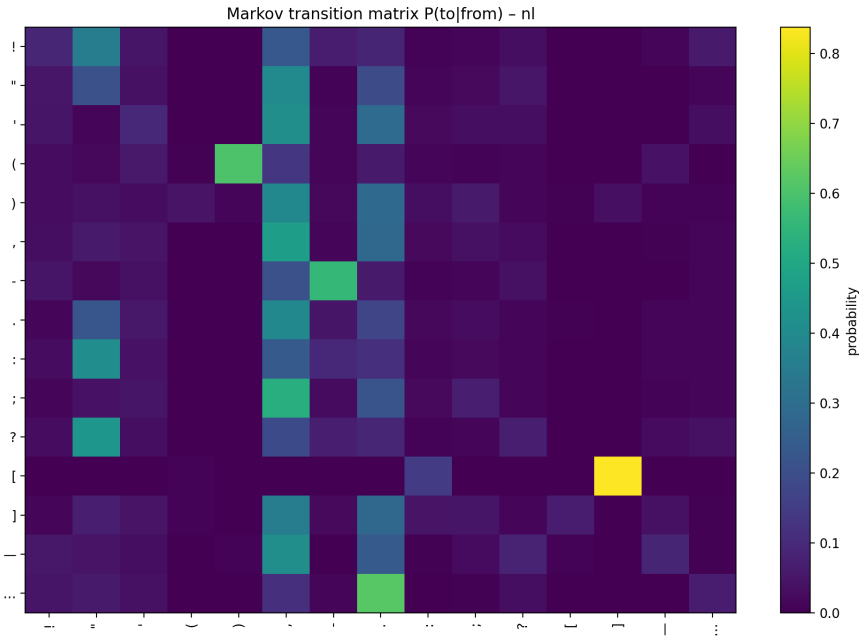
Na kvantitatívne porovnanie jazykových Markovovských modelov používame Jensen–Shannonovu divergenciu medzi priemernými prechodovými maticami. Pre každý pár jazykov počítame tzv. riadkovú Jensen–Shannonovu divergenciu (JSD) medzi príslušnými riadkami prechodových matíc a následne berieme vážený priemer cez všetky riadky. Takto získame jednu číselnú hodnotu, ktorá vyjadruje rozdielnosť Markovovských modelov interpunkcie daných dvoch jazykov. Výsledné párové vzdialenosti sú uložené v tabuľke `rowwise_language_distances.csv`.

Pre váženú priemernú riadkovú JSD dostávame približne:

- vzdialenosť medzi nemčinou a angličtinou $\approx 0{,}040$,
- vzdialenosť medzi nemčinou a holandčinou $\approx 0{,}017$,
- vzdialenosť medzi angličtinou a holandčinou $\approx 0{,}022$.

Hodnoty sú symetrické (vzdialenosť de–en je rovnaká ako en–de) a diagonálne prvky by sme definovali ako 0, keďže porovnávame jazyk s ním samým. Absolútna veľkosť týchto čísiel je relatívne malá (všetky vzdialenosti sú $\ll 0{,}1$), čo znamená, že aj najodlišnejší pár jazykov v rámci nášho korpusu má veľmi podobnú štruktúru prechodov interpunkčných znamienok.

Z uvedených hodnôt vyplýva, že:



Obr. 6.6: Priemerná Markovovská prechodová matica interpunkčných znakov pre holandčinu (nl).

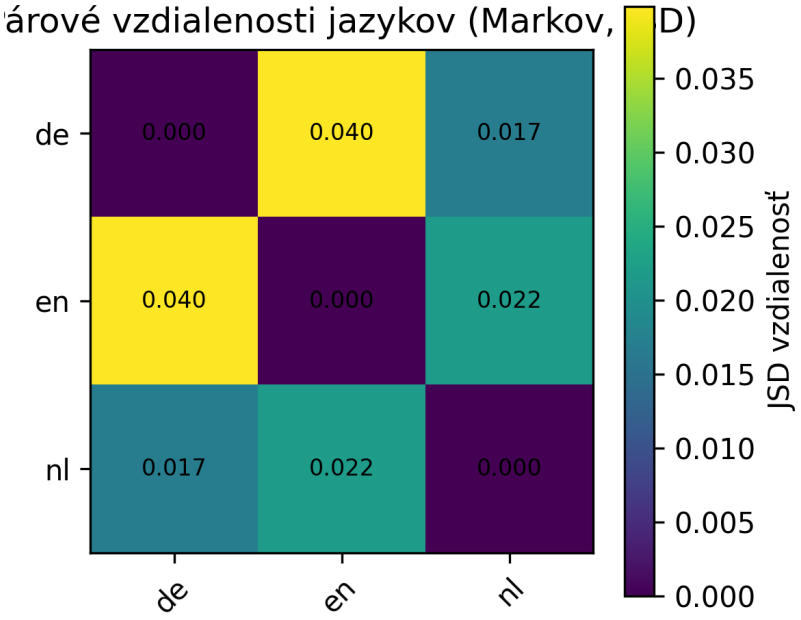
- nemčina a holandčina sú z hľadiska Markovovského modelu interpunkcie najbližšie (najmenšia JSD $\approx 0{,}017$),

- angličtina je od oboch kontinentálnych jazykov mierne vzdialenejšia (vzdialenosť $\approx 0{,}040$ voči nemčine a $\approx 0{,}022$ voči holandčine),

- celkové vzdialenosti sú však pomerne malé, čo naznačuje, že základná štruktúra prechodov interpunkčných znamienok je v rámci tejto germánskej jazykovej rodiny veľmi podobná.

Tepelná mapa na obr. 6.7 tieto hodnoty vizualizuje v podobe 3×3 matice. Diagonála obsahuje nulové vzdialenosti (jazyk porovnaný sám so sebou) a je zobrazená najbledšou farbou. Mimo diagonály vidíme tri relevantné prvky: dvojicu de–nl s najtmavším odtieňom (najmenšia vzdialenosť), dvojicu en–nl s mierne vyššou hodnotou a napokon dvojicu de–en, ktorá má najvyššiu JSD a je vizuálne oddelená výraznejším kontrastom.

Vizuálne tak heatmapa potvrdzuje číselné zistenia: nemčina a holandčina tvoria najkompaktnejší pár, zatiaľ čo angličtina zaujíma v priestore týchto Markovovských modelov mierne „odsunutú“ pozíciu. Zároveň však farebná škála ukazuje, že aj najväčšie rozdiely zostávajú v nízkom rozsahu, takže všetky tri jazyky zdieľajú veľmi podobný „interpunkčný štýl“ na úrovni prechodových pravdepodobností.



Obr. 6.7: Tepelná mapa párových vzdialeností medzi jazykmi na základe váženej riadkovej Jensen–Shannonovej divergencie prechodových matíc.

6.5 Stabilita jazykových modelov a odľahlé knihy

Okrem porovnania jazykových priemerov nás zaujíma aj to, ako veľmi sa jednotlivé knihy v rámci jedného jazyka od priemerného modelu odchyľujú. Pre každú knihu preto počítame vzdialenosť jej prechodovej matice od príslušného jazykového priemeru.

Súhrnné štatistiky pre jednotlivé jazyky sú uvedené v tabuľke 6.1. Táto tabuľka vychádza z dát v `markov_stability_summary.csv`.

Tabuľka 6.1: Súhrn vzdialeností kníh od jazykového Markovovského modelu. Pre každý jazyk uvádzame počet kníh, medián vzdialenosti, 95 % interval spoľahlivosti, minimum a maximum.

jazyk	n_{knih}	medián	$\text{CI}_{95}^{\text{lo}}$	$\text{CI}_{95}^{\text{hi}}$	min	max
de	80	0,0418	0,0385	0,0472	0,0161	0,1109
en	80	0,0368	0,0320	0,0431	0,0155	0,1383
nl	80	0,0536	0,0492	0,0587	0,0254	0,1444

Vidíme, že:

- medián vzdialenosti knihy od jazykového priemeru je u nemčiny a angličtiny o niečo menší než u holandčiny,

- holandčina vykazuje väčšiu variabilitu (vyšší medián aj maximum), čo naznačuje, že knihy v holandčine sú z hľadiska interpunkcie heterogénnejšie,
- v každom jazyku existujú knihy s výrazne väčšou vzdialenosťou (odľahlé hodnoty), ktoré sa od typického modelu odchyľujú.

Konkrétne príklady odľahlých diel možno identifikovať v tabuľke `per_book_distance_to_language_ma` kde každému Gutenberg ID prislúcha hodnota vzdialenosti od jazykového priemeru.

6.6 Interpretácia doterajších výsledkov

Doterajšie výsledky ukazujú niekoľko kľúčových faktov:

- Štatistika interpunkčných znamienok (frekvencie na 1000 slov) odhaľuje jazykovo špecifické rozdiely, no zároveň potvrdzuje podobnú základnú štruktúru v rámci germánskych jazykov, v zhode s pozorovaniami [5].
- Markovovské modely prechodov interpunkcie poskytujú prirodzený spôsob, ako kvantifikovať „štýl“ nadväzovania znamienok. Priemerné jazykové matice majú charakteristický tvar a ich porovnanie pomocou Jensen–Shannonovej divergencie ukazuje, že nemčina a holandčina sú si najbližšie, kým angličtina je mierne odlišná.
- Rozdelenie vzdialeností jednotlivých kníh od jazykového priemeru naznačuje, že v každom jazyku existuje „typické“ správanie, okolo ktorého sú knihy rozložené, a zároveň niekoľko odľahlých diel s netradičným používaním interpunkcie.

Z metodologického hľadiska sa ukazuje, že kombinácia frekvenčných štatistík, Markovovských modelov a informačno–teoretických mier vzdialenosti poskytuje konzistentný rámec pre porovnávanie jazykov na úrovni interpunkcie. V ďalších častiach práce bude možné tieto výsledky rozšíriť o detailnejšie sieťové ukazovatele a prípadne o porovnanie s inými typmi textov (napr. technické texty, texty s jazykovou poruchou), podobne ako v [4].

Kapitola 7

Interpretácia výsledkov a diskusia

Kapitola 8

Krátky manuál k softvérovému nástroju

Záver

Literatúra

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2 edition, 2006.

[2] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51(4):1079–1187, 2002.

[3] Martin Gerlach, Francesc Font-Clos, and Eduardo G. Altmann. A standardized project gutenberg corpus for statistical analysis of natural language. *Entropy*, 22(1):126, 2020.

[4] Tatiana Jajcayová, Jozef Kubik, Mária Markošová, and Soňa Senkovičová. Searching texts for signs of aphasia. In *ITAT 2021: Information Technologies – Applications and Theory*, volume 2962 of *CEUR Workshop Proceedings*, pages 171–175. CEUR-WS.org, 2021.

[5] Andrzej Kulig, Jarosław Kwapień, Tomasz Stanisz, and Stanisław Drożdż. In narrative texts punctuation marks obey the same statistics as words. *Information Sciences*, 375:98–113, 2017.

[6] Mária Markošová. Dynamika sietí. In *Umelá inteligencia a kognitívna veda II*, pages 321–377. Slovenská technická univerzita, Bratislava, 2010.

Príloha A: obsah elektronickej prílohy