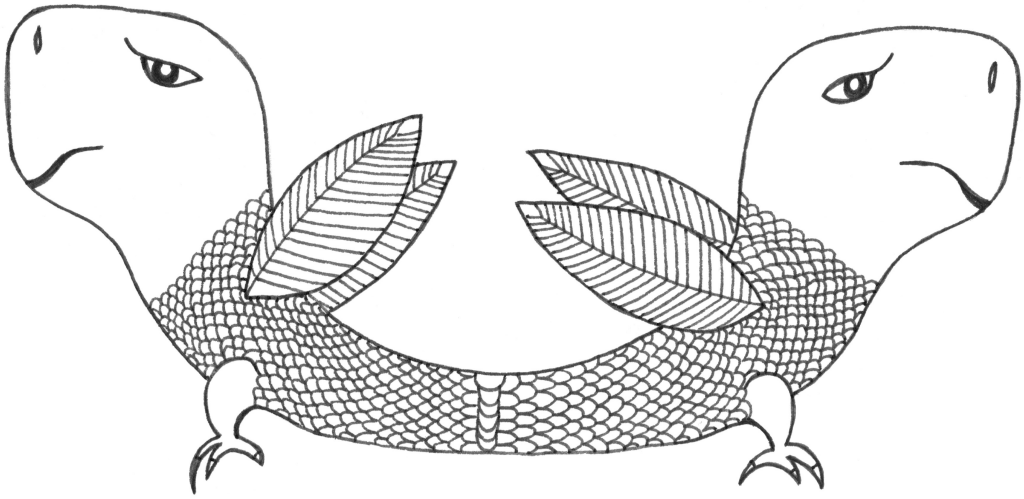
Univerzita Karlova
Filozofická fakulta

HABILITAČNÍ PRÁCE

Lexikální diverzita

Lexical Diversity

2022 Jiří Milička



Psát monografii bez spoluautorů je dneska docela anachronismus. Alespoň jsem se tedy opíral o spoustu kamarádů a kolegů. Zejména děkuji Václavu Cvrčkovi a Dominice Kováříkové, kteří mě k psaní průběžně dokopávali. Také děkuji Petru Zemánkovi za kolegiální podporu.

Ale hlavně děkuji slečně Viktorii, která s tímto spisem soutěžila o to, kdo bude dřív venku, a nakonec to vyhrála. A Hance, díky které se tenhle závod vůbec odehrál a která je autorkou úvodní ilustrace těchtýše lesního. Těchtýš je bájný tvor, s nímž se na stránkách této knihy setkáte dokonce dvakrát.

Také děkuji rodičům, kteří se tentokrát zapojili přímo do vědeckého procesu.

A jako obvykle i daňovým poplatníkům České republiky, kteří mě po dobu psaní nezištně podporovali a kterým tímto přeji hodnotný zážitek z četby.

Abstrakt

Lexikální diverzita je ústředním pojmem pro kvantitativní lingvistiku, ovšem nachází uplatnění i ve stylometrii, psycholingvistice či v didaktice jazyků a různých aplikovaných disciplínách.

Lexikální diverzita je netriviálně závislá na délce měřeného textu, což je fenomén, který během dvacátého století elicitoval vznik desítek metrik, jež tento problém více či méně neúspěšně řešily, přičemž obětovaly část své interpretovatelnosti.

Vzhledem k tomu, že pomocí metody klouzavého okna je možné zbavit libovolnou metriku lexikální diverzity její závislosti na délce textu, je možné se soustředit právě na interpretovatelnost metrik a další potřebné kvality, které byly opomíjeny, jako je intuitivní škálování, intersubjektivita či jednoznačně definovaná a srozumitelná jednotka.

Na tomto základě jsme vybrali právě ty metriky a metody měření, které v těchto parametrech excelují, a systematizovali je jednak racionálně, jednak empiricky na základě analýzy hlavních komponent (PCA). Z analýz v této knize představených vykrystalizovala definice lexikální diverzity jakožto intenzivní tenzorové lingvistické veličiny.

Klíčová slova: Kvantitativní lingvistika, lexikální diverzita, slovní bohatství, metriky diverzity, čeština, angličtina, arabština.

Summary

Lexical diversity is a central concept for quantitative linguistics and it is also used in stylometry, psycholinguistics, language didactics and various applied disciplines.

Lexical diversity is non-trivially dependent on the length of the measured text, a phenomenon that, during the twentieth century, elicited the emergence of dozens of metrics that more or less unsuccessfully addressed the problem, sacrificing some of their interpretability.

Since the moving window method can free any metric of lexical diversity from its dependence on text length, it is possible to focus on the interpretability of metrics and other necessary features that have been neglected, such as intuitive scaling, intersubjectivity or a clearly defined and comprehensible unit.

Building on this, we selected precisely those metrics and measurement methods that excel in these parameters and systematized them both rationally and empirically on the basis of principal component analysis. From the analyses presented in this book, a definition of lexical diversity as an intensive tensor linguistic variable emerged.

Keywords: Quantitative linguistics, lexical diversity, vocabulary richness, diversity metrics, Czech, English, Arabic.

Obsah

Úvod

Vy a já, má drahá čtenářko či čtenáři, jsme zcela jistě příbuzní. Představte si prosím teď jednoho z našich společných předků, jak sedí ve větvích a hledí do tmy, kde se zalesknou ostré leopardí zuby a zasvítí jeho oči. Náš předek, byť velmi vzdálený, je ostřílený nemilosrdným přirozeným výběrem, takže se nenechá ukolébat tím, že leopard zavře tlamu. Dobře ví, že trajektorie leopardích zubů koreluje s trajektorií jeho očí, že stačí sledovat jen je, chápe leoparda jako jeden celek, jeden objekt, jeden systém, jeden token. A taky ví, že leopardi umí lézt po stromech, že je to obecná vlastnost leopardů a že tato vlastnost se vztahuje i na onen objekt hledící na něj ze tmy, neboť právě on je instancí třídy leopardů. Nebo, jak budeme říkat v této knize, je *tokenem* daného *typu*. Žádní dva leopardi nejsou navlas stejní, ale jsou chvíle, kdy je to úplně jedno, chvíle, kdy na rozdílech mezi nimi nezáleží.[1]

Leopard! zakřičí náš dávný předek do tmy, aby varoval své blízké. Leopard! zakřičí pro jistotu ještě jednou. A ani teď na rozdílu mezi prvním a druhým zvoláním nezáleží. Obě zvolání jsou tokenem stejného typu. Diverzita oněch dvou zvolání je tedy proklatě nízká. Alespoň pro nás, kteří se na to díváme z bezpečí jednadvacátého století — těžko říci, jestli náš dávný předek a jeho blízcí mezi zvoláními nějaký zásadní rozdíl neviděli. Diverzita je totiž navýsost subjektivní záležitost, neboť každému záleží na jiných rozdílech. Vždyť se ani neshodneme na způsobu, jak z oněch rozdílů, na kterých záleží, diverzitu určit, jak ji vypočítat.

A přesto se už nejmíň sto let lidé pokoušejí diverzitu oné závislosti na subjektu co nejlépe zbavit, najít metodu jejího měření, která by příliš nezávisela na tom, kdo měří. Protože s takto změřenou intersubjektivizovanou diverzitou se dá pracovat dál, dají se díky ní formulovat testovatelné hypotézy, a z nich se dají stavět teorie. A jako vždycky při tom prošlapali spoustu cest, z nichž jen málo není slepých. Tato kniha budiž průvodcem po těchto cestách, průvodcem, který nás (snad!) zavede k cíli blíž než průvodci jiní. Některé cesty budeme muset prošlapat sami jako první. A prošlapání některých už zůstane jen na vás.

[1] Pokud znáte Batesonovu definici bitu jako (nejmenšího) rozdílu, na kterém záleží (Bateson, 1972, str. 46), asi už tušíte, že se v této knize budeme občas dotýkat teorie komunikace.

* * *

Jsem zvyklý komunikovat jazykem a stylem vědeckých časopisů vyžadujících akomodaci ke svému modelovému čtenáři a omezený rozsah. Kniha je jiné médium, kniha dokáže oslovit mnohem širší čtenářstvo a dovoluje vysvětlování od základů a redundanci, na všechno je v ní dost místa. K tomu je tahle kniha v mé mateřštině, ve které se rád rétoricky rozmáchnu. Slibuji, že se vynasnažím situace nezneužívat, jako jsem učinil na začátku této kapitoly. Ale zaručit nemůžu samozřejmě nic.[2]

Struktura práce

Lexikální diverzita je pojem, který přebývá na rozhraní kvantitativní a korpusové lingvistiky a různých dalších lingvistik. Má mnoho teoretických i čistě praktických uplatnění.

Díky této všestrannosti je lexikální diverzita prakticky nevyčerpatelné téma, pojmy jako *lexical richness*, *vocabulary richness, lexical diversity, lexical variability* a *lexical complexity* mají na Google Scholar desítky tisíc instancí. Tato kniha není vyčerpávajícím historickým exkurzem mezi tyto desetitisíce článků, zaměříme se jen na to nejdůležitější a kriticky prozkoumáme a systematizujeme různé smysluplné přístupy k tomuto košatému tématu.

O tom je vlastně celá první kapitola, která představuje různé rozumné metriky a snaží se zjistit, odkud se vzaly a k čemu slouží, srovnává jejich interpretace a omezení.

Druhá kapitola je extenzí té první, neboť u dobře zvolené metriky naše metodologické martyrium nekončí, je třeba se také vypořádat s tím, že metriky jsou inherentně a netriviálně závislé na délce textu. Abychom mohli srovnávat lexikální diverzitu různě dlouhých textů, je třeba metriku nějak normovat.

Během těchto dvou kapitol vykrystalizuje sada metrik, kterými se dál budeme zabývat — jsou jimi klasické metriky spadající pod Hillovo kontinuum a jejich křížové varianty, doplněné o průměrnou délku slova, rozdílnost a podíl autosémantik. Tyto metriky budeme normovat pomocí metody klouzavého okna.

Abychom mohli pokračovat dál a s klidným svědomím spolu metriky srovnávat, je třeba je napřed převést na stejnou škálu. O škálování vybraných metrik pojednává třetí kapitola.

Lemmatizovat či nelemmatizovat je klasické dilema, před které je postaven každý, kdo měří lexikální diverzitu. O důležitosti tohoto rozhodnutí je čtvrtá kapitola.

Metoda klouzavého okna nejenže bezpečně odstraňuje problém různě dlouhých

[2] Přiznávám, že v poznámkách pod čarou si občas pustím prsty na špacír. Jistě se najdou tací, kteří si z celé knihy budou číst právě pouze poznámky pod čarou, pro jejich neformální povahu. Doufám, že ani ty z vás nezklamu.

textů, ale také dovoluje měřit lexikální diverzitu na různých úrovních díky tomu, že si můžeme zvolit délku klouzavého okna. O tom se rozepisuji v kapitole páté.

Ve zmíněných kapitolách představuji nejméně pět důležitých metodologických rozhodnutí, která určují, co vlastně měříme. Těchto pět dimenzí, které charakterizují metody měření lexikální diverzity, shrnuji v šesté kapitole. Ony dimenze jsou určeny apriori, ovšem ke srovnatelným dimenzím přijdeme i empiricky pomocí analýzy hlavních komponent.

Na tomto základě pak na konci této kapitoly definuji lexikální diverzitu jako lingvistickou veličinu.

Korpusy, další zdrojová data a metodické poznámky

Vybral jsem tři jazyky, kterým rozumím — čeština, angličtina a arabština. Bylo pro mě zásadní, že jsem se mohl podívat do korpusu a případně si přečíst kusy textů, jejichž statistické vlastnosti byly podezřelé, udělat tak rychlý *sanity check*, jestli mám chybu v metodologii či její implementaci, či jestli je samotný text něčím neobvyklý.[3]

Zároveň je výhodné, že tyto jazyky shodou okolností dobře pokrývají tři pole ve skaličkovské typologické kvaternitě, neboť morfologie hraje v lexikální diverzitě zásadní roli. Tedy arabština má introflexivní, nonkonkatenativní morfologii, která se vyznačuje tím, že je schopna vložit morfémy do jiných morfémů, spojovat je tak, že splynou v jeden celek, koncovky však naopak věší poměrně pravidelně jednu za druhou, téměř aglutinativně. Češtinu asi nemusím představovat, jedná se o jazyk téměř prototypicky flexivní. Angličtina se vyznačuje morfologií izolační, tedy s omezeným inventářem koncovek a nepříliš bohatou derivací, což se projevuje mimo jiné tím, že její rodilí mluvčí mají problém vyslovit na první pokus termín *nonconcatenative morphosyntax*.

Čeština a arabština tedy sdílí různorodost tvarosloví, angličtina s češtinou pak patří do indoevropské rodiny a sdílí evropský kulturní areál. Pouhé tři jazyky samozřejmě nestačí k tomu, abychom mohli tvrdit, že naše hypotézy mají obecnou platnost, nicméně považuji za důležité, že je náš malý vzorek docela diverzifikován.

Výběr jazyků je oportunistický i z toho důvodu, že mám k dispozici jejich rozsáhlé a kvalitní korpusy. Technicky se při měření řídím defaultní tokenizací, lemmatizací a značkováním tak, jak je tvůrci těchto korpusů zamýšleli, s tou výjimkou, že ve všech korpusech je odstraněna interpunkce a tokeny obsahující číslice. Interpunkce je sice v korpusové tradici považována za token, respektive *pozici*, tedy prvek rovnoprávný s obyčejnými slovy, nicméně jejich význam je suprasegmentální a v češtině mají spíše

[3] Když jsem vzorky určené k měření automaticky náhodně vybíral z korpusu, zaznamenával jsem přesnou pozici, na které se nacházely. Pro pohodlné zobrazení této pozice, jejího kontextu a metadat textu, ze kterého pochází, jsem si vytvořil program `VerticalExplorer`, který je volně k dispozici společně s ostatními programy a skripty užívanými při tvorbě této knihy.

formální charakter. V případě arabského korpusu byla do většiny textů interpunkce dodána později při editorských zásazích, tedy nejedná se o původní součást textu. Číslice jsou pak v korpusech často součástí nějakých fragmentů či zbyly jako artefakty nedůsledného čištění, které vlastně nejsou integrální součástí textu — čísla stránek či poznámek pod čarou, záhlaví, editorské poznámky, telefonní čísla do redakce… jindy ovšem smysluplné místo v textu mají — třeba letopočty v historických knihách. Na rozdíl od odstranění interpunkce, odstranění číslic dokáže s mírou lexikální diverzity drasticky zahýbat, například tabulka se sportovními výsledky bude mít obrovskou míru lexikální diverzity, pokud čísla považujeme za slova, mnohem menší, pokud čísla odstraníme, a mizivou, pokud za slova považujeme jednotlivé číslice.

Odstranění číslic se tedy obhajuje poněkud hůře, ovšem je to jedno z mnoha arbitrárních rozhodnutí, která jsem musel učinit. Mohu však čtenářky a čtenáře uklidnit tím, že jsem různé statistiky v této knize uvedené měřil několikrát s různými detaily operacionalizace a výsledky se nelišily tak, aby bylo nutné změnit interpretaci a celkové vyznění textu.

SYN2015

SYN 2015 (Křen et al., 2015) je korpus česky psaných textů pocházejících z let 2010–2014 (ovšem několik je jich i starších, některé dokonce až z roku 1990). Velikostí i složením, které je pestré co do textových typů a žánrů (Cvrček et al., 2016), dobře zapadá mezi ostatní synchronní korpusy z Ústavu Českého národního korpusu, jejichž koncepce je historicky inspirována Britským národním korpusem. Korpus je pečlivě lemmatizován (Straková et al., 2014; Jelínek, 2008; Petkevič et al., 2014).

BNC

Britský národní korpus (British National Corpus) ve verzi, kterou je možné plně stáhnout (Consortium, 2007). Na rozdíl od SYNu neobsahuje celé texty, ale pouze jejich poměrně dlouhé fragmenty, což nás ovšem nijak neomezuje, neboť v této studii pracujeme se sekvencemi náhodně vybranými z textů, které respektují jejich hranice. Časový rozsah tohoto korpusu je mnohem větší než u SYNu 2015, zahrnuje prakticky celou druhou polovinu dvacátého století i začátek století jednadvacátého, nicméně stále je považován za synchronní. Na rozdíl od SYNů také obsahuje mluvenou složku. I tento korpus je pečlivě lemmatizován.

CLAUDia

Jedná se o korpus arabsky psaných textů nesrovnatelný s oběma předchozími, zejména tím, že je diachronní (obsahuje texty od počátku islámu až po dvacáté století), ale také velikostí (je několikrát větší) a žánrovým složením (Zemánek – Milička, 2017).

To ovšem nepřekvapí, neboť tematika, žánry a textové typy islámského středověku prostě vypadaly jinak než v Evropě dvacátého a jednadvacátého století. Korpus není lemmatizován ani morfologicky označkován a vzhledem k tomu, že automatické nástroje nedosahují pro arabštinu dobrých výsledků (stav roku 2021), v této studii je nepoužívám.

Korpus je převážně nevokalizován a pro konzistenci odstraňuji vokalické značky i tam, kde v korpusu původně jsou, takže vedle vcelku pravidelného „fonologického" zápisu češtiny a chaotického zápisu angličtiny je nevokalizovaný zápis arabštiny dalším zdrojem potenciálních rozdílů mezi naměřenými výsledky.

Pokud tedy mezi korpusy českých / anglických textů a korpusem textů arabských uvidíme nějaký rozdíl, může být zapříčiněn buďto rozdíly jazykovými, ale také rozdílností korpusů. O to víc nás potěší, pokud mezi těmito korpusy uvidíme stejné či aspoň podobné výsledky.

Další jednotlivé texty

Na některé grafy nebylo třeba velkého množství dat, neboť mají spíše ilustrační charakter. Pro tyto grafy jsem používal osvědčený román *The Last of the Mohicans* od Jamese Fenimora Coopera, protože jde o příjemný homogenní monotematický narativní text, na kterém různé kvantitativně lingvistické zákony fungují až podezřele očekávatelně. Obdobně jsem používal Jiráskovo *Temno* jakožto dlouhý a do značné míry monotematický narativní text.

Kapitola 1

Metriky

Abychom mohli vytvářet vědecké modely světa, který nás obklopuje, abychom mohli formovat hypotézy a empiricky je testovat, je nutné jeho vlastnosti nějak změřit. Tedy určit míry oněch vlastností co nejvíc bezrozporně a intersubjektivně.[1] Zjednodušeně — potřebujeme, aby měření vyšlo za stejných podmínek různým lidem stejně a aby o výsledcích mohli mezi sebou snadno komunikovat.

Může se zdát, že v tradičních přírodních vědách bylo snadné nalézt takové metriky a že problémy nastaly teprve, když jsme se začali pokoušet hledat obdobné způsoby měření ve vědách, které zkoumají člověka a jeho společenství. Ale není tomu tak, i historie přírodních věd je plná zmatku a bojů trvajících desítky let či staletí. Dnes bereme za samozřejmé, že existuje nějaká *teplota*, jejíž hodnota se odvíjí od průměrné rychlosti molekul a kterou chápeme jako okem neviditelnou vlastnost objektu, jež se ale viditelně demonstruje pomocí tepelné roztažnosti. Kvantifikovali jsme změnu teploty jako přímo úměrnou změně objemu materiálu, dejme tomu rtuti v teploměru, neboť tato konceptualizace vytváří málo nedorozumění, je jednoduchá na vysvětlení, snadno sdělitelná a prakticky jednoznačně určitelná. Tedy příhodně intersubjektivní.

Přitom původní představa, co to znamená, že něco je teplé a něco chladné, byla po tisíciletí v různých společenstvích různá a z pohledu dnešní fyziky velmi chaotická (McCaskey, 2020). Distinkce teplý — studený zahrnovala spoustu z dnešního pohledu chybných nebo metaforických interpretací, například spojovala teplotu s tepelnou vodivostí (dřevo je teplé, studený je kov) a vůbec schopností izolovat (teplé rukavice). Staletí trvalo vědě, než začala chápat teplo a chlad jako jednu škálu a než se zbavila filosofické kvaternity teplý — studený, vlhký — suchý, která opanovala představu

[1] *Intersubjektivitu* jsem si jako pojem vypůjčil od Poppera (2005). *Objektivita* je v epistemologii problematický termín a je těžké říct, co si pod ní vlastně představujeme. Ale intersubjektivita, tedy snaha nastavit metodologii tak, aby různí lidé došli za obdobných podmínek k obdobným výsledkům a mohli o nich snadno komunikovat, je rozumný požadavek.

o charakteru všeho neživého i živého v oblastech helénského vlivu a přilehlých krajích; tato představa byla rozvinuta později v alchymii a v některých zemích přetrvává v lidových představách dodnes.[2]

Zároveň je třeba si uvědomit, že moderní koncept teploty vlastně jen velmi přibližně pomáhá odpovědět na otázku, „jak je venku teplo“. Teplotu vzduchu totiž subjektivně nedokážeme vnímat samostatně, oddělit ji od vlivu jeho vlhkosti či síly větru.

Dnešní pojetí teploty je natolik odtažené od našeho vnímání, že vznikly subjektivizující metriky, které míchají mnoho moderních fyzikálních veličin do jedné, tzv. *zdánlivé teploty* (*apparent temperature*, též známé jako *pocitová teplota*, *feels like temperature*).[3] K tomu si připočtěme, že vnímání teploty není pro člověka lineární (například rozdíl mezi 60 a 65 °C je pro člověka jiný než mezi 35 a 40 °C) (Lautenbacher et al., 1992; Schweiker et al., 2017).

Lexikální diverzita se nachází na podobné křižovatce jako kdysi teplota. Naše konceptualizace je velmi neurčitá, široká a subjektivní. Podobně jako distinkce teplého a studeného byla kdysi uměle rozdělena mezi fyzikální veličiny jako teplota, teplo, tepelná vodivost atd., musíme i lexikální diverzitu rozdělit do několika intersubjektivních metrik. Úkol je to sice obtížný, nicméně úlevné je, že nemusíme hledat *jednu správnou* metriku. Nakonec samozřejmě můžeme ony metriky spojit nějak chytře zpátky do jedné, která bude podobně jako zdánlivá teplota lépe odpovídat našemu subjektivnímu vnímání.

V ideálním případě půjdou za pomoci oněch metrik dobře vyjádřit různé lingvistické zákony lexikální diverzity se dotýkající. Abychom zůstali u našeho příkladu s teplotou, je docela výhodné, že se nakonec konceptualizovala tak, že můžeme snadno spočítat, co se stane, když do vany, ve které je 50 litrů vody o 20 °C, vlejeme pětilitrový hrnec vřelé vody (vana bude pořád docela studená, ověřeno výpočtem i experimentálně). Se subjektivizující metrikou zdánlivé teploty nic takového snadno udělat nemůžeme.

Ještě jednu lekci si z konceptualizace teploty a historie jejího měření můžeme odnést. Teploměry v 16. a 17. století nedokázaly změřit teplotu, jak ji známe dnes, ale pouze kombinaci teploty a atmosférického tlaku (v tu dobu to ovšem nikdo netušil, viz McCaskey, 2020). Taková veličina vlastně může být docela užitečná, například při

[2] Například íránské kulinářství dodnes vyvažuje potraviny podle toho, jestli jsou „teplé“ nebo „studené“, dejme tomu „studený“ jogurt je vhodné smíchat s „teplým“ medem. V tomto případě to docela funguje.

[3] První komplexnější model subjektivizující teploty představil v roce 1984 Robert G. Steadman (Steadman, 1984), aniž by ovšem své výsledky testoval experimentálně na lidech. Od té doby vzniklo mnoho různých variant, v dnešní době je asi nejrozšířenější RealFeel od AccuWeather. Ani u něj se mi ovšem nepodařilo dohledat, že by jeho validita byla rigorózně testována (viz `https://www.accuweather.com/en/weather-news/what-is-the-accuweather-realfeel-temperature/156655`).

předpovídání počasí, neboť před deštěm obvykle klesne jak teplota, tak tlak. Ovšem v různých dalších aplikacích je to velmi nešťastné. Podobně lexikální diverzitu je těžké změřit samostatně a docela se nám do toho plete délka měřeného textu, jejíž vliv není snadné eliminovat nebo normalizovat. Nakonec se teploměr a barometr podařilo docela dobře rozdělit. Tomu, jak oddělit slovní diverzitu v textu od jeho délky, se bude věnovat celá 2. kapitola, nyní se ovšem podívejme na to, co pro nás *slovní diverzita* vlastně může všechno znamenat.

1.1 Počet typů (slovní bohatství)

Distinkci *typů* a *tokenů*, která byla jakožto základní koncept neformálně představena v úvodu, převzala kvantitativní a korpusová lingvistika od Charlese Sanderse Peirce (1931, §4.537),[4] ovšem ve skutečnosti není počet slovních typů ničím složitějším než prostým součtem *různých slov* v textu.[5] Tuto jednoduchou metriku můžeme nazvat i jako *slovní bohatství* (*lexical richness* či *vocabulary richness*).

Počet slovních typů (počet různých slov, slovní bohatství) je metrika snadno změřitelná, komunikovatelná i interpretovatelná. Tedy, ve skutečnosti je nejsnadněji změřitelná, komunikovatelná a interpretovatelná ze všech metrik, které si v této kapitole představíme, ale zase tak jednoduché to nebude, ďábel je ukryt v detailech i v základních principech.

Každý koncept, který pracuje s distinkcí typů a tokenů, se nepřekvapivě musí vyrovnat se dvěma základními problémy: tokenizací a typizací. Pokud si myslíte, že rozdělit text na jednotlivá slova je jednoduchá záležitost, na které se každý shodne, zkuste si rozdělit na slova větu *Ale nevěděls, že o „Československu“ se v devětatřicátém psalo jako o „Česko-Slovensku“, psalo-li se o něm vůbec…* Pak se na to zeptejte někoho dalšího a výsledky si porovnejte.

Shodnout se na typizaci je ještě obtížnější než na tokenizaci. Co všechno zakládá distinkci mezi typy, tedy v čem a jak moc se slova musí lišit, abychom uznali, že to

[4] Aby to nebylo tak jednoduché, v běžných pracech kvantitativní a korpusové lingvistiky se pojem token používá i tam, kde by jej Peirce nejspíš vůbec nepoužil a kde by spíše použil termín *výskyt* (*occurrence*), totiž k označení instance typu uvnitř jiného typu. Tedy pokud můžeme věřit výkladu a interpretaci Lindy Wetzel (2018), má smysl se ptát, kolik slovních tokenů má konkrétní vydání Máchova Máje, fyzická knížka, kterou právě teď beru do ruky. Nebo kolik tokenů měla recitace Máchova Máje, kterou jsem musel vyslechnout 4. listopadu 2014. Ovšem otázka „kolik tokenů má Máchův Máj jako takový?“ nedává smysl, neboť v této otázce chápeme samotný Máchův Máj jako typ, který tím pádem nemůže být rozdělený na tokeny. V diskurzu kvantitativní a korpusové lingvistiky jsou obvykle termíny *token*, *výskyt*, *instance* a *pozice* významově zaměnitelné a zaměňované a spíše jsou vázány kontextuálně.

[5] Obecně ovšem nemusí jít jen o slova, ale prakticky o jakékoli entity, u nichž má smysl rozlišovat typy a tokeny. Formálně tedy můžeme metriku definovat pomocí multimnožin, kde multimnožina $\mathcal{T}$ reprezentuje tokeny, její kardinalita reprezentuje počet tokenů, její support $\mathcal{V} = \mathrm{Supp}(\mathcal{T})$ reprezentuje typy a kardinalita tohoto supportu $|\mathcal{V}|$ je konečně počtem typů, tedy naší metrikou diverzity dané multimnožiny.

jsou *různá* slova? Je distinktivní velikost písmen? Jsou různé pravopisné nebo dialektální varianty distinktivní? A co překlepy? Co diakritika? Jsou distinktivní významy? Je kokrhající kohoutek stejným typem jako kohoutek vodovodní? Pokud není, pak se dostáváme na tenký led sémantiky, kde neexistuje ostrá hranice mezi homonymií, polysémií a metaforou.

Co když za distinktivní nebudeme považovat koncovky, takže různé tvary téhož slova nebudeme chápat jako různá slova? Tedy co když slova *vařím*, *vaříš* a *vaříme* budeme považovat za jeden slovní typ? Tato široká typizace, v korpusové lingvistice tradičně zvaná jako lemmatizace, je pro měření slovního bohatství překvapivě výhodná a podle experimentů Jarvise odpovídá intuitivnímu chápání lexikální diverzity lépe než typizace podle slovních tvarů (Jarvis – Hazhangmoto, 2021). Ovšem přináší obrovské množství otázek, které je třeba rozhodnout: patří ke stejnému lemmatu jako *vaříš* i slovo *uvaříš*? A co *uvařeno*, *uvařenými* nebo dokonce *vaření*? Přístup ČNK jde minimalistickou cestou a všechna tato slova typizuje jako vzájemně různá lemmata (Jelínek et al., 2021), což ovšem neznamená, že se na některé aplikace nemusí hodit komplexnější spojování do větších morfologicky či sémanticky spřízněných celků.

Ještě složitější je situace v diachronní lingvistice, neboť najednou je mnohem obtížnější se opírat o čistě formální aspekty, které se během dějin nutně měnily. Nejen že pravopis variuje v čase a prostoru, ale je obtížnější se opírat o sémantiku, která se také vyvíjela. Je slovo straſſnj ve významu „strašidelní“ rozdílné od slova *strašní* v dnešním významu, respektive v dnešních významech? Pokud ne, tak kde nakreslíme dělicí čáru?

A to mluvíme o psaném textu, kde samotný produktor udělal část práce za nás a stanovil jakés takés hranice pomocí mezer a fonémy typizoval do písmen, v mluvených textech je situace o mnoho složitější. Budeme se při jejich tokenizaci i typizaci řídit diktátem psaného jazyka a jeho obvyklých pravidel, nebo se pokusíme nějak diskretizovat přímo zvukové vlny?

Je třeba si uvědomit, že naše rozhodnutí, jak budeme tokenizovat a typizovat, ovlivňuje, co vlastně měříme, a mělo by přímo souviset s interpretací výsledků. V reálném světě je ovšem potřeba tyto otázky vyřešit nejen s respektem k výzkumnému záměru, ale zejména tak, aby naše měření bylo opakovatelné, a to jak na stejném textu, tak na srovnatelných textech, pokud možno i v textech srovnatelných jen obtížně, neboť právě takové obvykle srovnávat potřebujeme. Například tokenizace a typizace korpusů ČNK je do značné míry jednotná a řídí se podle písemné formy a relativně jednoduchých pravidel.[6] K sémantice přihlíží minimálně, i když stále je méně formální

[6]Například nepříliš vyčerpávající popis tokenizace momentálně nejnovějšího korpusu SYN2020 naleznete na adrese `https://wiki.korpus.cz/doku.php/cnk:syn2020:tokenizace`.
Obecné guidelines pro lemmatizaci synchronních korpusů ČNK se mi nepodařilo dohledat. Vzhledem k historickému vývoji morfologického značkování nemůžeme čekat přílišnou konzistenci, kupříkladu slova *zvířátko* a *zviřátko* patří k různým lemmatům, podobně jako další nářeční varianty slov (*být* — *bejt* atd.), zatímco dvojice *píča* a *piča* jsou přiřazeny ke stejnému lemmatu (*píča*). Diachronní korpusy ČNK naopak rozvíjejí poměrně systematicky několik úrovní sublemmat a hyperlemmat.

než tokenizace techničtěji založených formalismů, které vycházejí z tradic počítačové lingvistiky a NLP (zpracování přirozeného jazyka).[7]

Úlevné je, že všechna tato rozhodnutí mají na celkový výsledek menší vliv, než se na první pohled zdá. Vyzkoušíme si to na skutečných datech v kapitole 4. V ní standardní tokenizaci a typizaci v psaných korpusech ČNK porovnáme s výsledky, které získáme, když za slovní typy budeme považovat lemmata. Je to vlastně dost extrémní metodologický rozdíl, rozdíly v tokenizaci a typizaci budou obvykle subtilnější než rozdíly mezi slovními tvary a lemmaty. Přesto jsou naměřené rozdíly vcelku malé, jak si ukážeme. Alespoň tedy pro tuto základní metriku.

1.2 Počet hapax legomena

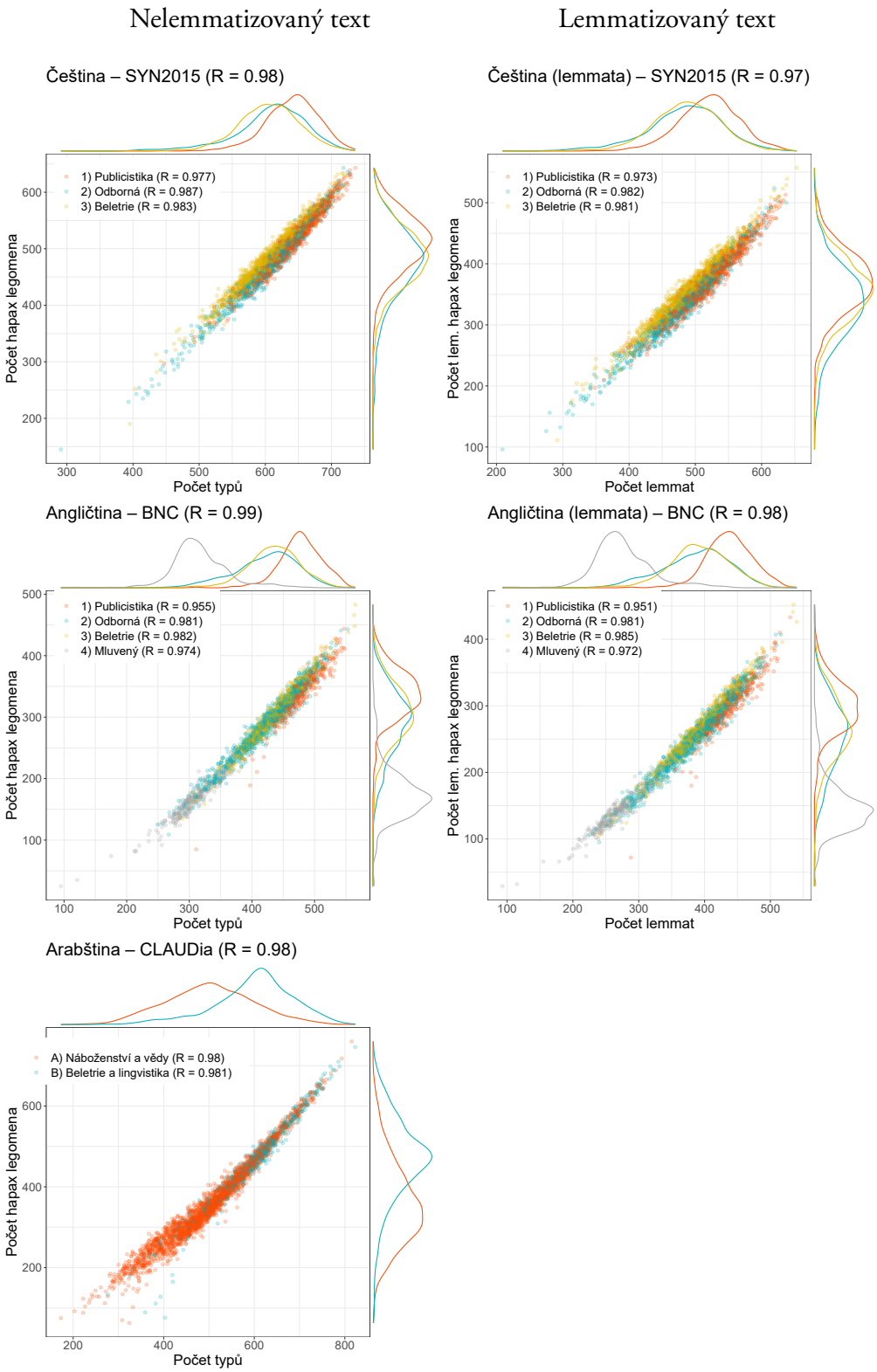
Slovní bohatství je metrika sice jednoduchá a skutečně základní, nicméně je možno ji různě variovat. Kromě zmíněných variací, které využívají toho, že si nejsme jistí, jak vlastně definovat token a typ, nabízí se ještě možnost zahrnout jen některá slova, slova vybraná podle nějakých kritérií. Například vyfiltrujeme pouze slova určité frekvenční hladiny — třeba ta, která jsou v měřeném textu velmi častá, nebo naopak vzácná. Získáme tak nový úhel pohledu, který sice bude s počtem všech slovních typů značně korelovat, nicméně podle zvoleného frekvenčního filtru zvýrazní určité charakteristiky.

Zajímavé jsou zejména počty slov s nízkou frekvencí, neboť mohou fungovat jako jakási míra tvořivosti či produktivity. Slavné jsou zejména studie o morfologické produktivitě od Haralda Baayena (počínaje 1992), které jako metriku produktivity užívají relativní počet hapax legomena (tedy počet slov vyskytujících se pouze jednou vydělený počtem všech tokenů). Ovšem tato metrika je již staršího data: je základem slavného Goodova článku (Good, 1953), podle něhož se její historie táhne minimálně až k Bletchley Parku a Turingovi, jako ostatně i spousta dalších principů, na které v této knize narazíme. Relativní počet hapaxů totiž v případě homogenního textu funguje jako aproximace růstu počtu typů.[8] Je-li tedy v textu relativně velké množství hapaxů, znamená to, že v něm rychle přibývají nové typy.

Tuto vlastnost hapaxů si můžeme ilustrovat na obrázku 1.1. Stejně jako v mnoha obrázcích, které budou následovat, každý datový bod zde reprezentuje jednu sekvenci o tisíci tokenech, náhodně vybranou z korpusu tak, aby respektovala hranice textů.

[7] Například guidelines pro Universal Dependencies jsou velmi obecné a prakticky nechávají rozhodnutí na anotátorech jednotlivých jazyků a jejich nástrojích. Viz `https://universaldependencies.org/u/overview/tokenization.html`

[8] V podstatě jde o diferenciaci počtu typů v závislosti na počtu tokenů (diferenciace je něco jako derivace, ale na diskrétních datech). Pokud jste se ještě nesetkali s grafem závislosti počtu typů na počtu tokenů, doporučuji obrázky 1.12 a 1.13 (červená linie). Pokud by vás téma zaujalo, s diferenciací této závislosti jsem si trochu pohrál v článku (Milička, 2013).



Obrázek 1.1: Korelace typů a hapaxů ve vzorcích o tisíci tokenech.

Barvy reprezentují druh textu, ze kterého byla ona sekvence vybrána.[9] Jak je vidět, přestože počet typů a počet hapaxů spolu velmi dobře lineárně korelují (Pearsonův korelační koeficient se blíží jedné), dokáže tato dvojice proměnných ony textové typy docela hezky separovat. Zatímco samotný počet typů či samotný počet hapaxů nám dejme tomu publicistiku a beletrii v češtině příliš nerozliší, což je vidět na distribucích na okraji grafů, jejich kombinace už ano. Pokud totiž vezmeme vzorky z beletrie o určitém počtu typů a srovnáme je se vzorky z publicistiky, které mají přibližně stejný počet typů, zjistíme, že beletristické texty mají systematicky vyšší počet hapax legomena. Což docela dobře odpovídá naší představě o počtu hapaxů jako o indexu lexikální kreativity, právě tu bychom totiž u beletrie čekali vyšší než u publicistiky. Pozorování platí jak pro lemmatizovaný, tak pro nelemmatizovaný text.

Jako indikátor produktivity můžeme používat všechna slova s relativně nízkou frekvencí, ovšem používají se právě hapax legomena, slova, jež se v textu vyskytují pouze jednou[10] — jednak kvůli zmiňovanému vztahu objevenému minimálně již Turingem, ale nejspíše i z numerologických důvodů. Určitou fascinaci hapaxy jsme zdědili již z předkvantitativních dob, kdy byly noční můrou klasických filologů zabývajících se mrtvými jazyky, neboť význam slova, jež má v celém díle či dokonce kánonu pouze jeden výskyt, prakticky není možné odhadnout z jeho kontextu. Tento zájem pak pokračoval i v časech kvantitativní a korpusové lingvistiky — viz například Herdan (1959). Hapaxy se podle mě používají zejména proto, že o konstantě velikosti jedna se v naší kultuře nediskutuje. Pokud byste si vybrali číslo dejme tomu osm, museli byste vysvětlit proč. Přitom ovšem při porovnávání nesouměřitelně dlouhých textů dává smysl tuto konstantu přizpůsobit velikosti textu, neboť pro velmi krátké texty jsou slova s frekvencí osm velmi frekventovaná, pro velmi dlouhé texty může tento práh být naopak málo velkorysý. V krátkých textech jsou hapaxy jednoduše vzácná slova, v dlouhých textech mohou být spíše kolekcí překlepů. Velikost prahu ovšem není jednoduchá lineární závislost, budeme tedy dále předpokládat, že texty, které chcete srovnávat, jsou alespoň řádově stejně velké, popřípadě že se s rozdíly ve velikosti textů dokážete nějak vypořádat — nejlépe pomocí metod popsaných v kapitole 2.2.1. A že tedy nebude přílišnou metodologickou chybou postupovat tradičně a prostě použít ve všech případech počty hapaxů.

[9] Respektive užívám klasifikaci, které se v korpusové lingvistice říká *textové typy*. V případě anglického korpusu BNC pak ještě vyděluji mluvené texty, takže jde o jakousi smíšenou oportunistickou typově-modální klasifikaci, která ovšem, vzhledem ke složení korpusu a anotaci jeho metadat, dává docela smysl. Hlavní snahou bylo, aby počet kategorií nebyl příliš velký a graf se tak nestal nepřehledným, zároveň aby užité kategorie dobře kontrastovaly. Poněkud exotická kategorie *beletrie a lingvistika* v arabštině vychází z toho, že lingvistické práce se chovaly jinak než ostatní vědy a spíše splývaly s beletrií, tuto klasifikaci jsem použil už v (Milička, 2018).

[10] Formálně opět můžeme počet hapaxů definovat pomocí multimnožin, kde multimnožina $\mathcal{T}$ reprezentuje tokeny, množina $\mathcal{H}$ reprezentuje všechny typy, které mají multiplicitu právě jedna, takže $\mathcal{H} = \{x \in \mathcal{T} \mid m_{\mathcal{T}}(x) = 1\}$, a kardinalita této množiny $|\mathcal{H}|$ je konečně počtem hapax legomena, tedy naší metrikou diverzity dané multimnožiny.

1.3 Pravděpodobnost opakování (repeat rate)

Pokud text posuzujeme podle samotného počtu typů, či dokonce jen jejich podmnožiny, pak ztrácíme obrovské množství informace, informace o tom, který typ má jakou frekvenci. Přitom právě to je základním prvkem diverzity. Intuitivně: pokud budeme mít skupinu lidí, kde je čtyřicet mužů a jedna žena, tak bude méně diverzifikovaná než skupina o jednadvaceti mužích a dvaceti ženách, přestože počet pohlaví je v obou skupinách stejný.

Mohli bychom jako metriku použít průměrnou frekvenci typů? Pokud bychom jednoduše sečetli všechny frekvence typů ve vzorku, dostali bychom počet všech tokenů. A součet relativních frekvencí typů je roven jedné. Tudy tedy cesta nevede. Co ale můžeme udělat, je, že spočítáme průměrnou relativní frekvenci tokenů. Tedy jednoduše u každého jednotlivého tokenu určíme relativní frekvenci typu, ke kterému náleží, tyto relativní frekvence sečteme a podělíme počtem tokenů. Tento postup vyjadřuje vzorec 1.1.

$$\lambda = \frac{\sum_{i=1}^{N} \frac{f(t_i)}{N}}{N} \tag{1.1}$$

Ten je ekvivalentní následující variantě (1.2), kde nesčítáme relativní frekvence tokenů, ale čtverce relativních frekvencí typů:

$$\lambda = \sum_{i=1}^{V} p(t_i)^2 \tag{1.2}$$

V obou těchto vzorcích, jako ostatně ve všech následujících, N značí počet tokenů, V značí počet typů, $f(t)$ je pak absolutní frekvence jednotlivých slov.[11] Relativní frekvence slov pak značím jako $p(t_i)$, tedy platí, že $p(t_i) = \frac{f(t_i)}{N}$.

Tímto se nám podařilo frekvence typů před sečtením transformovat tak, že výsledek uchovává určitou informaci o vnitřní struktuře slovníku, zároveň si pod tímto číslem dokážeme něco představit.

Ovšem ono číslo se dá interpretovat ještě zajímavěji: shodou okolností totiž funguje i jako pravděpodobnost opakování slov, tedy konkrétně pravděpodobnost, že dvě slova náhodně vytažená z textu nebo jeho vzorku budou stejná. Pravděpodobnost, že dva náhodně zvolené tokeny budou patřit ke stejnému typu.

[11] Dané značení je běžné v kvantitativně lingvistické literatuře. Abychom byli preciznější, můžeme pro popis proměnných opět použít multimnožiny: $\mathcal{T}$ je multimnožina reprezentující tokeny, $\mathcal{V}$ je množina reprezentující typy, takže $\mathcal{V} = \mathrm{Supp}(\mathcal{T})$, pak $N = |\mathcal{T}|$ a $V = |\mathcal{V}|$. Dále $f(t)$ reprezentuje multiplicitu typu t v multimnožině $\mathcal{T}$. Ve vzorci 1.1 iterujeme všemi tokeny, pro každý z nich nalezneme typ, jehož je instancí, a k tomuto typu nalezneme jeho multiplicitu. V případě vzorce 1.2 iterujeme přímo jednotlivými typy t_i, takže $f(t_i) = m_{\mathcal{T}}(t_i)$.

Relativní frekvence typu t, tedy $p(t)$, představuje totiž pravděpodobnost, že z textu náhodně vytáhneme token reprezentující právě typ t. No a když tento vytažený token vrátíme zpátky do textu a operaci opakujeme, tak pravděpodobnost, že vytáhneme opět nějaký token toho stejného typu t, je rovna $p(t)p(t)$, tedy $p(t)^2$.

Tato metrika diverzity je proto výborně interpretovatelná a skutečně se používá jak v lingvistice, tak v ekologii, ekonomii, nebo prostě kdekoli je potřeba měřit diverzitu, obvykle pod názvem *Simpsonův index* či *Simpsonova koncentrace* (Simpson, 1949; Rousseau, 2018), nebo jednoduše jako *repeat rate*, pravděpodobnost opakování typů.

Takto popsaná metrika skutečně předpokládá, že tokeny taháme z pomyslného pytlíku a pokaždé je zase házíme zpět. Ve skutečnosti ale originální Simpsonův vzorec (1949) chápe *repeat rate* jako pravděpodobnost, že se dva tokeny stejného typu potkají v náhodně zpřeházeném konečném textu.[12] Vzniká tak vzorec 1.3,[13] který sice vypadá velmi podobně jako ten předchozí, ale pro typickou distribuci frekvencí slov se může zásadně lišit, obzvláště na kratších textech. V krátkých textech totiž drtivou většinu slovní zásoby tvoří hapax legomena, která ovšem mají nulovou pravděpodobnost, že v textu potkají slovo stejného typu, záleží tedy na drobných změnách v distribuci těch několika málo typů, které hapaxy nejsou, a metrika se chová velmi chaoticky. V delších textech je tento chaos menší, nicméně je třeba brát v úvahu, že velkého počtu hapaxů se nezbavíme ani u opravdu dlouhých textů či dokonce jejich kolekcí.[14]

$$\lambda = \sum_{i=1}^{V} \frac{f(t_i)\,(f(t_i)-1)}{N(N-1)} \tag{1.3}$$

Jelikož chceme, aby metrika stoupala se stoupající diverzitou (obdobně jako slovní bohatství v předchozí kapitole), bude lépe využít převrácené hodnoty pravděpodobnosti opakování, jak si ukážeme na konci této kapitoly. Ovšem historicky se často

[12] Všimněte si, že nemáme na mysli tokeny, které jdou v textu za sebou. Pravděpodobnost, že dva sousedící tokeny jsou stejného typu, je v přirozených jazycích mnohem nižší, než že se tak stane u skutečně náhodně zvolených tokenů. Tedy pokud nejde o proslulou větu *Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo* (Rapaport, 2014).

[13] Pokud máte rádi obecné definice, můžeme repeat rate uvnitř textů konečné délky opět definovat pomocí multimnožin. Pokud $\mathcal{T}$ je multimnožina reprezentující tokeny a $\mathcal{A}$ je její náhodnou podmultimnožinou o mohutnosti 2 (tedy $\mathcal{A} \subseteq \mathcal{T} \wedge |\mathcal{A}| = 2$), pak repeat rate je definován jako pravděpodobnost, že support této podmultimnožiny má mohutnost rovnu jedné, tedy $|\mathrm{Supp}(\mathcal{A})| = 1$.

[14] Podíl hapaxů v typech v přirozených textech a jejich korpusech nikdy neklesne až k nule, ale konverguje k nějaké (poměrně vysoké) hodnotě, kolem které pak osciluje (Cvrček, 2011). Například v češtině se pohybuje okolo 44 procent i pro opravdu velké kolekce textů, jako je SYN 2015. V analytické angličtině klesne níž a hapaxy zabírají cca 38,5 % typů. Naopak syntetická až aglutinační arabština (reprezentovaná korpusem CLAUDia) má ve slovní zásobě cca 47,2 % hapaxů. Tento fakt je poměrně neintuitivní, pokud si neuvědomíme důležitost dynamiky jazyka nebo pokud závislost empiricky neověříme; například ještě v roce 2010 Popescu píše „[…] while very long texts in which all words are repeated would have richness 0“(Popescu et al., 2011, str. 1), což jednoduše není možné.

používala nikoli převrácená hodnota, ale pravděpodobnost opačného jevu, tedy že naopak dvě slova po sobě v náhodném textu budou různá (vzorec 1.4):

$$\rho = 1 - \lambda = 1 - \sum_{i=1}^{V} \frac{f(t_i)\,(f(t_i) - 1)}{N(N-1)} \tag{1.4}$$

Tato obrácená metrika se často označuje jako *Gini-Simpson index*, ale je docela možné, že tyto vzorce znáte pod nějakým jiným názvem, podle různých jiných domnělých objevitelů.[15]

Na událostech okolo jejich vzniku si můžeme ilustrovat, jak komplikovaná je historie jednotlivých metrik a jejich variant. Zkonstruování tohoto indexu nepředstavuje nijak výjimečný intelektuální výkon, základní principy, na kterých stojí, byly známy už minimálně od generace Pascala, Huygense a Jacoba Bernoulliho a nejdůležitější je vlastně otázka motivace, tedy nejen schopnost uvědomit si, že pravděpodobnost opakování dvou typů může sloužit jako metrika pro diverzitu systému, ale hlavně potřeba takové metriky. Jak napsal I. J. Good (1982), „any statistician of this century who wanted a measure of homogeneity would take about two seconds to suggest $\sum p_i^2$“.

Taková potřeba nastávala opakovaně a opakovaně byla touto metrikou naplňována. Kryptoanalytici ji nejspíš používali minimálně už v devatenáctém století (Rousseau, 2018), odtud se (možná přes polské kryptoanalytiky, viz Rejewski (1981)) dostala do Bletchley Parku, kde se běžně užívala a kde jí sám Turing říkal *repeat rate*, alespoň podle svědectví I. J. Gooda (1982). V citovaném článku Good dále uvádí (včetně referencí) několik dalších domnělých objevitelů této metriky (A. Sinkov z roku 1968 a S. Kullback z roku 1976), přičemž Kullback citovanou knihu podle Gooda sepsal už ve čtyřicátých letech, jenomže byla klasifikována. Podobně Simpson (1949), podle kterého je dnes metrika známá, se s ní podle Gooda seznámil u Turinga a necitoval ho kvůli režimu utajení, kterým informace z Bletchley podléhaly. Další kryptoanalytická větev podle některých autorů vedla od J. F. Friedmanna (1922), kde je metrika nazývána *index of coincidence*, ovšem abych se přiznal, to, co je zde popisováno, mi připadá jako něco trochu jiného.

Podobně mnozí autoři připisují index Ginimu (Gini, 1912), jiní naopak tvrdí, že s touto metrikou nemá nic společného (Rousseau, 2018). Čímž se dostáváme k ekonomické větvi, neboť touto metrikou je možné měřit například diverzifikaci bohatství ve společnosti. Tato větev začíná už v hloubi devatenáctého století (Lexis, 1879). Lexis je ve své době poměrně citovaný a zná jej i vlivný Keyness (1921, s. 398–399, zde také Keyness podrobně rozebírá práce von Bortkiewicze na toto téma), jenže pak se

[15] Popřípadě pod názvem *logická entropie* (*logical entropy*), který se snaží propagovat David Ellerman, jenž na tomto vzorci prakticky zakládá konkurenční teorii informace. Výsledek výpočtu opačného repeat rate (ve formě vzorce 1.2) pak nebere jako bezrozměrnou veličinu, ale jako číslo vyjádřené v jakýchsi *ditech* (což by měla být jednotka distinkce či diverzity), které v jeho teorii nahrazují bity (Ellerman, 2017).

linie ztrácí a po druhé světové válce se metrika vynořuje znovu, tentokrát ovšem jako *Herfindahlův index*, přičemž Herfindahl ji nejspíš poprvé použil ve své disertaci (Herfindahl, 1950). Vtipné je, že Otto (1964) si stěžoval, že index je po Herfindahlovi pojmenován neprávem, neboť jej vymyslel on, a to dokonce už roku 1945, a tedy že mu bylo prvenství ukradeno (Rousseau, 2018).

Zmíněné indexy nejsou zcela totožné, liší se různými kosmetickými doplňky, jako je vynásobení nějakou konstantou, odmocnění apod., čímž se ovšem fakticky nemění pořadí textů, když je podle těchto indexů seřadíte, a jediné, čeho se tak docílí, je možná tak horší interpretovatelnost.

V lingvistice se tato metrika asi nejčastěji používá pod jménem Simpsonův index, ale známý je i pojem repeat rate — takto ho zmiňuje třeba Radek Čech (2016, s. 92). Osobně souhlasím s Goodem, který horuje za to, aby se nepoužívaly u indexů jména jejich domnělých objevitelů (Good, 1982),[16] a budu tedy také používat termín repeat rate (RR, česky pravděpodobnost opakování) a pro jeho opačnou hodnotu pak termín *distinction rate* (DR, pravděpodobnost distinkce).

* * *

Častěji než pravděpodobnost opakování budu však používat její převrácenou hodnotu (1.5, značím jako RRR, *reciprocal repeat rate*, v literatuře asi častěji potkáte *inverted* či *inverse Simpson index*).

$$\mathrm{RRR} = \frac{1}{\lambda} = \frac{1}{\sum_{i=1}^{V} p(t_i)^2} \tag{1.5}$$

Jsou k tomu důvody estetické, neboť pravděpodobnost opakování je prakticky vždy nějaké číslo velmi blízké nule, převrácením hodnoty se tak stane mnohem čitelnější. Ale více než estetika mě k tomu vede škálování, neboť tato převrácená hodnota lineárně koreluje s počtem typů ((Hill, 1973), což potvrzuji také empiricky v kapitole 3). Tato korelace s počtem typů a dalšími metrikami je navíc pozitivní, tedy čím větší je převrácená pravděpodobnost opakování typů daného systému, tím různorodější a diverzifikovanější daný systém je.

[16] Pojmenovávání fenoménů po jejich domnělých objevitelích, tradiční obsese našeho civilizačního okruhu, dosahuje v případě stylometrických indexů, mezi něž lexikální diverzitu počítám, absurdních rozměrů. Například občas citovaný *Fucks' Index* či *Fucks Stilcharakteristik* (Briest, 1974; Michalke et al., 2021), pokud mohu věřit své němčině, není nic jiného než průměrný počet slabik ve větě (Fucks, 1955). Text je možné nasegmentovat na několika různých úrovních, například: hláska či písmeno → slabika či morfém → slovo → věta → souvětí. Takže můžete snadno „objevit" a po sobě nechat pojmenovat spoustu metrik: počet hlásek ve slabice, počet hlásek ve slově, počet písmen ve větě… jenom v našem příkladě je jich 18.

Díky tomuto převrácení navíc pravděpodobnost opakování přestává být bezrozměrnou veličinou a získává svou jednotku: *efektivní počet typů*,[17] která se používá i pro další metriky, které lineárně škálují s počtem typů a se kterou se tak setkáme i v následujících kapitolách. Efektivní počet typů můžeme v tomto případě interpretovat tak, že kdybychom vytvořili pseudotext, který by měl stejnou pravděpodobnost opakování jako náš měřený text a v němž by měly všechny typy stejnou frekvenci, pak by se počet efektivních typů rovnal počtu typů v onom pseudotextu.

1.4 Perplexita, entropie a komplexita

Metrika popsaná v předchozí kapitole se dá interpretovat jako aritmetický průměr relativních frekvencí tokenů v celém textu. Aritmetický průměr se často chápe jako default pro popis střední hodnoty, tedy jako něco, co použijeme, když nad popisovaným fenoménem příliš nepřemýšlíme, nebo když jiné možnosti ani neznáme. Přitom v mnoha případech dává mnohem větší smysl použít pro střední hodnotu jiné statistiky — modus, medián nebo geometrický průměr. A právě na geometrický průměr frekvencí tokenů se podíváme v této kapitole.

Geometrický průměr je vhodné používat tam, kde průměrované hodnoty nabyly své velikosti exponenciálním růstem či poklesem, což by při troše dobré vůle mohl být případ frekvence slov — alespoň podle klasického zákona Piotrovských (Piotrovskaja – Piotrovskij, 1974). Geometrický průměr ovšem nepoužijeme jen proto, že by byl pro popis střední hodnoty frekvencí slov inherentně lepší než průměr aritmetický, ale zejména proto, že jeho převrácená hodnota, takzvaná perplexita (vzorec 1.6),[18] nás vede přímo do lůna shannonovské teorie komunikace, kterážto je základním kamenem současné informatiky (Shannon, 1948).

$$P = \prod_{i=1}^{V} p(t_i)^{-p(t_i)} \tag{1.6}$$

Tato metrika „zmatení“[19] totiž není definována tak, aby souvisela s reálným pocitem zmatenosti či překvapení (dejme tomu měřeným v rámci nějakého psycholingvistického experimentu), ale tak nějak tiše předpokládá, že překvapivost úzce souvisí se

[17] Anglicky *effective number of types*, v ekologii se velmi často setkáváme s konkrétní variantou e*ffective number of species*. Pod názvem *equivalent number of species* tento pojem zavedl MacArthur (1965), přičemž ukázal, jak transformovat různé metriky tak, aby vyjadřovaly diverzitu právě v této jednotce. Pro ekonomii stejnou metodologii znovuobjevil Adelman (1969), ovšem jednotku pojmenovává jako *numbers equivalent*.

[18] Pro jistotu připomínám, že Π značí, že je třeba iterovat násobení, podobně jako v předchozích vzorcích Σ značila sumaci, tedy iteraci sčítání.

[19] Volba termínu není právě ideální, vlastně nám nejde ani tak o zmatenost, popletenost či dokonce propletenost (význam asi nejbližší etymologii tohoto slova), ale o překvapení, které jednotlivá slova průměrně posluchači způsobí.

Shannonovou entropií. Geometrický průměr relativních frekvencí tokenů totiž po zlogaritmování odpovídá odhadu entropie, a jde tak o stejné indexy lišící se pouze škálou (viz vzorec 1.7).

$$H = -\sum_{i=1}^{V} p(t_i) \log_2 p(t_i) \tag{1.7}$$

Tento rozdíl ve škálování je ovšem zásadní pro interpretaci obou hodnot. Entropii měříme v bitech,[20] tedy jakýchsi minimálních binárních rozdílech, na které jsme ochotní členit realitu, vlastně nám říká, kolik bitů informace bychom potřebovali průměrně k zapsání jednoho slova z daného textu, kdybychom dokázali onen text zakódovat co nejefektivněji. Jednotkou perplexity je pak, stejně jako v předchozí kapitole, efektivní počet typů. Tedy, kdybychom vytvořili pseudotext, který by měl stejnou délku a byl vytvořen systémem o stejné entropii jako náš měřený text, v němž by ovšem měly všechny typy stejnou frekvenci, pak by se počet efektivních typů rovnal počtu typů v onom pseudotextu.

Všimněte si, jak jsem se v předchozí větě vyhnul sousloví „entropie textu“ a místo něj použil podivně komplikované „text, který byl vytvořen systémem o stejné entropii“. Z definice totiž nic jako entropie textu neexistuje, pouze entropie *systému*, v rámci kterého byl daný text vytvořen — respektive v našem případě entropie distribuce slovních typů, ze kterých byl daný text vytvořen. Podobně jako nemá smysl mluvit o fyzikální entropii jednotlivých mikrostavů, ale pouze celého systému, který se jimi manifestuje — text je v tomto pojetí vlastně množina mikrostavů a otázka je, co chápeme jako systém, ke kterému se shannonovskou entropií v lingvistice vztahujeme.[21]

Entropii systému tedy jenom odhadujeme podle distribuce typů v textu, tato distribuce je jenom odleskem distribuce slovních typů v systému, kterou skutečně můžeme popsat pomocí entropie. Vlastně předpokládáme, že existuje nějaká ideální distribuce typů, ze které do textu vybíráme tokeny. Tento pohled na jazyk je poněkud absurdní, a nejspíš je absurdní i pro popis ekosystémů, neboť ekologové místo „odhad entropie ekosystému“ používají raději termín *Shannonův index*.[22]

[20] Respektive v shannonech. Skutečně, podle normy IEC 80000-13:2008 jsou bity pouze jednotkou informační kapacity, zatímco entropie se měří v shannonech. Můžeme-li věřit Googlu, na celém internetu se vyskytují pouze tři instance užití pojmu kiloshannon, z toho dvakrát ironicky a jednou jako chyba OCR. Jelikož tedy shannony nikdo nepoužívá, budu se držet bitů.

[21] Zde raději zdůrazním, že pokud máte bohemistické vzdělání, tak si pod pojmem *systém* představíte pravděpodobně něco radikálně jiného než Claude Shannon.

[22] Respektive Shannon–Wiener index, popřípadě Shannon–Weaver index, aniž by ovšem měl Weaver na jeho autorství nějaký podíl, nejspíš se jednalo o chybu, která se následně propsala do dalších prací (Spellerberg – Fedor, 2003). Wiener se Shannonem na jeho nejvlivnějším článku sice nespolupracoval, nicméně vzájemně se ovlivňovali a stejný vzorec vydal v témže roce jako Shannon (Wiener, 1948; Kullback – Leibler, 1951). Sám Shannon ovšem jako zdroj inspirace neuvádí jeho, ale Hartleyho práci (Hartley, 1928), která je i dnes, po sto letech, aktuální, byť určitý podíl Wienerovi v pozdější

Je to slovíčkaření, ale občas důležité. Například entropie systému je už z definice nezávislá na velikosti vzorku, na kterém ji měříme. To se ovšem nedá říct o jejím odhadu, tedy „Shannonově indexu“, který měříme na reálném textu. Tam totiž naopak velikost textu roli hraje — s přibývající velikostí textu odhad entropie roste. Teoreticky by mohl dokonvergovat pro dlouhé texty k nějaké rozumné stabilitě, ovšem vzhledem k tomu, že jazyk je dynamický systém,[23] se to nestane (Lozano et al., 2017). Nepomůže nám tedy ani neshannonovský odhad entropie, který ke kýžené hodnotě konverguje o něco rychleji.[24]

Ještě se krátce vrátím k perplexitě, kterou jednoznačně preferuji před entropií jako metriku lexikální diverzity, neboť její měřítko (počet efektivních typů) lineárně škáluje s počtem typů. Nicméně občas má smysl použít i logaritmické měřítko, tedy entropii, neboť nás může skutečně zajímat otázka, kolik informace lexikální systém kóduje. Entropie navíc souvisí s komplexitou, například pokud ji definujeme kolmogorovovsky,[25] přičemž komplexita lexika vyjádřená v bitech může charakterizovat vlastnosti, které chceme měřit, lépe než efektivní počet typů. Například pokud stavíme nějakou teorii ohledně toho, jak namáhavé je pro nás zpracovat či vyprodukovat určité množství informace.

1.5 Kontinuum metrik

Pokud jste se v posledních třech kapitolách mírně ztratili, navzdory mé upřímné snaze podat věci pokud možno nekomplikovaně, věřte, že v tom nejste sami, literatura kolem metrik diverzity je nesmírně košatá, spletitá a v mnoha případech redundantní.

práci přiznává (Shannon – Weaver, 1949, poznámka pod čarou 1). Představa, že diverzita systému by šla popsat pomocí binárního logaritmu všech stavů, kterých daný systém může nabývat, tedy pomocí jakýchsi distinktivních rysů, bitů, je ještě mnohem starší a první stopy se dají nalézt nejméně v sedmnáctém století — v esoterickém spisku Johna Wilkinse *Mercury or the Secret and Swift Messenger* z roku 1641 (Gleick, 2011, s. 161), i když tam ještě nemůžeme mluvit o váženém průměru logaritmů frekvencí, tím méně o entropii nějak anachronicky spjaté s entropií fyzikální. Jak již bylo zmíněno, společně s Goodem (1982) bych indexy po jejich tvůrcích raději nepojmenovával.

[23] Přesněji řečeno, nestacionární a neergodický. Lingvista nechť si představí smývání hranice mezi synchronicitou a diachronicitou. Synchronní popis jazyka je zjednodušením — podobně jako si ve fyzice pomáháme okřídleným „tření zanedbejte“, i toto zjednodušení je často plauzibilní pro cíl, kterého chceme dosáhnout. Pojem *synchronní korpus* je vlastně metonymickou zkratkou pro *korpus vhodný pro vytváření popisu jazyka, jenž zanedbává jeho dynamiku*.

[24] Jeden z posledních pokusů je Zhangův odhad entropie (Zhang, 2012). Shi – Lei (2022) sice tvrdí, že konverguje tak rychle, že je na velikosti textu nezávislý, to ovšem už z dat, která autoři sami uvádějí, očividně není pravda. Navíc, jak jsem již zmínil a jak ještě připomenu mnohokrát, hledání metriky nezávislé na délce textu je zbytečné, protože existují snadno použitelné techniky, jak kteroukoli metriku této závislosti zbavit (kapitola 2.2).

[25] Ke Kolmogorovově komplexitě (Kolmogorov, 1965) v lingvistickém kontextu nemůžu nedoporučit svou disertaci, kde ji rozebírám (Milička, 2016, str. 10). Přímo komplexitu lexika pak pomocí kolmogorovovské komplexity studoval Patrick Juola, viz například (Juola, 1998) nebo (Juola, 2008).

Úspěšný pokus o určité vyčištění prostoru a systematizaci provedl už v sedmdesátých letech M. O. Hill (1973), který převedl metriky, jež jsem zde zatím popsal, na společnou jednotku (zmiňované efektivní typy) a sjednotil je pod jedním vzorcem (1.8):

$$ {}^{q}D = \left(\sum_{i=1}^{V} p(t_i)^q \right)^{1/(1-q)} \tag{1.8} $$

Když za parametr q v tomto vzorci dosadíme číslo nula, vyjde nám jako výsledek metriky jednoduše prostý počet typů, neboť cokoli na nultou je rovno jedné a součtem takto vzniklých jedniček dostaneme počet typů, který následně umocníme na prvou, respektive $1/(1-0)$, čímž se nic nezmění.

Když se q bude blížit jedničce,[26] situace bude poněkud komplikovanější, nicméně výsledkem bude vzorec pro perplexitu, kterou známe ze vzorce 1.6 (Hill má ve zmiňovaném článku v příloze elegantní důkaz).

Když za q dosadíme dvojku, opět velmi jednoduše dojdeme k převrácené hodnotě pravděpodobnosti opakování, kterou známe ze vzorce 1.5.

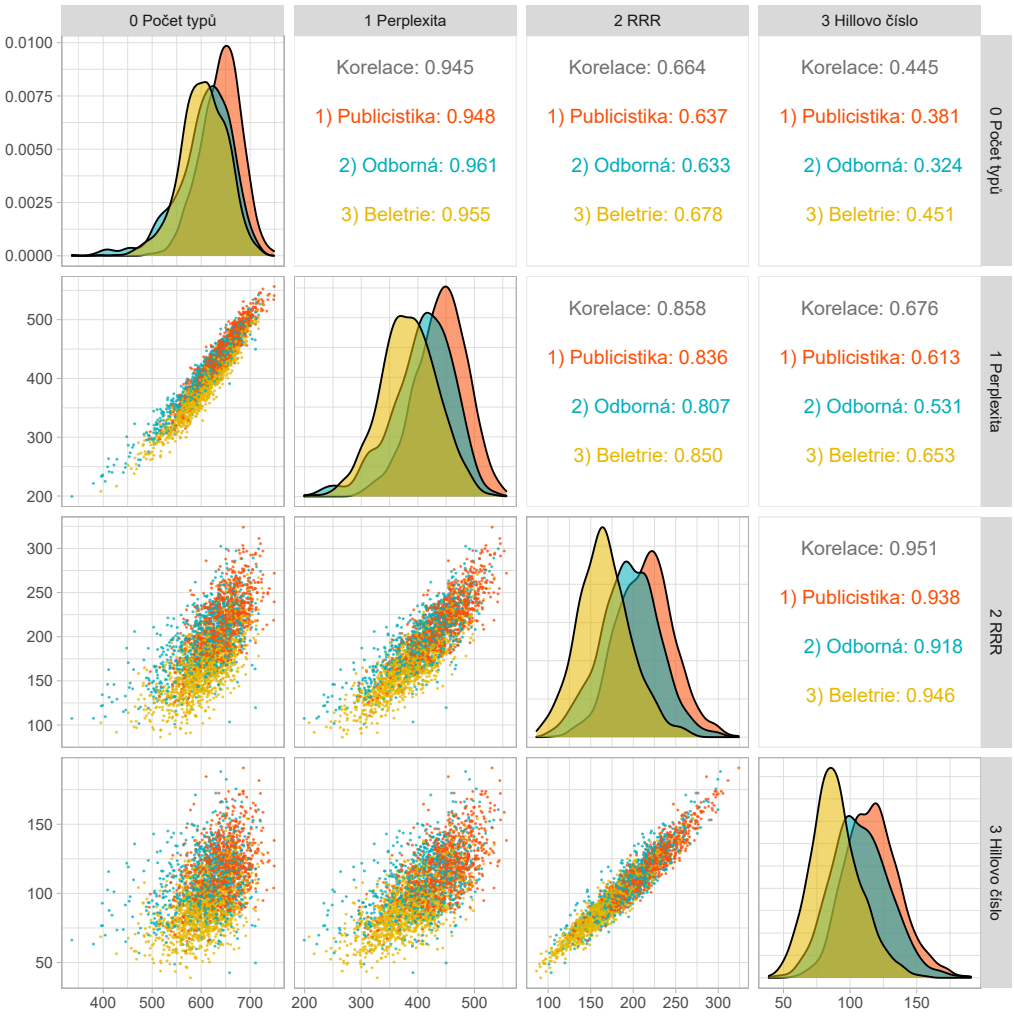
Všimněte si, že čím větší je hodnota q, tím víc mají na výslednou hodnotu vliv typy s větší frekvencí — při nekonečně velkém q by vzorec znamenal jednoduše převrácenou hodnotu relativní frekvence nejčastějšího typu.[27] Mohli bychom tedy jít dál a dosadit i vyšší hodnoty q, než je dvojka, nicméně byla by otázka, jak tyto hodnoty pojmenovat a interpretovat.

Rovněž je docela dobře možné využít i jiných než celočíselných hodnot, tedy například nastavit q jako 1,427 a podobně. Přestože sám autor před něčím takovým varuje,[28] přijde mi škoda nevyužít Hillova kontinua jako skutečného kontinua. V roce 1973, v čase, kdy bylo vlastně docela obtížné každou hodnotu spočítat a ještě k tomu bylo možno prezentovat čtenáři odborného článku data jen staticky a obvykle jen černobíle, samozřejmě dávalo smysl uvést do tabulky jen několik málo pečlivě vybraných hodnot, nicméně dnes nemáme problém vygenerovat pro čtenáře interaktivní graf, kde si může měnit parametr q podle vlastního uvážení. Koneckonců můžeme změřit diverzitu textu pro všechny parametry q z určitého intervalu (řekněme 0 až 3) a vynést ji do grafu, samotná křivka je zajímavá, protože čím je plošší, tím rovnoměrněji jsou druhy zastoupeny. Touto veličinou (tzv. evenness) se v této knize sice nezabývám, nicméně s diverzitou úzce souvisí (Chao et al., 2014, přehledně znázorněno na obrázku 1).

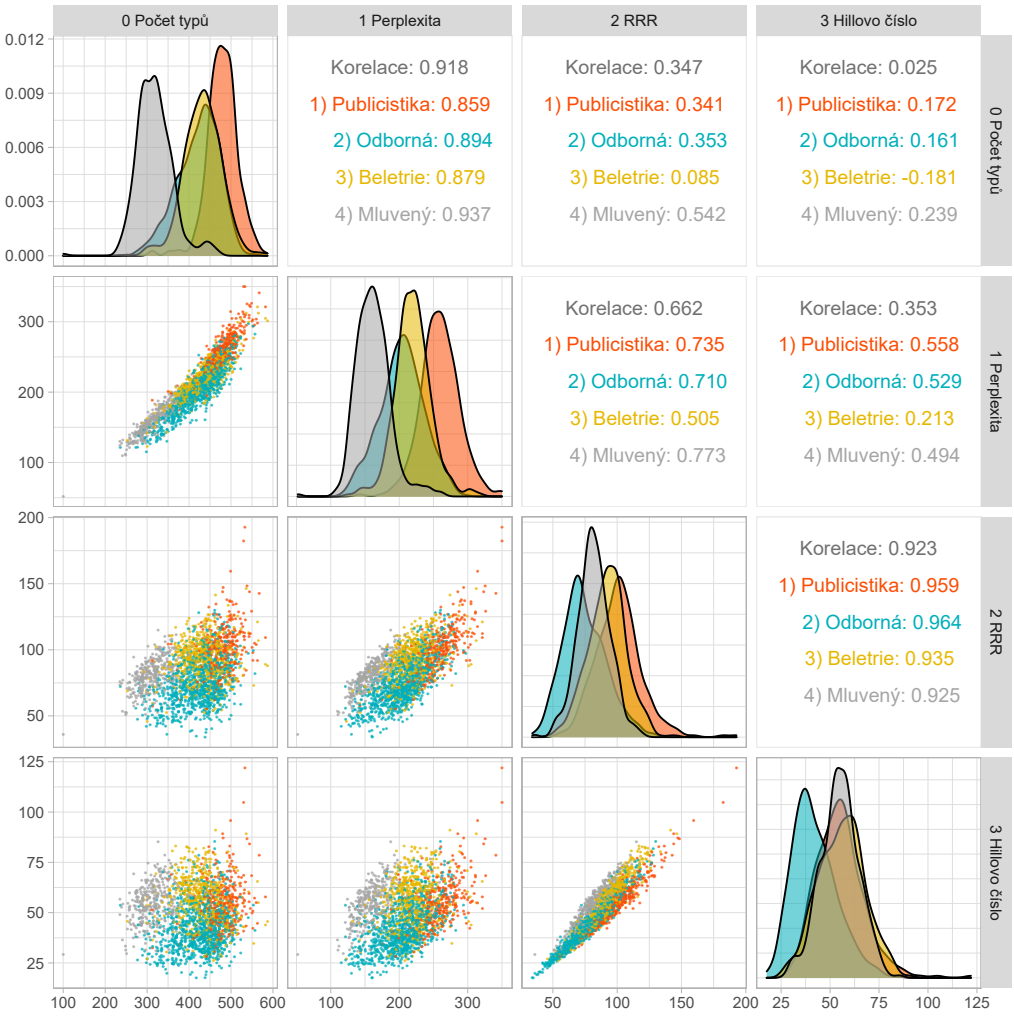
[26] V tomto případě nemůžeme jednoduše dosadit jedničku, neboť v exponentu bychom dostali $1/0$ a nulou nelze dělit, proto využíváme limitní hodnoty.

[27] Obdobně při nekonečně malém q by vzorec vyplivl převrácenou hodnotu relativní frekvence nejméně častého typu. Vzhledem k tomu, že nejméně častý typ je v reálném textu prakticky vždy hapaxem, převrácenou hodnotou jeho relativní frekvence je prostě délka textu, což je nám při měření diverzity úplně k ničemu.

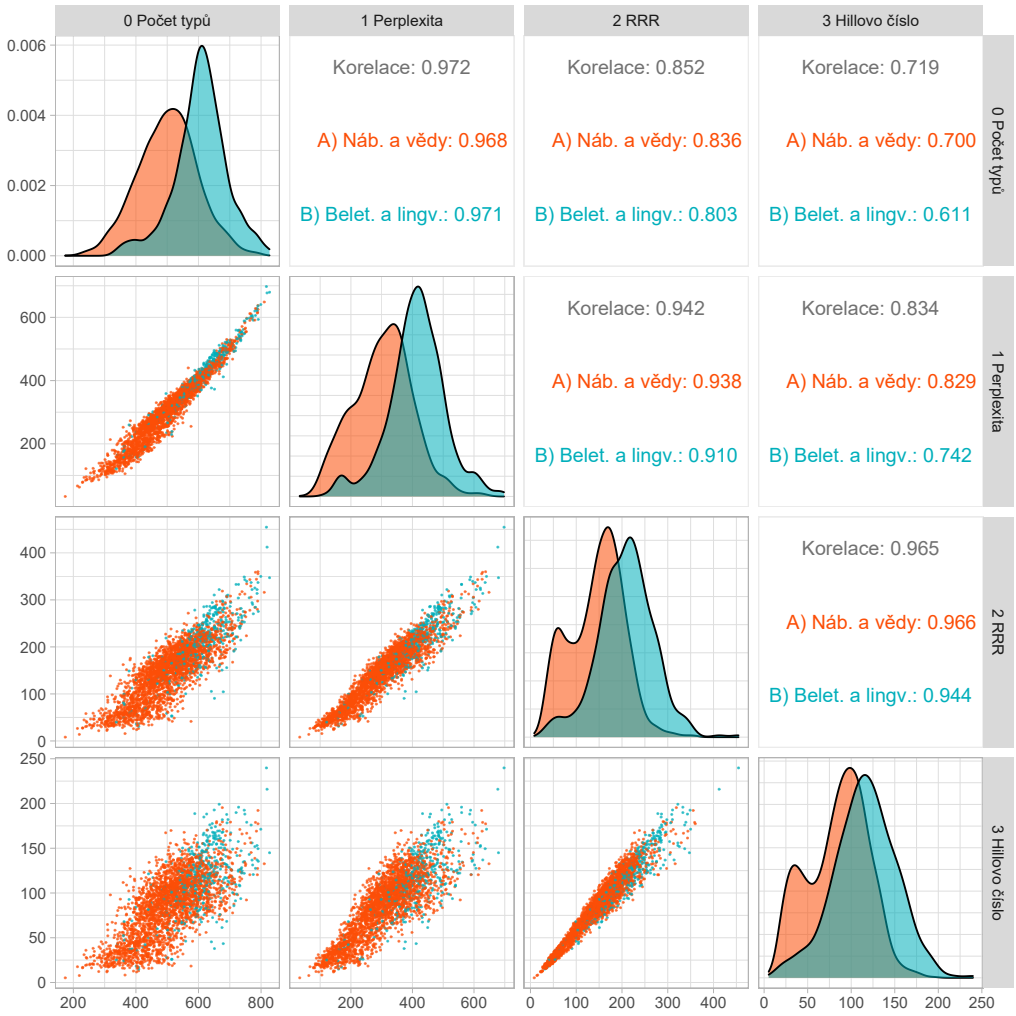
[28] Doslovně zavádí jakousi Hillovu břitvu, jak jinak než latinsky, řka *Indices non sunt multiplicandi praeter necessitatem*.



Obrázek 1.2: Korelogram Hillových čísel o různých hodnotách parametru q, počínaje nulou, konče trojkou (čeština, SYN2015).



Obrázek 1.3: Korelogram Hillových čísel o různých hodnotách parametru q, počínaje nulou, konče trojkou (angličtina, BNC).



Obrázek 1.4: Korelogram Hillových čísel o různých hodnotách parametru q, počínaje nulou, konče trojkou (arabština, CLAUDia).

V ekologické literatuře má pojem Hillovo kontinuum (respektive Hillovo číslo, Hill's number) prominentní místo, ovšem Hill vlastně jen aplikoval koncept o dvanáct let starší Rényiho entropie (Rényi, 1961). Jak vidíte ze vzorce 1.9, od Hillova čísla se liší pouze škálou, která je logaritmická. Její jednotkou tak nejsou efektivní typy, ale bity. Pro $q = 0$ nám vyjde logaritmus počtu typů (tzv. Hartleyova entropie), pro $q \to 1$ získáme vzorec pro Shannonovu entropii atd.

$$^{q}H = \log_2(^{q}D) = \frac{1}{1-q} \log_2 \left(\sum_{i=1}^{V} p(t_i)^q \right) \tag{1.9}$$

Podobně jako Shannonovu entropii v předchozí kapitole, i tuto obecně definovanou entropii je vhodné užít, pokud chceme nějak aproximovat lexikální komplexitu spíše než lexikální diverzitu.

Korelogramy na obrázcích 1.2–1.4 nám ukazují, že kontinuum se chová hezky předvídatelně, tedy korelace mezi metrikami s podobným q jsou menší než korelace mezi metrikami, které jsou v kontinuu vzdálenější (například korelace počtu typů a perplexity vs. korelace počtu typů a převrácené pravděpodobnosti opakování), nejvzdálenější body kontinua, které zde uvádím (počet typů a Hillovo číslo s $q = 3$), pak korelují suverénně nejmíň, v angličtině se dokonce blíží nule. I tady si povšimněme, že dvojice metrik, obzvláště těch v kontinuu vzdálenějších, mají schopnost hezky separovat klastry podle barev, tedy podle typů textů, ze kterých byly vzorky náhodně vybrány. Rozestavení oněch klastrů v prostoru se mění v závislosti na absolutní hodnotě parametru q, tedy na grafu pro dvojici počet typů — perplexita mají texty různých typů a modalit vzájemně trochu jinou pozici než na grafu znázorňujícím dvojici RRR – Hillovo číslo s $q = 3$.

1.6 Metriky srovnávající text s referenčním korpusem

Dejme tomu, že chceme získat hodnotu lexikální diverzity pro nějaké velmi krátké texty. Třeba tweety nebo příspěvky na jiných sociálních sítích, reklamní slogany, politická prohlášení, krátké dialogy — například korpus interakcí se zmrzlinářkou. Je nepravděpodobné, že se v takto krátkých textech bude vůbec nějaké slovo opakovat, a i kdyby, bude to spíš otázka náhody. Tedy pro většinu takových textů je počet typů roven počtu tokenů, ba i počet efektivních typů, ať už je měříme jakkoli, se bude blížit počtu tokenů.

Přesto i u takto krátkých textů dokážeme intuitivně odhadnout, který autor bude mít bohatší slovní zásobu — totiž ten, kdo používá obecně méně častá slova.

Z opačného úhlu pohledu, v předchozích kapitolách popisované metriky lexikální diverzity se vůbec neptají, jestli slova, kterými je krmíme, jsou vlastně vůbec slova. Například pacient s Wernickeovou parafázií se bude vyznačovat projevem, který ony

metriky budou považovat za lexikálně velmi bohatý, přitom se může jednat spíše o kolekci fragmentů slov beze smyslu a pseudoslov (Cho et al., 2021).

K tomu, abychom určili, která slova jsou málo častá, která hodně a která nejspíš slovy vůbec nejsou, potřebujeme referenční korpus, jehož prizmatem se na dané texty budeme dívat. Mezi žánry a modalitami nalezneme obrovské rozdíly, a jelikož „vyvážený korpus“ je svatým grálem korpusové lingvistiky ve všech smyslech toho slova, nutně budeme stát před otázkou, který referenční korpus vybrat. Osobně si nemyslím, že by to byl nějaký nepřekonatelný problém, ostatně můžeme referenčních korpusů vyzkoušet víc a právě z rozdílů mezi nimi může leccos vyplynout.[29]

Metrika, která bere v potaz nějaký referenční rámec, diskurz, ve kterém text figuruje, může také lépe odpovědět na otázku, jak náročný je daný text na percepci, neboť i texty, které mají vnitřní lexikální diverzitu malou (opakují stále stejné typy), mohou být na čtení velmi obtížné, pokud jsou ony typy pro čtenáře vzácné a málo používané.

1.6.1 Křížové verze metrik

Poněkud anachronicky začněme s křížovou verzí Hillova kontinua a Rényiho entropie, které známe z předchozí podkapitoly. Tyto vzorce hezky systematizují různé metriky tak, aby jednotně dokázaly vzít v potaz referenční korpus, ve skutečnosti byl ovšem vývoj křížových verzí jednotlivých metrik podobně chaotický jako vývoj jejich nekřížových variant.

Na křížovou variantu Hillova kontinua jsem v literatuře nenarazil, pročež jsem strávil hezké dopoledne tím, že jsem ji odvozoval pomocí metody, kterou Hill popisuje v příloze svého nejslavnějšího článku (Hill, 1973). Podařilo se, viz vzorec 1.10, kde $p(t)$ značí relativní frekvenci daného typu v našem zkoumaném textu a $p(r)$ relativní frekvenci daného typu v referenčním korpusu. Všimněte si, že pokud mají typy ve zkoumaném textu nachlup stejnou relativní frekvenci jako typy v referenčním korpusu, vzorec je ekvivalentní své nekřížové variantě.

$$ {}^{q}D = \left(\sum_{i=1}^{V} p(t_i) p(r_i)^{q-1} \right)^{1/(1-q)} \tag{1.10} $$

Následně jsem zjistil, že k ekvivalentnímu vzorci došel před šedesáti lety přede mnou Rényi (1961). Musím uznat, že díky tomu mám ke křížové Rényiho entropii (vzorec 1.11) mnohem lepší vztah, než kdybych si o ní jen přečetl. Je možné, že v literatuře narazíte i na jiné vzorce pro křížovou Rényiho entropii, nicméně tyto vzorce jednak nefungují jako zobecnění pro klasické metriky lexikální diverzity, jednak mají se samotným Rényim pramálo společného.

[29] Jelikož v této knize pracuji s texty a vzorky vybranými z nějakého konkrétního korpusu, jako referenční korpus systematicky používám přímo korpus, ze kterého je takový vzorek vzat. Od tohoto korpusu ovšem odčítám frekvence slov ve vzorku.

$$ {}^{q}H = \log_2({}^{q}D) = \frac{1}{1-q} \log_2 \left(\sum_{i=1}^{V} p(t_i) p(r_i)^{q-1} \right) \tag{1.11} $$

Opět platí, že diverzitu měřenou pomocí křížového Hillova kontinua vyjadřujeme v efektivních typech, zatímco Rényiho křížovou entropii v bitech.

Z křížového Hillova kontinua můžeme separovat metriky, které nás zajímají.

Dosadíme-li tak do vzorce 1.10 parametr q roven dvěma, dostaneme křížovou převrácenou pravděpodobnost opakování (reversed cross repeat rate, vzorec 1.12), která je známější ve své zlogaritmované variantě jakožto *collision entropy*. Všimněte si, že tato metrika je symetrická, tedy je jedno, který z textů či korpusů bereme jako zkoumaný a který jako referenční. Samotná křížová pravděpodobnost opakování má velmi jednoduchou intuitivní interpretaci: ze zkoumaného textu vytáhneme náhodně jeden token, z referenčního textu či korpusu také náhodně vybereme jeden token. Jaká je pravděpodobnost, že oba tokeny budou patřit ke stejnému typu? Díky této jednoduché interpretaci je to opět jedna z nejstarších metrik podobnosti dvou souborů dat, používá se už ve zmiňovaných pracech Lexise (1879) a Keynesse (1921, s. 398–399).

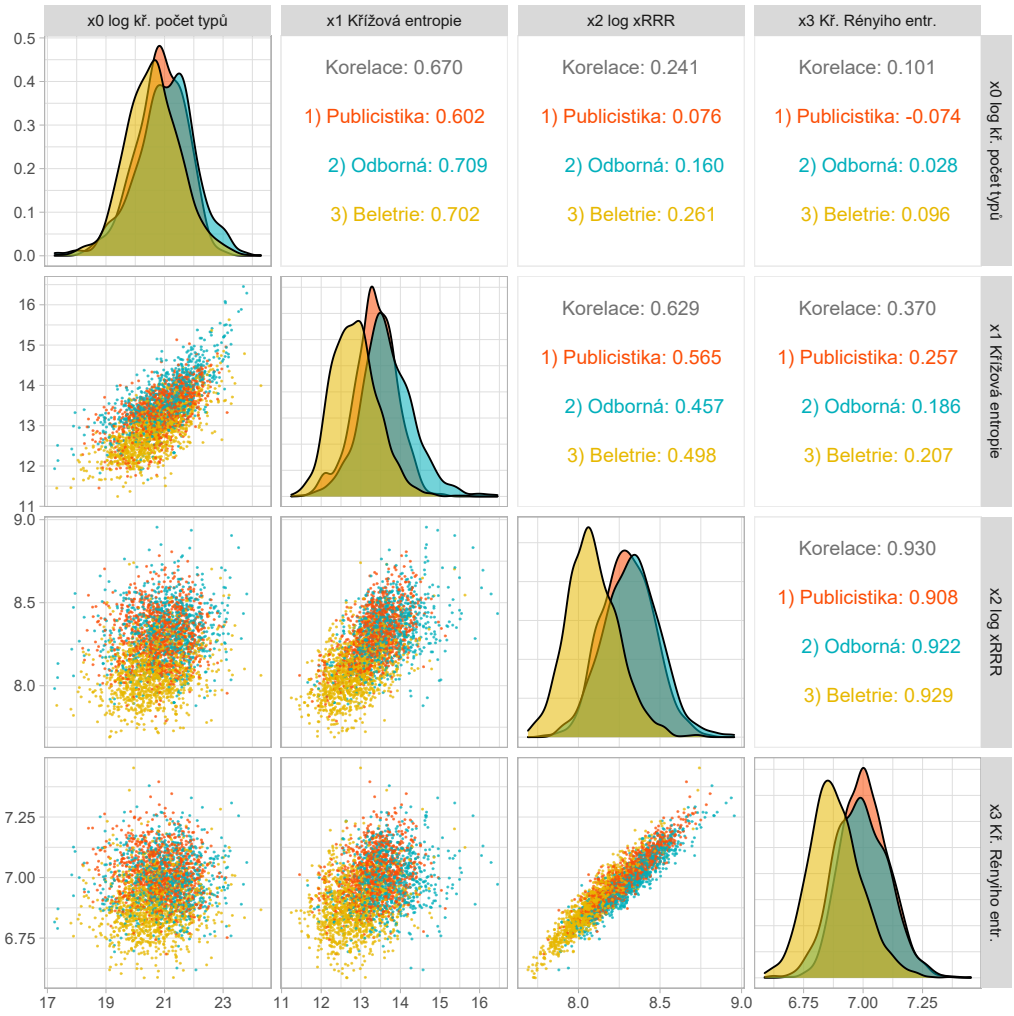
$$ {}^{2}D = RRR(T, R) = \frac{1}{\sum_{i=1}^{V_T} p(t_i) p(r_i)} \tag{1.12} $$

Pro $q \to 1$ získáme křížovou perplexitu (vzorec 1.13),[30] pojem opět dobře ukotvený v rámci Shannonovy teorie komunikace. Když totiž tento vzorec zlogaritmujeme, získáme klasickou Shannonovu křížovou entropii, číslo, které udává, kolik bitů informace bychom potřebovali průměrně k zapsání jednoho slova z měřeného textu, kdybychom ho kódovali kódováním vyvinutým tak, aby co nejefektivněji zakódovalo referenční korpus. Můžeme tedy křížovou perplexitu interpretovat tak, že ukazuje, jak moc překvapivý je měřený text, když se na něj podíváme prizmatem referenčního korpusu.

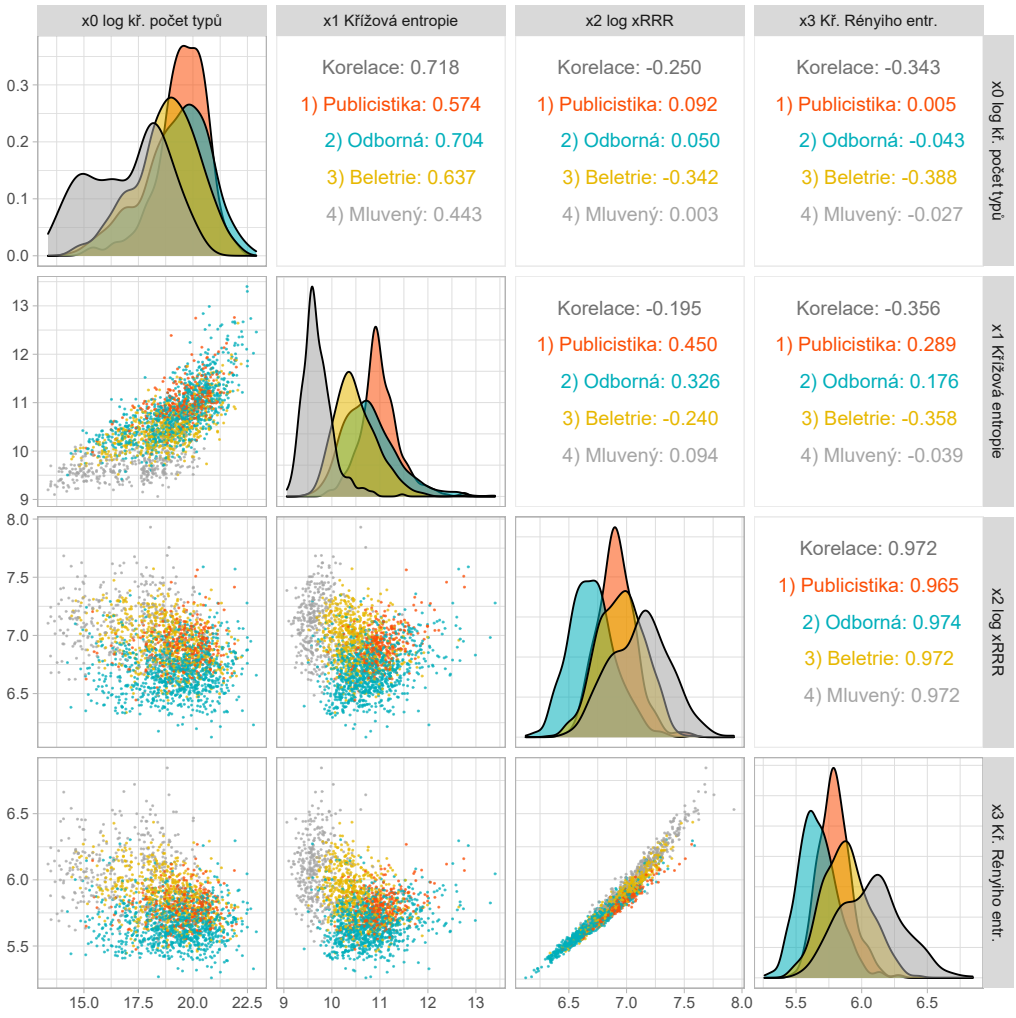
Všimněte si, že pokud křížovou entropii či perplexitu chápeme jako míru rozdílnosti měřeného a referenčního korpusu, tak není symetrická, tedy že záleží na tom, který korpus srovnáváme s kterým. Podle matematického názvosloví tak křížovou entropii a perplexitu při takovémto užití nemůžeme označit za *metriku*, neboť nesplňuje axiom symetričnosti. Avšak pokud tuto míru užijeme k porovnávání lexikálních diverzit dvou textů za užití stejného referenčního korpusu, tak se o metriku jedná. Další terminologické otázky rozebírám v kapitole 6.4.

Nahlédnutím do vzorce si všimnete, že nastane problém, když se v měřeném textu vyskytnou slova, která v referenčním korpusu nejsou. Perplexita pak je rovna nule

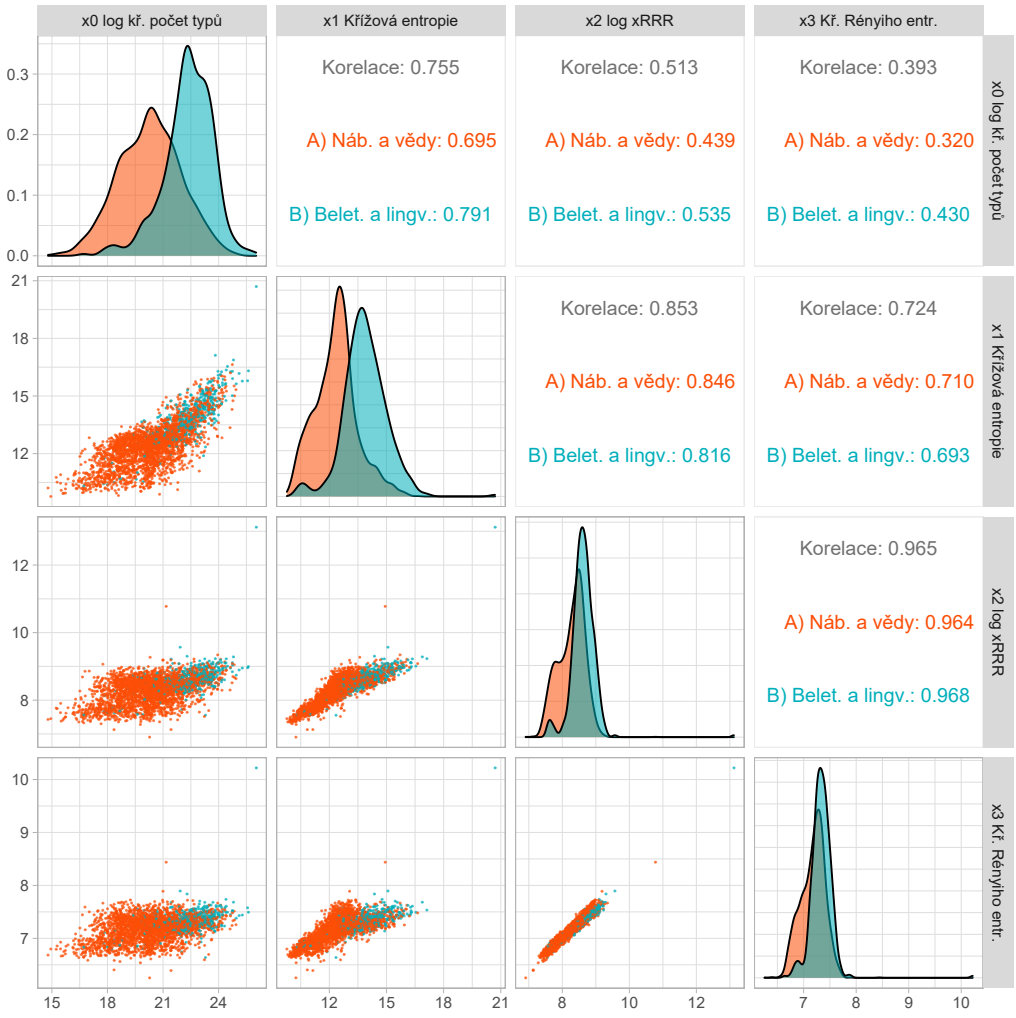
[30] Opět nemůžeme říct $q = 1$, protože pak bychom dělili nulou, tedy pracujeme s limitou. Aby terminologie nebyla příliš nudná a předvídatelná, občas se v literatuře křížová perplexita označuje jako perplexita bez přívlastků.



Obrázek 1.5: Korelogram Rényiho entropií různých hodnot parametru q, počínaje nulou, konče trojkou (čeština, SYN2015).



Obrázek 1.6: Korelogram Rényiho entropií různých hodnot parametru q, počínaje nulou, konče trojkou (angličtina, BNC).



Obrázek 1.7: Korelogram Rényiho entropií různých hodnot parametru q, počínaje nulou, konče trojkou (arabština, CLAUDia).

a křížová entropie se limitně blíží nekonečnu. Tento výsledek sice velmi dobře reflektuje definici křížové entropie i způsob, jakým ji interpretujeme, nicméně metrika pak selhává, protože při skutečném měření se tohle stává prakticky pořád. Existuje spousta technik, jak se s tímto problémem vypořádat, asi nejjednodušší je přidat ke všem absolutním frekvencím v referenčním korpusu nějakou konstantu (metoda se obvykle uvádí pod názvem add-k smoothing nebo Laplace smoothing).[31] Ve všech měřeních v této knize přidávám $k = 0{,}5$. S tímto problémem se musíme vypořádat pro všechna q menší než jedna.

$$^{1}D = P(T, R) = \prod_{i=1}^{V_T} (p(r_i))^{-p(t_i)} \tag{1.13}$$

Pro $q = 0$ získáme křížový počet typů, který se po zlogaritmování promění v Hartleyho entropii. Nezaznamenal jsem, že by se tato metrika v lingvistice v praxi používala, nicméně není důvod, proč by se použít nedala. I zde je nutno nějak řešit situace, kdy se typ nevyskytuje v referenčním korpusu, abychom nedělili nulou.

$$^{0}D = V(T, R) = \sum_{i=1}^{V_T} \frac{p(t_i)}{p(r_i)} \tag{1.14}$$

Podobně jako v případě Hillova kontinua, i tady si ukážeme korelogramy (na obrázcích 1.5–1.7). Tentokrát ovšem využijeme nikoli křížové Hillovo kontinuum, ale jeho logaritmizovanou variantu, tedy křížové Rényiho entropie (z důvodů popsaných v kapitole 3, která je věnována právě otázce škálování). Za referenční korpus použijeme zbytek korpusu, ze kterého daný vzorek pochází.

Opět platí, že dvojice metrik se sousední hodnotou parametru q korelují obecně lépe než dvojice metrik, jež se v tomto parametru liší více, ovšem celkový obrázek vypadá trochu víc chaoticky, neboť dvojice s parametry $q = 2$ a $q = 3$ mají vysokou korelaci, zatímco ostatní dvojice mají korelaci mnohem nižší, a to i ve srovnání s Hillovým kontinuem. U angličtiny dokonce dosahuje záporných hodnot, což ovšem může být dáno heterogenním složením BNC.

[31] Striktně řečeno, Laplace smoothing popisuje situaci, kdy $k = 1$, ovšem v praxi se používá konstanta obvykle menší. Důvod, proč prosté přidání jedničky funguje, je popsán v klasické Lidstoneově práci (Lidstone, 1920), nicméně Lidstone, Laplace a tím méně Bayes samozřejmě netušili nic o křížové entropii.

1.6.2 Relativní lexikální diverzita Kullback–Leiblerova divergence

Můžeme zajít ještě dál a plně se soustředit na rozdíl mezi měřeným textem a referenčním korpusem. V takovém případě vyjdeme z relativní entropie, zvané též Kullback-Leiblerova divergence (Kullback – Leibler, 1951), ať už konkrétně shannonovské entropie, nebo obecné entropie Rényiho (vzorec 1.16).

$$H_{rel}(T, R) = H(T, R) - H(T) \tag{1.15}$$

Jak již víme z předchozích kapitol, Rényiho entropie není nic jiného než logaritmizované Hillovo kontinuum metrik diverzity, tedy snadno můžeme odvodit relativní lexikální diverzitu pro všechny metriky z tohoto kontinua (počet typů, perplexita, převrácená pravděpodobnost opakování a všechno mezi tím). Z odčítání se tak stane dělení (vzorec 1.16), čímž jdeme vstříc běžné představě toho, co znamená relativní metrika (podobně jako je relativní frekvence).

$$D_{rel}(T, R) = \frac{D(T, R)}{D(T)} \tag{1.16}$$

1.7 Průměrná délka slov

Jazyk je poměrně efektivní komunikační prostředek. Asi těžko můžeme říct, že je dokonale efektivní, ale rozhodně dost na to, aby zhruba platilo, že čím je slovo překvapivější, tím delší průměrně je (Piantadosi et al., 2011). Je to nejspíš dvojsečná zbraň, tedy čím lépe může recipient dané slovo v jeho typickém kontextu predikovat a čím častější slovo je, tím větší je pravděpodobnost, že se během vývoje jazyka zkrátí nebo že jeho místo zaujme jeho kratší synonymum. A naopak, čím kratší slovo je, tím častěji ho produktor bude dávat do kontextů, kde nezpůsobí přílišné překvapení. A tím častěji ho bude používat, často používaná slova totiž nejsou tak překvapivá jako slova vzácná.

Frekvence ovšemže není jediný faktor, který má vliv na to, jestli může recipient snadno uhodnout, že dané slovo se vyskytne v daném kontextu, nicméně je to faktor silný, neboť samotný vztah frekvence slova a jeho délky je velmi výrazný, tak výrazný, že si ho všiml už G. A. Zipf (Zipf, 1935; Bentz – Ferrer-i Cancho, 2016).

Průměrná délka slova v textu je tedy něco jako křížová entropie, kde referenční text je jakýsi ultimátní všeobjímající korpus, který působil na vývoj daného slova a jemuž se délka daného slova přizpůsobovala.

Slova v přirozeném jazyce ovšem nefungují jako ideální kódy, řekněme ve smyslu Huffmanova kódování (Huffman, 1952), ale poněkud primitivněji (non-singular coding, viz Ferrer-i Cancho et al. (2022)). Délka slova proto lineárně škáluje s logaritmem *pořadí* daného slova, seřadíme-li je sestupně podle jejich frekvence. Respektive,

abychom byli přesní, pokud pořadí daného typu označíme jako i, pak ideální délka slova daného pořadí zakódovaná pomocí abecedy o A znacích se počítá podle vzorce 1.17 (Ferrer-i Cancho et al., 2022, str. 15):

$$L(t_i) = \left\lceil \log_A \left(\left(1 - \frac{1}{A} \right) i + 1 \right) \right\rceil \tag{1.17}$$

Tento vzorec vypadá trochu zmateně, ale ve skutečnosti se nejedná o nic složitého: pokud seřadíme slovní typy v textu či korpusu sestupně podle jejich frekvence a chceme je zakódovat abecedou o šestadvaceti znacích, tak na prvních 26 slov nám stačí jeden znak, na dalších 26^2 potřebujeme dva znaky atd. Takže vzorec pro křížovou entropii bychom měli podle této teorie upravit na vzorec 1.18, pokud chceme, aby lépe korelovala s průměrnou délkou slova:

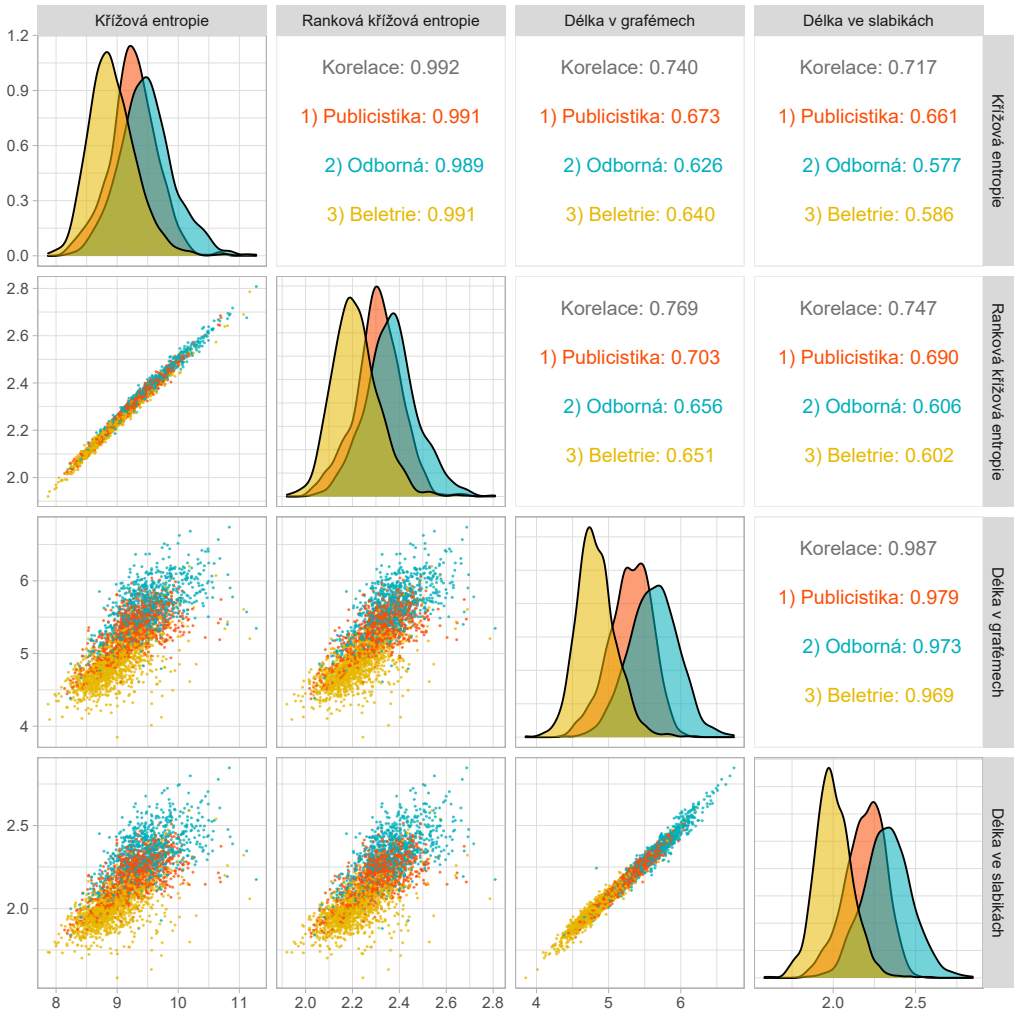
$$H_r(T, R) = \sum_{i=1}^{V_T} p(t_i) L(t_i) \tag{1.18}$$

Jak vidíme na obrázcích 1.8 a 1.9, Shannonova křížová entropie a naše nová rankova křížová entropie spolu velmi dobře korelují, takže mezi nimi není zas až takový rozdíl. Průměrná délka slova s rankovou entropií koreluje o trochu líp než s entropií Shannonovou, ovšem rozdíl není právě velký, takže je otázka, zda to vůbec stálo za tu námahu.

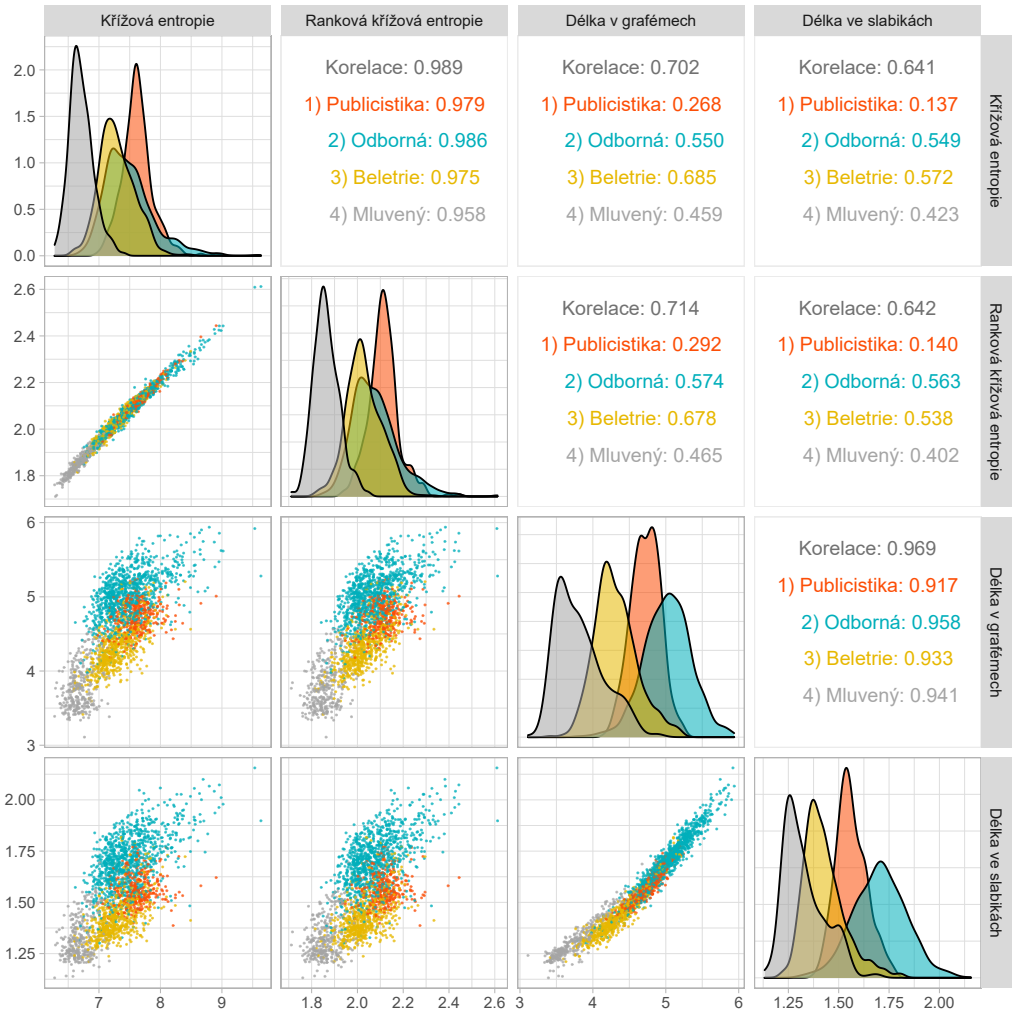
Nedokonalou lineární korelaci mezi křížovou entropií a průměrnou délkou slova můžeme přičíst nejméně třem fenoménům: 1) náš referenční korpus je příliš vzdálen onomu ultimátnímu korpusu, na němž se tříbilo kódování slov; 2) kódování v jazyce není efektivní, pouze *dostatečně* efektivní; 3) naše teorie nepostihuje dobře, co pro lidský kognitivní aparát skutečná efektivnost kódování znamená. Ovšem je otázka, jestli nás má ona nedokonalost příliš trápit, neboť pokud se podíváme na korelaci průměrné délky slova s dalšími metrikami z Hillova kontinua a jeho křížové varianty, tak křížová entropie v souladu s teorií suverénně vede (viz kapitolu 6.2).

V každém případě kombinace křížové entropie a průměrné délky slova docela obstojně separuje textové typy. Jak v angličtině, tak v češtině mají sekvence vybrané z odborných textů poměrně delší slova než sekvence o stejné křížové entropii vybrané z textů beletristických či publicistických. Tento jev bychom mohli interpretovat tak, že v odborných textech je na lexikální úrovni více redundance, popřípadě bychom poněkud odvážně mohli tvrdit, že kódování lexika v odborných textech není tak efektivní, což by souznělo s tím, že se jazyk vyvíjel zejména na jiných textových typech a jeho exaptace na funkci komunikace odborných textů není právě ideální.

I bez velkého teoretizování je průměrná délka slova jako index lexikální diverzity intuitivně používána. Respektive není nutně označena přímo jako index lexikální diverzity, ale je často součástí různých systémů, které mají za úkol odhadnout čitelnost



Obrázek 1.8: Korelogram různých operacionalizací délky a křížové entropie (čeština, SYN2015).



Obrázek 1.9: Korelogram různých operacionalizací délky a křížové entropie (angličtina, BNC).

textu (readability), vyspělost slovní zásoby studentů a podobně.[32] Rovněž se k těmto účelům používá podíl slov s velkým počtem slabik (polysyllables), který s průměrnou délkou slov úzce souvisí.

V angličtině k oné intuici přispívá určité povědomí o diachronní perspektivě, neboť dlouhá slova jsou obvykle normanského původu, a tedy patří do novější slovní zásoby, navíc byla hojněji používána vyššími společenskými vrstvami a intelektuály.

Podobným diachronním vývojem si ovšem prošly i jiné jazyky, jmenovitě byla zkoumána angličtina, čínština a arabština (Chen – Liu, 2014; Chen et al., 2015; Milička, 2018), a u všech byla zjištěna tendence, že s přibývajícími staletími přibývala slova delší a delší.

Při použití této metriky odpadá problém s typizací, tedy jak určit, které slovo patří k jakému typu. Ovšem problému s tokenizací jsme se nezbavili, naopak musíme text segmentovat na dvou úrovních: na úrovni slova, stejně jako ostatní metriky, a dále délku onoho slova musíme změřit v nějakých jednotkách, například v čase (u mluveného jazyka) nebo v nějakých jeho subsegmentech. Výběr oněch subsegmentů je nutně arbitrární, byť můžeme argumentovat, že se z nějakého důvodu pro určité účely hodí určité jednotky. Například můžeme měřit délku slova ve slabikách, čímž se do určité míry přiblížíme tomu, jak velký prostor slovo zaujímá v mluveném jazyce, pročež je tento způsob v kvantitativní lingvistice oblíbený — ovšem opravdu má slovo „máta“ dvojnásobnou délku slova „lusk“? Můžeme tedy použít hlásky, nějaké formy distinktivních rysů, morfémy… ovšem všechny tyto možnosti před nás kladou další a další volby. Když se vrátíme k obrázkům 1.8 a 1.9, vidíme, že průměrná délka slov měřená ve slabikách velmi dobře koreluje s průměrnou délkou měřenou v grafémech, a to i v angličtině,[33] takže s klidným svědomím můžeme, pokud nemáme lepší řešení, použít jako jednotku délky slov prostě počet písmen.

V každém případě ale odpadá otázka, jestli použít lemmatizovaný text. Slovní formy, které dané lemma v lemmatizovaném textu reprezentují, jsou totiž více méně arbitrárně zvoleny, a přestože jsou obvykle interpretovány jako „základní tvar“ daného paradigmatu, ona *základnost* je dána tradicí a náhodou, nikoli délkou, komplexitou nebo frekvencí.[34] Například u angličtiny bychom mohli argumentovat, že jde obvykle o nejjednodušší a nejkratší tvar daného slova, ovšem ve flektivnějších jazycích, tedy v těch, kde na volbě více záleží, to tak obvykle není.

[32] Například hojně používané, byť tak trochu numerologicky konstruované, *flesch reading ease score* a jeho varianty.

[33] Což je s podivem, neboť „sa anglické písmo natoľko rozchádza s akustickou rečou, že môžeme hovoriť o dvoch paralelných, len čiastočne korelujúcich kódoch“ (Krupa – Genzor, 1989, str. 30).

[34] První pád podstatného jména obvykle není ten nejvíce užívaný, infinitiv slovesa už vůbec ne, není ani nejkratším tvarem a v klasických slovnících mnoha jazyků se jakožto základní tvar používá to, co bychom označili za třetí osobu indikativu (Kováříková et al., 2020).

1.8 Rozdílnost (dissimilarity)

Hillovo kontinuum, respektive Rényiho entropie, docela dobře fungují jako syntéza všemožných smysluplných metrik, nicméně jedná se o syntézu nutně neúplnou, neboť bere v potaz jen metriky, které jsou závislé pouze na počtu typů a jejich frekvencích. Tyto metriky se neptají, jak jsme k oněm typům přišli a jak vlastně vypadají. Přitom intuitivně cítíme, že pokud se v textu opakují slova odvozená od stejného základu, bude tak nějak lexikálně méně diverzifikován než text, ve kterém je každý typ velmi nepodobný ostatním typům.

V případě, že nás tento problém skutečně trápí, v lingvistice bych neváhal a použil diverzitu menších segmentů, než je lexikon, tedy například diverzitu morfematickou,[35] nebo jednoduše a mechanicky diverzitu písmenných n-gramů. Ovšem v biologii nic takového provést nemůžeme, těžko bychom rozkládali zvířátka na trigramy a morfologie rostlin je poněkud nepodobná morfologii slov. Vznikly a vznikají proto metriky, které berou v potaz, že některé typy jsou si podobnější než jiné, proto jim říkáme metriky rozdílnosti (*dissimilarity*).

Metriky rozdílnosti se tedy v lingvistice zatím příliš neužívají (Jarvis, 2013), mohly by ale pomoci i zde, zejména tam, kde si nejsme jistí typizací. Binární typizaci totiž nahrazují nějakou spojitou funkcí, kterou si můžeme zvolit podle vlastního uvážení. Ona funkce se může týkat vnější podobnosti, kde použijeme nějakou formu editační vzdálenosti, třeba klasickou Levenshteinovu vzdálenost (1965; 1966), nejdelší společný nepřerušený řetězec, nebo nejdelší společnou sekvenci.[36] Nebo může jít o vzdálenost více spjatou s preferovanou lingvistickou teorií, například distance derivační — kolik přípon a předpon musíme změnit, abychom ze slova A udělali slovo B. Popřípadě se ona funkce může týkat sémantiky, ať už vyjádřené kvalitativně a ručně, například vzdáleností ve Wordnetu (Miller et al., 1990), nebo moderněji pomocí vektorů (Mikolov et al., 2013).[37]

Pro typy $\mathtt{i}$ a $\mathtt{j}$ tedy získáme nějakou distanční funkci $\mathtt{d}_{\mathtt{i},\mathtt{j}}$, kterou dosadíme do

[35] Pro češtinu můžete k automatické segmentaci použít skvělou morfematickou databázi od Pelegrinové et al. (2021).

[36] Nejdelší společná sekvence může být i přerušovaná, což se hodí zejména pro jazyky s nonkonkatenativní morfologií. Ve skutečnosti ale chceme metriky rozdílnosti, nikoli podobnosti, tedy u nejdelších společných řetězců i sekvencí nás zajímá jejich převrácená hodnota. Tyto metriky, včetně Levenshteinovy distance, je třeba normovat na délku oněch dvou slov, neboť pokud máme dvě slova o dvou písmenech, která se liší v obou písmenech, tak si zrovna moc podobná nejsou, zatímco dvě desetipísmenná slova, která se liší také jen ve dvou písmenech, jsou si podobná docela dost, přitom editační vzdálenost je stejná. Metod pro normalizaci je celá řada, používám (Yujian – Bo, 2007).

[37] Měření rozdílnosti pomocí vektorové reprezentace slov si můžete vyzkoušet díky Covingtonovi (ten samý Covington, který má na svědomí MATTR, a o kterém tak ještě v této knize uslyšíte). Jeho software (Covington, 2016), navzdory výborné technické dokumentaci, ovšem není otevřený a není tak úplně jasné, jaký vzorec konkrétně používá. Nicméně vektory jsou přebrány z otevřených zdrojů (Pennington et al., 2014).

vzorce, který udělá vážený průměr všech těchto distancí pro všechny dvojice typů (1.19). Takto aspoň vypadá klasická metrika rozdílnosti, kterou už v osmdesátých letech navrhl Radhakrišna Rao (Rao, 1982) a kterou si oblíbili biologové.[38]

$$Q = \sum_{i,j=1}^{N} p(t_i)p(t_j)d_{i,j} \tag{1.19}$$

I rozdílnost může mít své kontinuum vycházející z Rényiho entropie a Hillova kontinua a je s podivem, že bylo popsáno teprve nedávno (Chao et al., 2014; Rocchini et al., 2021).

Rozdílnost má větší asymptotickou složitost algoritmu než obvyklé metriky diverzity, neboť s počtem typů v textu roste komputační náročnost kvadraticky, protože porovnáváme každý typ s každým, zatímco u ostatních dosud popsaných metrik byla závislost lineární. Když k tomu připočteme relativně pomalé porovnávání slov pomocí Levenshteinovy distance, zjistíme, že při běžném nasazení bývá počítání rozdílnosti řádově pomalejší než výpočet ostatních metrik diverzity, takže je otázka, jestli to vůbec stojí za to.

1.9 Podíl autosémantik

Podíl autosémantik (respektive tokenů, které řadíme k tak či onak definovaným *content words*) je jako jeden z indexů pro profilování slovní zásoby používán minimálně od sedmdesátých let (Ure, 1971), přičemž je označován jako lexikální hustota (*lexical density*).

Podobně jako délka slov ani lexikální hustota není běžně chápána jako index lexikální diverzity, nicméně vzhledem k tomu, že synsémantika jsou průměrně krátká a velmi frekventovaná, tak s lexikální diverzitou a průměrnou délkou slova jednoznačně souvisí — čím víc autosémantik, tím větší lexikální diverzita.

Tato korelace, pravda, nemusí platit vždy, dokážu si představit text, ve kterém se budou hojně opakovat stále stejná autosémantika, takže by nebyl zrovna lexikálně diverzifikován navzdory vysokému podílu autosémantik. Ovšem právě díky tomu má smysl používat tento index společně s dalšími metrikami lexikální diverzity, neboť se tím zvýší šance odhalit něco zajímavého.

Je ještě jeden způsob, jak využít toho, že máme k dispozici korpusy s morfologickým značkováním, a že tedy u každého tokenu můžeme snadno určit jeho slovní druh: můžeme změřit lexikální diverzitu zvlášť pro synsémantika a zvlášť pro autosémantika

[38] Shodou okolností ve stejném roce ekvivalentní koncept publikoval i I. J. Good (1982) v článku, se kterým jsme se na těchto stránkách několikrát setkali. Na to, že je pouhým komentářem k (Patil – Taillie, 1982), je vlastně docela nabitý užitečnými a ve své době objevnými koncepty.

(respektive klidně i pro každý slovní druh zvlášť). Získáme tím další dimenzi lexikální diverzity, další čísla, která nás mohou něčím překvapit.

Předpokládám, že lexikální diverzita autosémantik by se dala dobře použít jako vhodná metrika tematické koncentrace, tedy toho, jestli se v textu pojednává pouze jedno téma, nebo jestli naopak autor přelétá od jednoho tématu k druhému. Kubát a Čech zkoušeli, jak klasická metrika tematické koncentrace (nazvaná příhodně *tematická koncentrace*) koreluje s lexikální diverzitou celého textu, a to s více méně negativním výsledkem (Kubát – Čech (2016); posléze Čech (2016, kapitola 7)). Nejspíš proto, že nízká diverzita témat byla vyvážená vyšší nápaditostí v lexikálním plánu, což by právě, alespoň částečně, mohlo řešit změření diverzity pouze na autosémantikách.

1.10 Další metriky

Zatímco metriky představené v předchozích kapitolách vycházejí z potřeby měřit lexikální diverzitu na různých škálách a s váhou rozloženou na různé aspekty tohoto košatého fenoménu, následující metriky vycházejí z prosté potřeby vypořádat se s problémem závislosti na délce textu. Tento problém rozebírám podrobně v kapitole 2 a navrhuji, že způsob normalizace by měl být obecný a uplatnitelný na všechny metriky, nezávisle na jejich kvalitách. Nicméně tato myšlenka není tak samozřejmá, jak se zdá, a desítky let probíhal boj o to, kdo nalezne *tu správnou* metriku. Svým způsobem se mohlo zdát, že *ta správná metrika lexikální diverzity* se pozná právě podle toho, že je nezávislá na délce textu. Vznikaly tak metriky s hodnotami obtížně interpretovatelnými nebo rovnou zavádějícími, z nichž vybírám pouze malý zlomek.

Následující metriky jsem proto implementoval pouze pro účely této kapitoly a dále se jimi nezabývám.

1.10.1 Poměr typů a tokenů (type-token ratio)

Type token ratio, zkráceně TTR,[39] je asi nejčastější způsob, jakým se dnes v literatuře měří slovní bohatství, respektive lexikální diverzita jako taková, pročež ho není možné v této publikaci opominout, přestože bych to rád udělal. Základní myšlenka je velmi jednoduchá — prostě vydělíme počet typů, jak ho známe z kapitoly 1.1, počtem tokenů (vzorec 1.20).

$$\mathrm{TTR} = \frac{V}{N} \tag{1.20}$$

[39] Tato zkratka bohužel koliduje s TTR, které značí type-token relation, vztah mezi počtem typů a tokenů, což je na rozdíl od type-token ratio velmi užitečný koncept.

Podle mých zkušeností si lidé, kteří se nevěnují kvantitativní nebo korpusové lingvistice, pod pojmem *slovní bohatství* představí právě tuto metriku. Přesto ji vůbec nedoporučuji používat a v této kapitole se pokusím vysvětlit proč.

Hlavním důvodem je, že vyvolává zdání, že se jedná o počet typů normovaný na délku textu, tedy že je možné díky ní porovnávat texty různé délky. V současné literatuře samozřejmě nikde nenajdete přímé vyjádření myšlenky, že TTR je nezávislé na délce textu, tohle tvrzení je neudržitelné minimálně od konce padesátých let, kdy vyšly Somersovy modely (Somers, 1959) a zejména Herdanova *Type-token mathematics* (Herdan, 1960). Nicméně stále potkáte velké množství článků, které tuto metriku používají, *jako by to byla pravda*. Panuje určité přesvědčení, že TTR sice není na délce textu nezávislé, ale jaksi nezávislejší než samotný počet typů. Přitom není ani jasné, jak tuto nezávislost vůbec měřit a porovnávat, a je možné, že TTR je za určitých podmínek více ovlivněno délkou textu než prostý počet typů.

Pojďme se podívat, jak délka vzorku vnáší relativní systematickou chybu do TTR, prostého počtu typů a několika dalších metrik, které byly vymyšleny proto, aby tento typ systematické chyby eliminovaly. Použijeme stejnou metodiku pro určení míry systematické chyby, jakou podrobně popisuji v kapitole 2.1.1, tedy zjednodušeně z korpusu vybereme velké množství vzorků o určité délce a ze stejných míst vzorky o menší velikosti (přičemž respektujeme hranice textů). Následně ze vzorků uděláme dvojice (jeden kratší a jeden delší tak pochází z pozice A, jeden kratší a jeden delší z pozice B) a vzájemně je porovnáme: pokud má kratší vzorek z pozice A menší lexikální diverzitu než kratší vzorek z pozice B a zároveň má kratší vzorek z pozice A větší diverzitu než delší vzorek z pozice B, pak takovou dvojici započítáme jako systematicky chybnou (a vice versa). Samozřejmě „chybné“ dvojice budou vznikat i u metrik lexikální diverzity, které jsou na délce textu z definice nezávislé, například u délky slov. Takovouto metriku tedy budeme brát jako baseline — čím víc se k ní bude zkoumaná metrika blížit, tím méně je ovlivněna délkou textu. V tomto případě budeme porovnávat vzorky vždy o pětinu kratší (respektive o čtvrtinu delší) a o polovinu kratší (respektive dvojnásobně dlouhé).

Z grafů na obrázku 1.10 vidíme, že počty typů a TTR se chovají v různých jazycích různě, nicméně u všech tří zkoumaných platí, že systematická chyba u počtu typů je ze začátku opravdu vysoká a postupně pomalu klesá.[40] Naproti tomu systematická chyba u TTR je od začátku menší, následně velmi rychle klesne — pro texty či vzorky o velikosti řádově tisíce tokenů je na svém minimu, které ovšem ani zdaleka nedosahuje baseline a je několikrát větší. Klesání se tedy velmi rychle zastaví a pak křivka naopak pomalu stoupá, u vzorků, které se liší o polovinu (pravý sloupec), roste docela znatelně. Pro vzorky a texty řádově desetitisíce tokenů dlouhé (běžný evropský román) je sice TTR stále o něco méně chybové než prostý počet typů, ovšem podobně nepoužitelné.

[40] Nenechte se překvapit dalšími metrikami, které na obrázcích najdete, bude o nich řeč později.

Všimněte si, že morfologicky bohatší čeština a arabština favorizuje TTR oproti prostému počtu typů více než morfologicky analytická angličtina, kde křivka vztahu mezi typy a tokeny stoupá mnohem méně strmě a není tak „rovná“ (srovnej obrázky 1.12 a 1.13, které tento vztah znázorňují u jednoho českého a jednoho anglického textu), tedy správněji řečeno, je hůře aproximovatelná lineárním modelem, na kterém metrika TTR stojí. Přesto i v češtině je ona chyba řádově srovnatelná, tudíž použitím TTR se závislosti metriky lexikální diverzity na délce textu rozhodně nezbavíme.

Z grafů na obrázku 1.11 vidíme, že lemmatizací textu si mnoho nepomůžeme, právě naopak, neboť z morfologicky bohaté češtiny lemmatizací uděláme něco jako angličtinu — tedy chybovost prostého počtu typů a TTR se k sobě přiblíží rychleji a více.

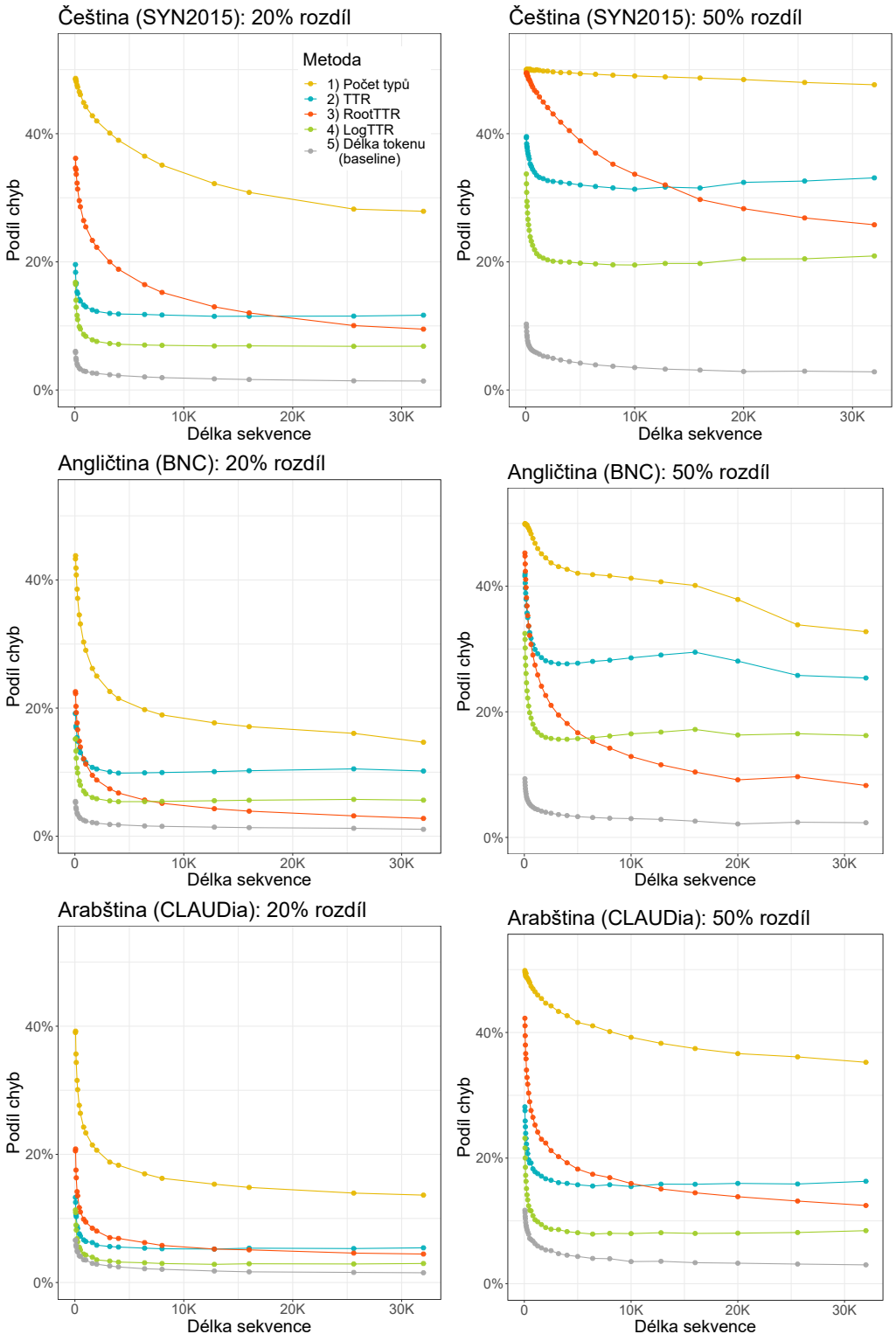
* * *

Když uvážíme výše uvedené, vyvstane nám otázka, odkud oblíbenost TTR vlastně pramení. Pokusím se najít nějaká vysvětlení.

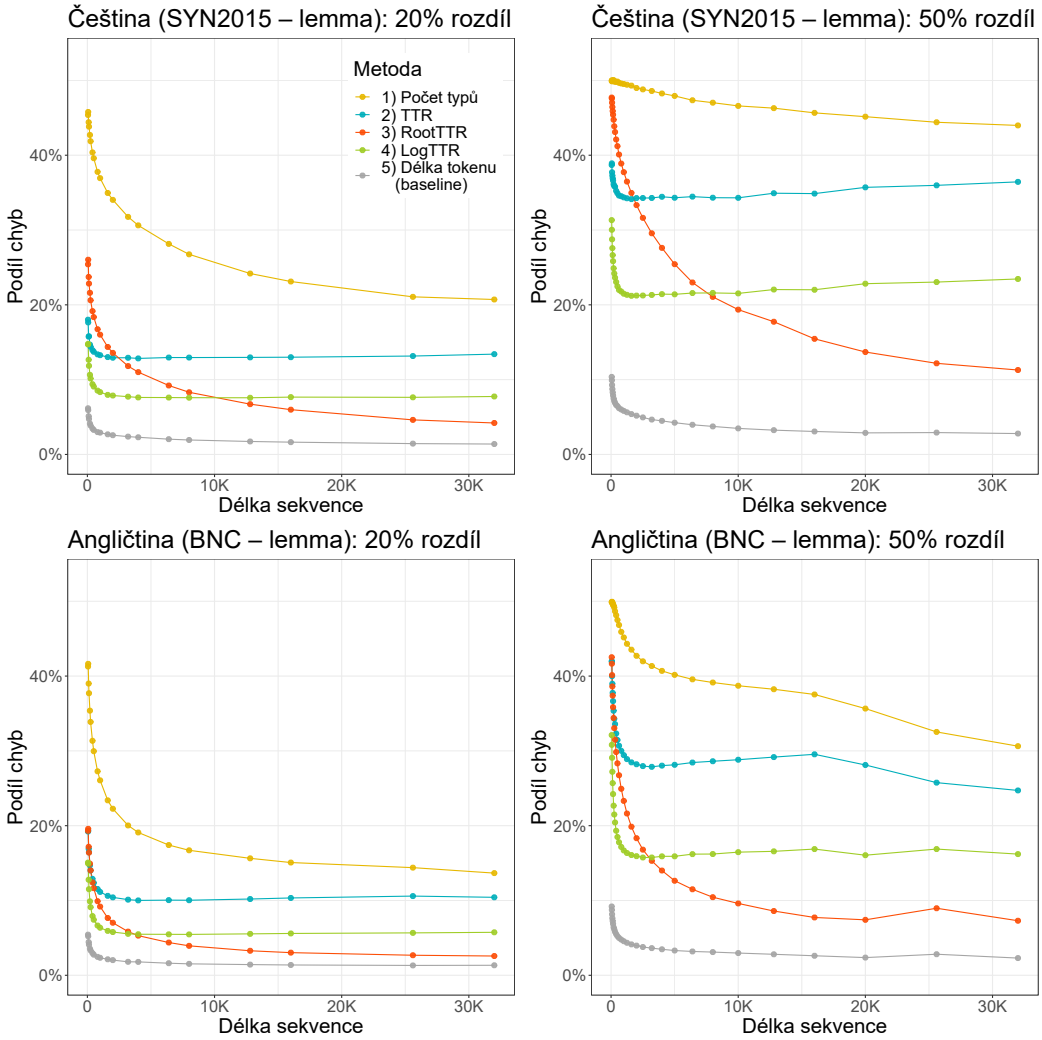
Nenáročné srovnání krátkých textů přibližně stejné délky. Pokud se rozdíl délek dvou textů pohybuje v intervalu, ve kterém křivka vypadá opticky lineárně (srovnej obrázky 1.12 a 1.13 vlevo), může se zdát použití TTR jako dobrý nápad. Dejme tomu, že jsme nechali účastníky nějakého experimentu vytvořit texty v češtině o 10 000 slovech $\pm$ 1 000 slov, pak bias vnesený použitím TTR je okolo tří procent chyb, což je sice zhruba šestinásobek oproti metrice, která je na velikosti vzorku skutečně nezávislá (jako je délka slov nebo podíl autosémantik), nicméně pro některé použití to prostě stačí.

Zásadní problém nastává, pokud někdo bude chtít onen experiment zopakovat a publikovaná data srovnat se svým vlastním souborem textů, který má jiné parametry, například texty se pohybují někde kolem dvou až čtyř tisíc slov. Korpus s takto velkými rozdíly mezi texty bychom přitom v klidu nazvali „korpusem srovnatelně dlouhých textů“.[41] Takto rozdílné texty by ovšem měly systematickou chybu již osmiprocentní — shodou okolností též šestinásobnou oproti baseline metrice skutečně nezávislé na délce textu. Navíc TTR získaný z těchto nových textů se bude pohybovat ve fundamentálně jiných hodnotách, neboť rozdíl mezi dvěma tisíci a deseti tisíci tokeny je už značný. Ovšem hodnoty TTR budou nesrovnatelné s TTR z předchozího výzkumu i z jiných důvodů: přestože se metrika jmenuje stejně, vlastně se jedná o dvě různé metriky, které měří něco trochu jiného, neboť měřením počtu typů na delších sekvencích se projevují jiné efekty než na sekvencích krátkých (viz kapitolu 5).

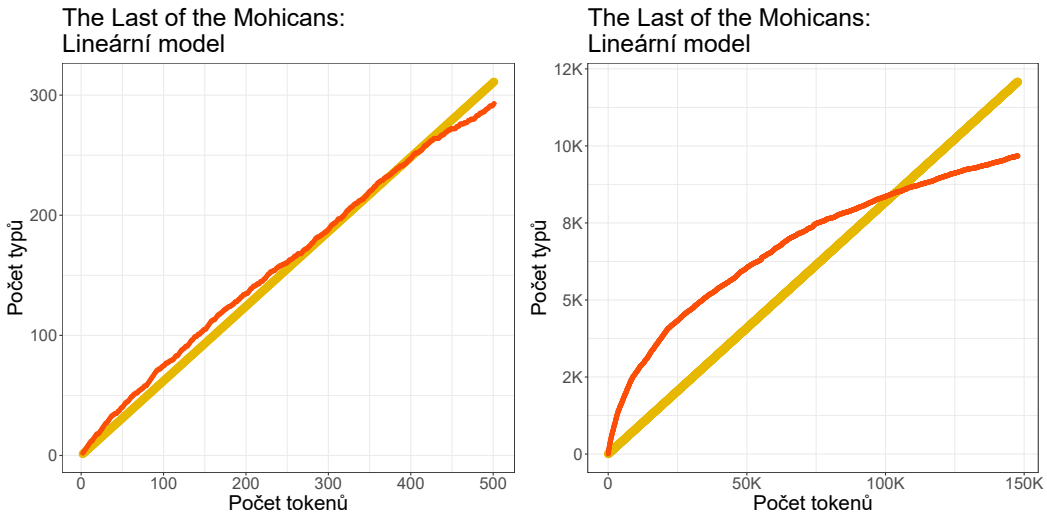
[41] Například v žákovském korpusu SKRIPT2015 (popsán podrobněji v příloze) jsou soubory textů, z nichž některé vznikaly v prakticky totožných podmínkách: žáci měli stejné zadání, stejné téma, stejný doporučený rozsah, stejnou časovou dotaci, přesto se jejich práce rozsahem liší i několikanásobně.



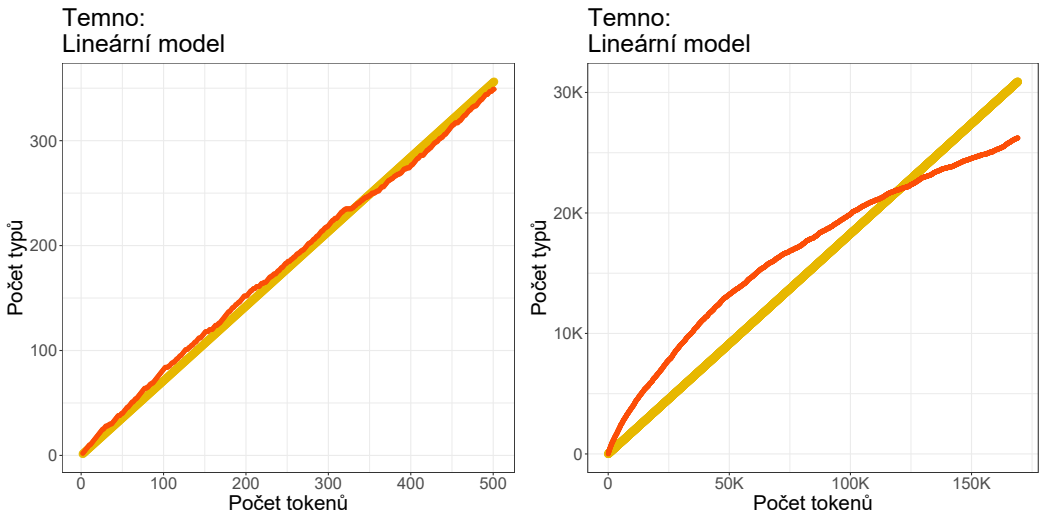
Obrázek 1.10: Srovnání, jak jednotlivé metriky ovlivňuje rozdíl ve velikosti textu.



Obrázek 1.11: Srovnání, jak jednotlivé metriky ovlivňuje malý rozdíl ve velikosti lemmatizovaného textu.



Obrázek 1.12: Vztah typů a tokenů proložený lineárním modelem — krátká sekvence 500 tokenů (vlevo), celý text (vpravo). The Last of the Mohicans.



Obrázek 1.13: Vztah typů a tokenů proložený lineárním modelem — krátká sekvence 500 tokenů (vlevo), celý text (vpravo). Temno.

Přitom pokud bychom použili normování pomocí pohyblivého okna (viz kapitolu 2.2.1), tak můžeme oba soubory docela dobře srovnat, navíc na několika různých úrovních.

Je důležité si uvědomit, že právě pro nemožnost globálního srovnání si nemůžeme zkušeností vytvořit nějakou intuici ohledně toho, jak bude vypadat text, který má určitou hodnotu TTR. Otázky typu „moji studenti mají po třech letech studia angličtiny průměrnou TTR okolo 0,6, je to moc, nebo málo?“ tak vůbec nedávají smysl.

Použití TTR „v bezpečných intervalech“ je také ošemetné v kombinaci s paradigmatem inferenční statistiky založeném na vyvracení nulové hypotézy, které je stále ještě mnohdy používáno. Pokud se dvě populace liší v průměrné délce textů, tak se budou lišit i v TTR, statistická signifikance je tedy pak už jen otázka velikosti vzorku.[42]

Cargo cult. Hlavní výhoda TTR je, že *vypadá* jako dobrá metrika, tedy má různé vnější aspekty jiných známých metrik.[43] Podobně jako korelační koeficienty, determinační koeficienty a různé metriky založené na relativních četnostech a pravděpodobnostech nabývá TTR hodnoty mezi nulou a jedničkou. Na rozdíl od těchto metrik však nula a jednička nepředstavují meze, které mají nějaký význam. Například pokud se Pearsonův korelační koeficient blíží jedničce, pak máme co do činění s extrémně dobrou lineární korelací dvou proměnných. Nic takového ovšem nemůžeme říct o TTR. Pokud se TTR rovná jedné, může jít buď o úplně běžný jev (u velmi krátkého textu, dejme tomu jednoho tweetu), nebo o projev mimořádně vysoké lexikální diverzity (řekněme v textech o několika desítkách slov), nebo to taky může znamenat, že daná sekvence nepředstavuje přirozený text, či že jsme udělali při měření nějakou chybu (pro delší texty).

Další cargo-cultickou výhodou této metriky je, že je označována zkratkou. *TTR* na první pohled vypadá jako termín a zcela jednoznačně určuje, o jaký index se jedná, zatímco *počet typů,* anglicky dejme tomu *number of types,* je nenápadná kolokace dvou slov, která se používají i mimo vědecký diskurz, přičemž *typ / type* má v běžném jazyce poněkud posunutý význam. U *type-token ratio* naproti tomu recipient jednoznačně ví, v jakém významu jsou ony *typy* myšleny, a pokud neví, tak má alespoň klíč k tomu, aby věděl, že neví. Zkratka pro *počet typů* (například *NoT* pro *number of types*) zase není v běžném kontextu dost jednoznačná, protože písmeno *T* může označovat jak

[42] K tomuto tématu doporučuji nekompromisní článek od Hess et al. (1989), který pojednává i o dalších variantách TTR, jež představuji dále.

[43] Pojmenování některých metod nebo jejich užití jako cargo-cultických je ve vědě a statistice oblíbené (Stark – Saltelli, 2018), nicméně tato metafora docela kulhá. Lidé stavějící slaměné řídicí věže si nemůžou nevšimnout, že u nich letadla prostě nepřistávají. Pilot se možná může splést a na některé falešné letiště omylem dosednout, ale je obtížné si představit, že by ani poté nic podezřelého nezaznamenal, odletěl některým z místních slaměných letadel a na letiště se dále omylem po léta vracel, jako se to děje například uživatelům t-score.

typy, tak tokeny. Často tak vidíme, že autoři mají tendenci prostý počet typů ještě nějak pojmenovat, ovšem tyto názvy jsou pak zase příliš obecné (například *richness*, která se ovšem používá i obecně pro jakýkoli index variability, nebo třeba exotická *abundance*, například v Kyle et al. (2021)).

TTR tak funguje jako označení pro „slovní bohatství“, takže dokonce i tam, kde je počet typů normován vůči délce tokenů nějak rozumně, ze zvyku naprosto zbytečně počet typů dělíme počtem tokenů a označujeme jako nějakou variantu TTR; například u MATTR (*moving average type token ratio*, viz 2.2.1) ono podělení počtu typů počtem tokenů vůbec nedává smysl a vlastně by stačilo zavést MAT nebo nějakou jinou zkratku pro *moving average number of types*.[44] Legračně pleonastické je pojmenování i koncept sTTR, tedy *standardized type-token ratio* (viz 2.2.1), kdy počet typů ve vzorku standardizujeme vydělením počtem tokenů, čímž vznikne TTR, jenomže jelikož víme, že taková standardizace je k ničemu, tak TTR standardizujeme ještě jednou a pořádně, čímž vznikne sTTR.

Neznalost funkce popisující vztah počtu typů a tokenů. Také asi není třeba přeceňovat znalosti základních kvantitativně-lingvistických zákonů u odborné veřejnosti. Zatímco kvantitativní nebo korpusoví lingvisté téměř jistě někdy v životě viděli křivku charakterizující růst počtu typů v závislosti na počtu tokenů, u jiných odvětví lingvistiky to není tak samozřejmé. Pokud se naše intuice toho, co je čím třeba normovat, vyvíjela v prostředí řekněme frekvencí slov nebo gramatických jevů, pak se nám může zdát jako klíčová otázka, co je třeba vydělit čím, abychom z absolutních četností, které máme, dostali relativní četnosti či pravděpodobnosti, které potřebujeme. Tak můžeme snadno nabýt dojmu, že na vydělení počtu typů počtem tokenů není nic špatného, nemusí nás vůbec napadnout, že bychom potřebovali mnohem složitější transformaci, než je prosté dělení.[45]

[44] Na tomto místě bych si měl popelem trochu posypat hlavu i já: v článku, ve kterém jsme s Miroslavem Kubátem MATTR dále rozvíjeli (Kubát – Milička, 2013), jsme v této praxi pokračovali. Absurdní je, že jsme v popiscích obrázků psali o TTR, přestože je na ose x neoddiskutovatelně *počet typů*. Nemyslím, že by šlo o překlep, spíš o zafixovaný pocit, že právě TTR je synonymem pro lexikální bohatství.

[45] Častou bezradnost ohledně toho, jak normovat metriky tak, aby fungovaly pro texty různé délky, si zde můžeme ilustrovat na příkladu jedné studie, jejíž autoři vynaložili nemalé úsilí, aby vytvořili index popisující pokročilost slovní zásoby u studentů (Daller et al., 2003). Nejprve nechali odborné hodnotitele vybrat ze seznamu typů ty, které považují za pokročilé, a následně podíl pokročilých typů v jednotlivých textech podělili počtem tokenů, tedy vytvořili jakousi obdobu TTR. Samozřejmě že daná metrika byla kvůli tomu velmi citlivá na počet tokenů a nepoužitelná na různě dlouhé texty (Kojima – Yamashita, 2014; Kyle, 2019). Přitom stačilo normovat počet pokročilých typů nikoli počtem tokenů, ale celkovým počtem typů, a tento problém by byl radikálně menší. S přibývajícím počtem typů by samozřejmě nakonec nastal, neboť množina pokročilých typů byla definována jako konečná, ovšem řádově později než v případě normování pomocí tokenů. Úplné nezávislosti na délce by bylo možné dosáhnout tak, že bychom nepočítali pokročilé typy, ale pokročilé tokeny, tedy instance pokročilých typů, a ty následně podělili počtem všech tokenů. Ovšem tato metrika by počítala něco trochu jiného

1.10.2 Metriky odvozené od TTR

Nevhodnost TTR samozřejmě nezůstala odbornou veřejností nepovšimnuta a velmi záhy se začaly hromadit různé alternativní metriky pro lexikální diverzitu, jejichž hlavní devízou byla domnělá nezávislost na délce textu, čemuž byla obětována interpretovatelnost. Pojďme se teď podívat na dvě klasické metriky: rootTTR a logTTR.

Poněkud anachronicky můžeme prohlásit, že obě metriky se dají odvodit z Herdanova modelu pro vztah mezi typy a tokeny (type-token relation), který byl téměř dvacet let po Herdanovi znovuobjeven Heapsem, a jelikož vědecké objevy málokdy nesou jméno svých původních objevitelů, je dodnes znám jako Heapsův zákon (Herdan, 1960; Heaps, 1978). Klasická forma tohoto modelu vypadá takto (vzorec 1.21; N značí počet tokenů, V počet typů, a a b jsou parametry):

$$V = aN^b \tag{1.21}$$

Hledanou metrikou je pak vždy jeden z parametrů tohoto modelu, přičemž ostatním parametrům přiřadíme nějakou defaultní hodnotu. Dokonce i samotné TTR můžeme chápat tak, že vlastně implicitně vychází z tohoto modelu, pokud totiž je parametr b roven jedné, pak TTR je rovno parametru a (vzorce 1.22 a 1.23):

$$a = \frac{V}{N^b} \tag{1.22}$$

$$\mathrm{TTR} = \frac{V}{N} \tag{1.23}$$

Takové defaultní nastavení hodnoty parametru b ovšem neodpovídá realitě (a to je taky důvod, proč TTR selhává), neboť tento parametr se obvykle pohybuje někde okolo čísla 0,6, tedy alespoň v evropských jazycích. Toho využívá rootTTR, známý též jako Guiraudův index.

Guiraudův index (rootTTR)

Guiraudův index (Guiraud, 1954) vzniká tak, že do vzorce 1.22 dosadíme za parametr b konstantu 0,5, kterážto byla vybrána částečně z empirických, částečně z numerologických důvodů, neboť umocnění jednou polovinou je ekvivalentní s druhou odmocninou — vzniká tak tedy vzorec 1.24, který vypadá, že žádnou konstantu vlastně ani neobsahuje. Pokud bychom místo jedné poloviny zvolili nějaké jiné číslo, které by třeba lépe odpovídalo danému jazyku, nutně by vzorec budil otázku „proč právě tato

než ta původní a vlastně by předpokládala, že k textu bohatému na pokročilou slovní zásobu stačí neustále dokola opakovat jeden jediný pokročilý slovní typ, například *paradigma*. Asi ideální řešení proto představuje normování počtu „pokročilých typů“ pomocí obecných metod popsaných v 2. kapitole.

hodnota konstanty?“, zatímco takhle se můžeme tvářit, že ona odmocnina má nějaké teoretické opodstatnění.

$$\text{RootTTR} = \frac{V}{\sqrt{N}} \tag{1.24}$$

Jak je vidět na obrázcích 1.14 a 1.15, odmocninový model sice sedí lépe než čistě lineární, ovšem pro krátké texty je dost nevhodný, což se také projevuje na jeho chybovosti (vraťme se prosím ke grafům na obrázcích 1.10 a 1.11), která je pro kratší texty či vzorky ještě větší než TTR, nicméně v angličtině postupně klesne na docela přijatelné hodnoty — díky tomu, že právě v angličtině je parametr b obvykle blízko hodnotě jedné poloviny (obdobně je tomu ve francouzštině, na které Guiraud metriku původně testoval). V nelemmatizované češtině a arabštině zůstává systematická chyba stále vysoká, což by šlo nejspíš zlepšit úpravou parametru b.

LogTTR

Druhá možnost, kterou nám Herdanův model nabízí, je využití parametru b. Tedy ze vzorce 1.21 za parametr a dosadíme jedničku a vyjádříme parametr b (1.25; připomínám, že je nutné zlogaritmovat jak podle osy x, tak podle osy y, pouhým zlogaritmováním samotného TTR bychom nedosáhli vůbec ničeho, pouze změny měřítka):

$$\text{LogTTR} = b = \frac{\log V}{\log N} \tag{1.25}$$

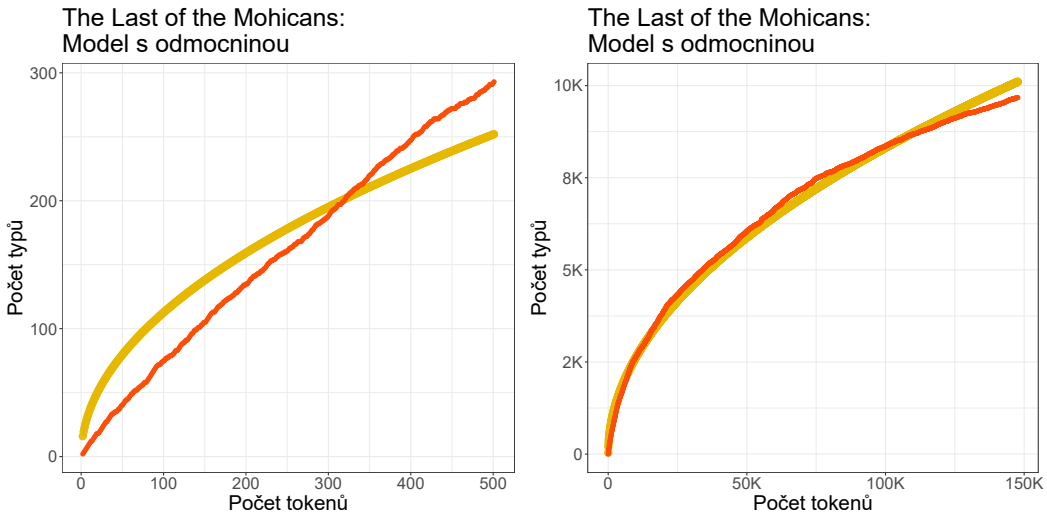
Kromě označení logTTR se v literatuře setkáte i s označením *Herdanovo C* (Herdan's C). Podle grafů na obrázcích 1.10 a 1.11 si tato metrika vede překvapivě dobře — systematická chyba začíná nízko, klesá rychle, i když na delších anglických textech má navrch rootTTR a baseline se vůbec neblíží.

V literatuře se pracuje s ještě kurioznější metrikou — dvojitě zlogaritmovaným TTR (tedy podle vzorce 1.26):

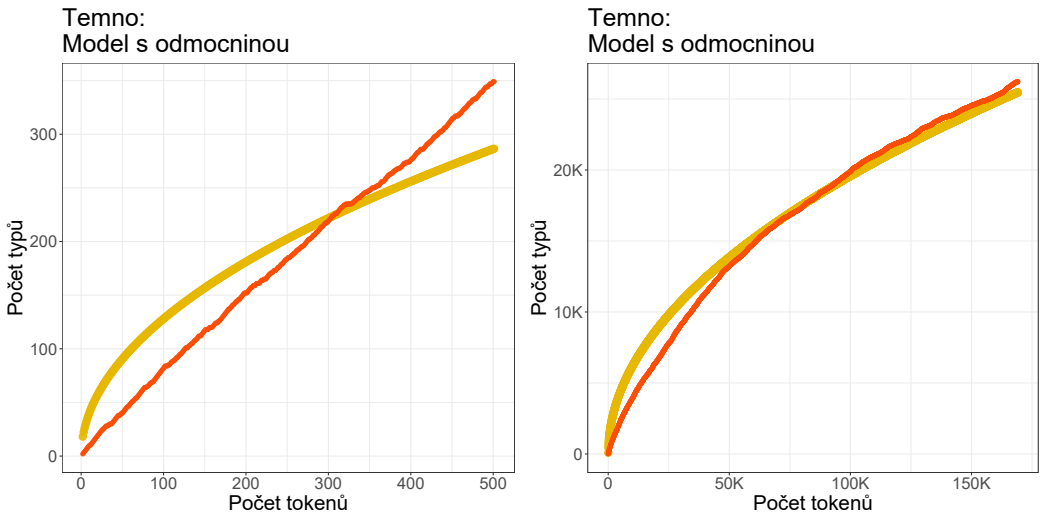
$$\text{LogLogTTR} = \frac{\log\log V}{\log\log N} \tag{1.26}$$

Vznikne tak tzv. *Somersův index* (*Somers' Index*, Somers (1959)), který překvapivě dokáže vzdorovat nástrahám textů různých délek ještě líp než logTTR (viz grafy na obrázcích 1.16 a 1.17). Tento vzorec již samozřejmě nevychází z prostého Herdanova modelu, ale z modelu poněkud komplikovanějšího (1.27, kde c je jediný parametr odpovídající metrice loglogTTR; nepodařilo se mi zjistit, jaký má model teoretický základ):

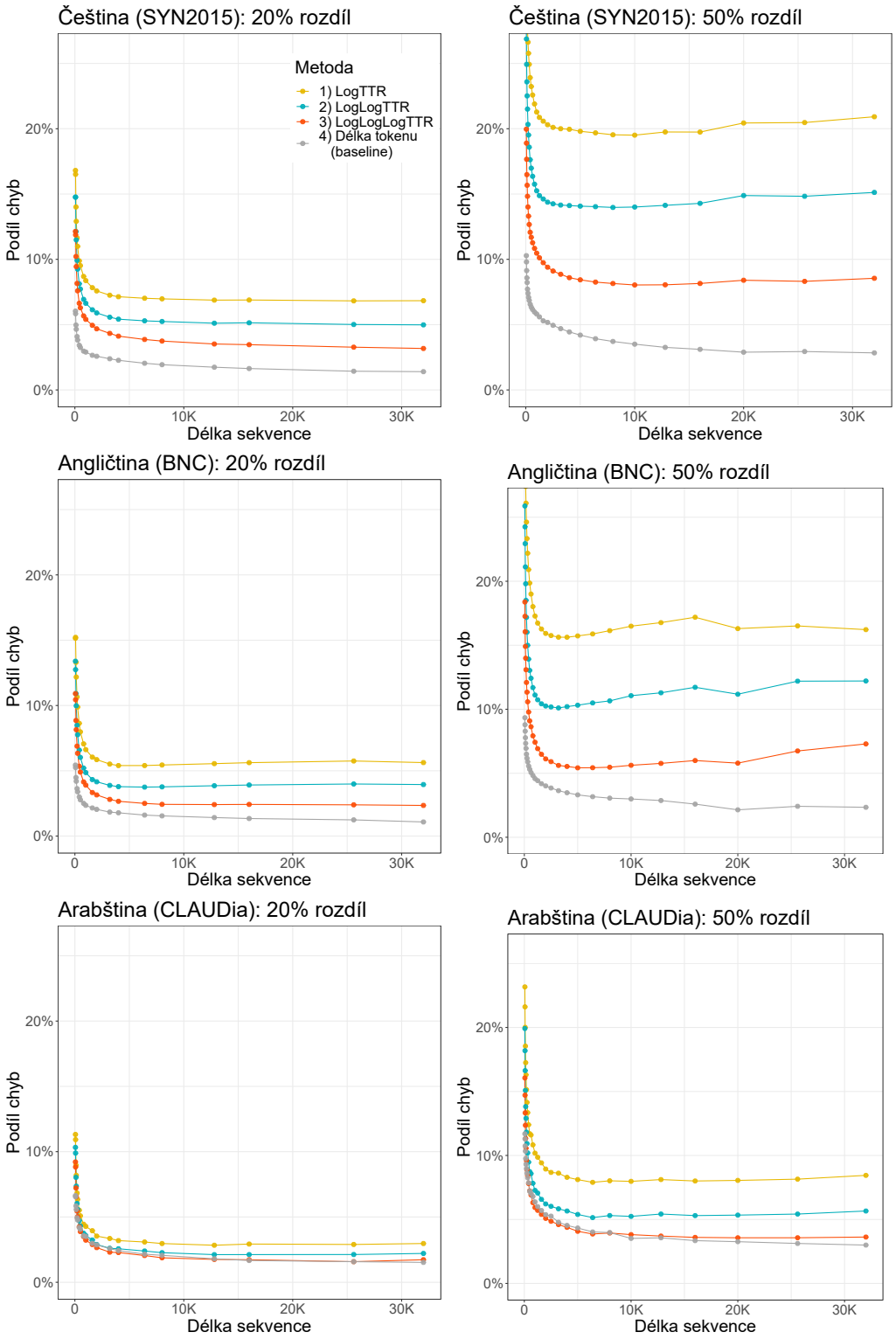
$$V = e^{\sqrt[c]{\log N}} \tag{1.27}$$



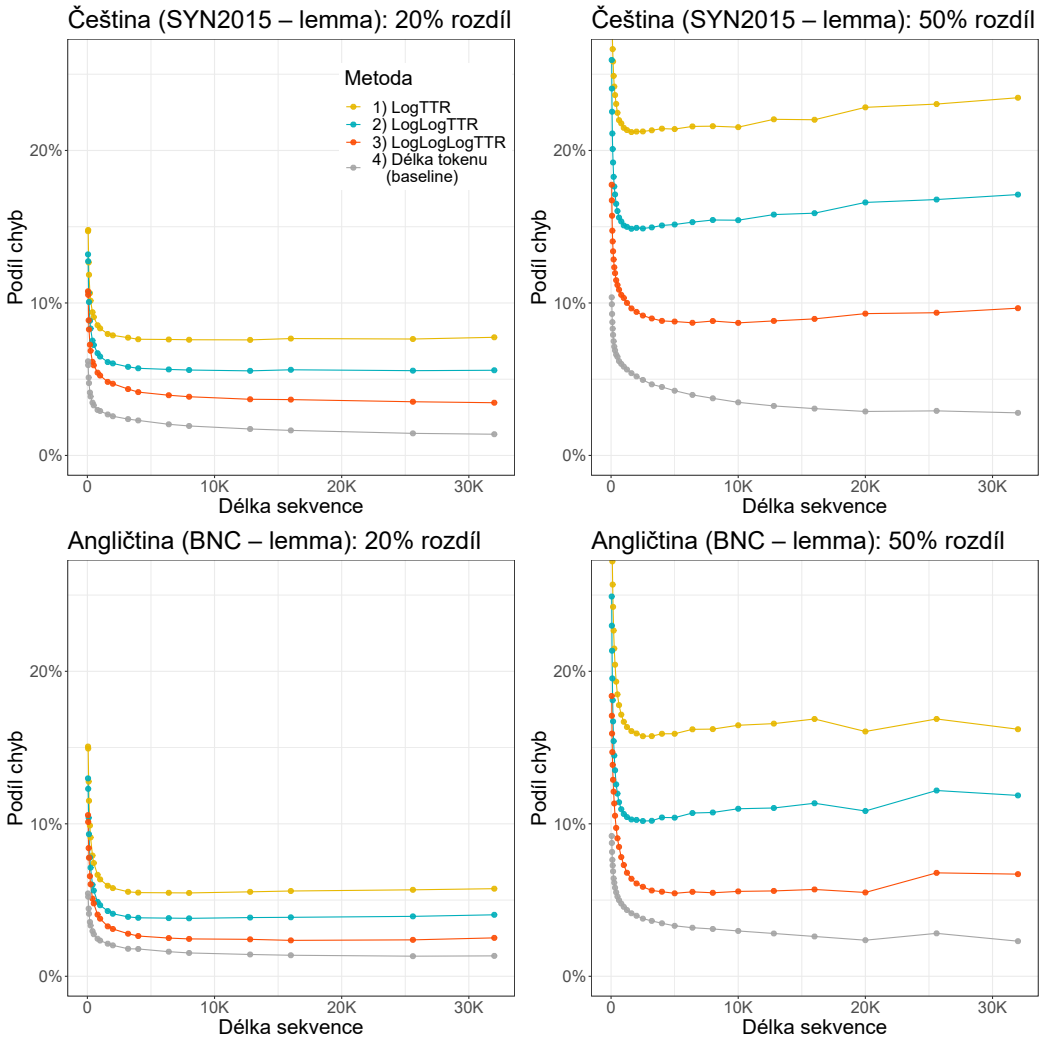
Obrázek 1.14: Vztah typů a tokenů proložený odmocninovým modelem — krátká sekvence 500 tokenů (vlevo), celý text (vpravo). The Last of the Mohicans.



Obrázek 1.15: Vztah typů a tokenů proložený odmocninovým modelem — krátká sekvence 500 tokenů (vlevo), celý text (vpravo). Temno.



Obrázek 1.16: Srovnání, jak jednotlivé metriky ovlivňuje rozdíl ve velikosti textu.



Obrázek 1.17: Srovnání, jak jednotlivé metriky ovlivňuje rozdíl ve velikosti lemmatizovaného textu.

Vzhledem k tomu, že i tento model má menší procento chyb než předchozí (viz obrázky 1.16 a 1.17), nedalo mi to a zkusil jsem obě osy z legrace zlogaritmovat ještě jednou (1.28).

$$\text{LogLogLogTTR} = \frac{\log\log\log V}{\log\log\log N} \tag{1.28}$$

Tato monstrozita ovšemže nevychází z žádného rozumného modelu pro vztah typů a tokenů.[46] Nicméně úspěšnost dané metriky ohledně délky textů je ještě vyšší než v předchozích dvou případech a v arabštině je dokonce srovnatelná s úspěšností délky slov, kterou používáme jako baseline metriku fundamentálně nezávislou na délce textu. Kdybych zde používal jiný způsob měření úspěšnosti, byl bych na pochybách, jestli to není úspěch jen zdánlivý, daný pouze radikální změnou měřítka, nicméně mám za to, že zvolená metodologie je, co se měřítka týče, neprůstřelná. Navíc rozhodně neplatí, že čím víckrát osy zlogaritmuji, tím lepší výsledky získám, když proces zopakujeme počtvrté, lepší výsledky již nedostaneme a páté zlogaritmování je již kontraproduktivní.

1.10.3 Lambda

Jak je vidět, obě předchozí metriky zlepšují závislost na délce textu, nicméně problém jako takový neřeší a nemyslím si, že by takovéto zlepšení stálo za ztrátu možnosti index jednoduše interpretovat. Namísto prostého *počtu typů* tu najednou máme *parametry určitého modelu pro vztah typů a tokenů*, číslo, které nemá žádnou jednotku a pod kterým si není možné nic konkrétního představit.

Metrik, jejichž interpretace je podobně neintuitivní jako u rootTTR a logTTR a jejichž autoři tvrdí, že jsou nezávislé na délce textu, přičemž ovšem tento problém řeší také nepřesvědčivě, najdeme v literatuře desítky. Poněvadž problém délky navrhuji řešit systematicky a věnuji mu celou kapitolu 2, nebudu se těmito více či méně složitě konstruovanými[47] metrikami zabývat každou zvlášť, vydalo by to na samostatnou monografii.

Nicméně rád bych se na chvíli zastavil u *indexu lambda* jakožto jejich prototypického zástupce a taky kvůli jeho české stopě, neboť Radek Čech stál jak u jeho zrodu,

[46] Respektive modelem je vzorec 1.29, což je ovšem jen ad hoc vyjádření proměnné V z již hotové metriky, bez jakékoli teorie nebo fundamentálního principu.

$$V = e^{e^{\sqrt[c]{\log(\log N)}}} \tag{1.29}$$

[47] McCarthy a Jarvis to nazývají „sophisticated approaches to lexical diversity“ (McCarthy – Jarvis, 2010), ovšem osobně bych onu „sofistikovanost“, nejde-li ruku v ruce se snadnou interpretovatelností, považoval spíše za nevýhodu.

tak u jeho konce. Lambda (Popescu et al., 2010, 2011) vychází z celé distribuce frekvencí, konkrétně eukleidovské délky zipfovské křivky, která charakterizuje daný text. Tato délka křivky, která je samozřejmě na délce textu závislá, je následně normována vydělením délkou textu a vynásobením dekadickým logaritmem délky daného textu. Nepodařilo se mi zjistit, z jaké teorie dané normování vychází nebo jaké má předpoklady, nejspíš je daný vzorec prostě empiricky odvozený z dat, v citovaných publikacích podrobnosti chybí, nicméně velmi autoritativně se v nich tvrdí, že lambda je prokazatelně nezávislá na délce textu.

Toto tvrzení bylo rigorózně empiricky testováno až o pět let později (Čech, 2015) a ukázalo se, že jednoduše neplatí. Tedy, platí pouze na velmi omezeném intervalu.[48]

Pokusme se nyní onen empirický důkaz doplnit teoretickou explanací. Abychom pochopili, proč daná metrika ve své úloze selhává, pojďme se napřed podívat na její původní vzorec (1.30) (Popescu et al., 2011, str. 1 a 2).[49]

$$\lambda = \frac{\log_{10}(N)}{N} \sum_{i=1}^{V-1} \sqrt{(f(t_i) - f(t_{i+1}))^2 + 1} \tag{1.30}$$

Zlomek na začátku udává normování indexu podle délky textu, zbytek rovnice jednoduše s využitím Pýthagorovy věty počítá eukleidovskou délku křivky značící zipfovskou distribuci (správnější by bylo mluvit o rank-frequency relation, neboť striktně vzato o Zipfův model zde vůbec nejde).

Když se podíváme na skutečná empirická data zipfovské distribuce na nějakém běžném textu v angličtině (obrázek 1.18, graf vlevo),[50] zjistíme, že křivka nejprve velmi rychle klesá, v podstatě vizuálně kopíruje osu *y*, a následně kopíruje osu *x*. Prakticky to znamená, že na začátku je úhlopříčka značící vzdálenost mezi jednotlivými datovými body, kterou pomocí Pýthagorovy věty počítáme, téměř rovna vertikální vzdálenosti (vzorec 1.31), na konci pak vzdálenosti horizontální (vzorec 1.32), takže s výjimkou velmi krátkého úseku někde uprostřed můžeme eukleidovskou vzdálenost nahradit vzdáleností manhattanskou (1.33).

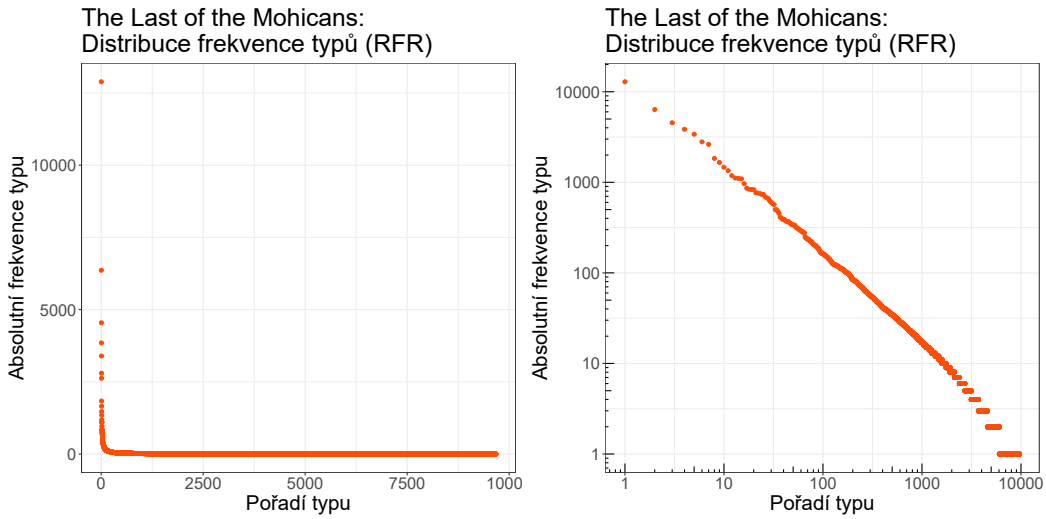
$$\sqrt{(f(t_i) - f(t_{i+1}))^2 + 1} \approx f(t_i) - f(t_{i+1}) \tag{1.31}$$

$$\sqrt{(f(t_i) - f(t_{i+1}))^2 + 1} \approx 1 \tag{1.32}$$

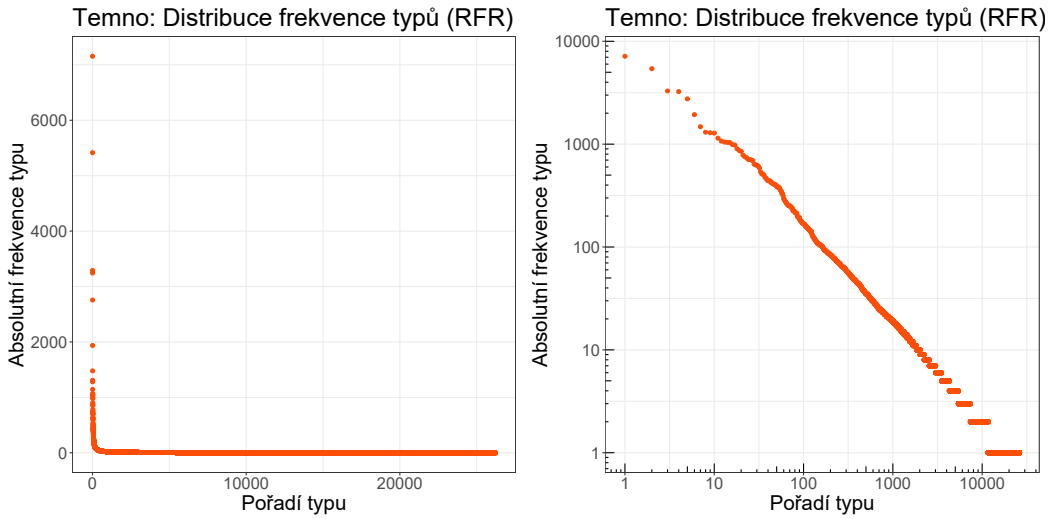
[48] Nejsem si jistý, proč bylo nutné lambdu zabíjet podruhé, nicméně učinili tak Poiret – Liu (2017).

[49] Stejně jako v dalších vzorcích v této publikaci N značí délku textu, V počet typů a $f(t_i)$ absolutní frekvenci typu t_i.

[50] V tomto případě jde o román *The Last of the Mohicans* od Jamese Fenimora Coopera. Za normálních okolností se takováto distribuce zobrazuje tak, že obě osy jsou logaritmovány, aby byl graf přehlednější, tak jak to vidíme na obrázku vpravo. Zde nám ale jde právě o to názorně vizualizovat, že distribuce je extrémně nevyrovnaná, proto na obrázku vlevo osy nijak netransformujeme.



Obrázek 1.18: The Last of the Mohicans. Distribuce frekvencí jednotlivých typů (rank-frequency relation).



Obrázek 1.19: Temno. Distribuce frekvencí jednotlivých typů (rank-frequency relation).

$$\lambda \approx \frac{\log_{10}(N)}{N}\,(f(t_1) + V) \tag{1.33}$$

Takto manhattansky spočítaná délka se od eukleidovské příliš neliší: u textu *The Last of the Mohicans* je to 22562 jednotek podle eukleidovské vzdálenosti versus 22398 jednotek podle manhattanské vzdálenosti. Pokud z těchto délek vypočítáme lambdu, vyjde nám 0,783 versus 0,789, tedy kolem sedmi promile rozdílu. Jak vidíme na obrázku 1.19, ani český text, který má distribuci slovních frekvencí vyrovnanější (z důvodů komplexnější morfologie), nebude mít rozdíl mezi eukleidovskou a manhattanskou vzdáleností nijak oslnivý. A skutečně, například pro Jiráskovo *Temno* vychází lambda 0,630, zatímco manhattanská pseudolambda 0,635, tedy rozdíl činí zhruba 7,7 promile.

Jelikož autoři metriky považují λ za konstantní a relativní četnost nejčastějšího slova ($p(t_1)$) je v delších textech také více méně konstantní, můžeme metriku dále zjednodušit na vzorec 1.34.

$$\frac{\lambda}{\log_{10} N} \approx p(t_1) + \frac{V}{N} \tag{1.34}$$

A vyjádřením V se dostáváme k teoreticky nijak neopodstatněnému modelu pro type-token relation 1.35.[51]

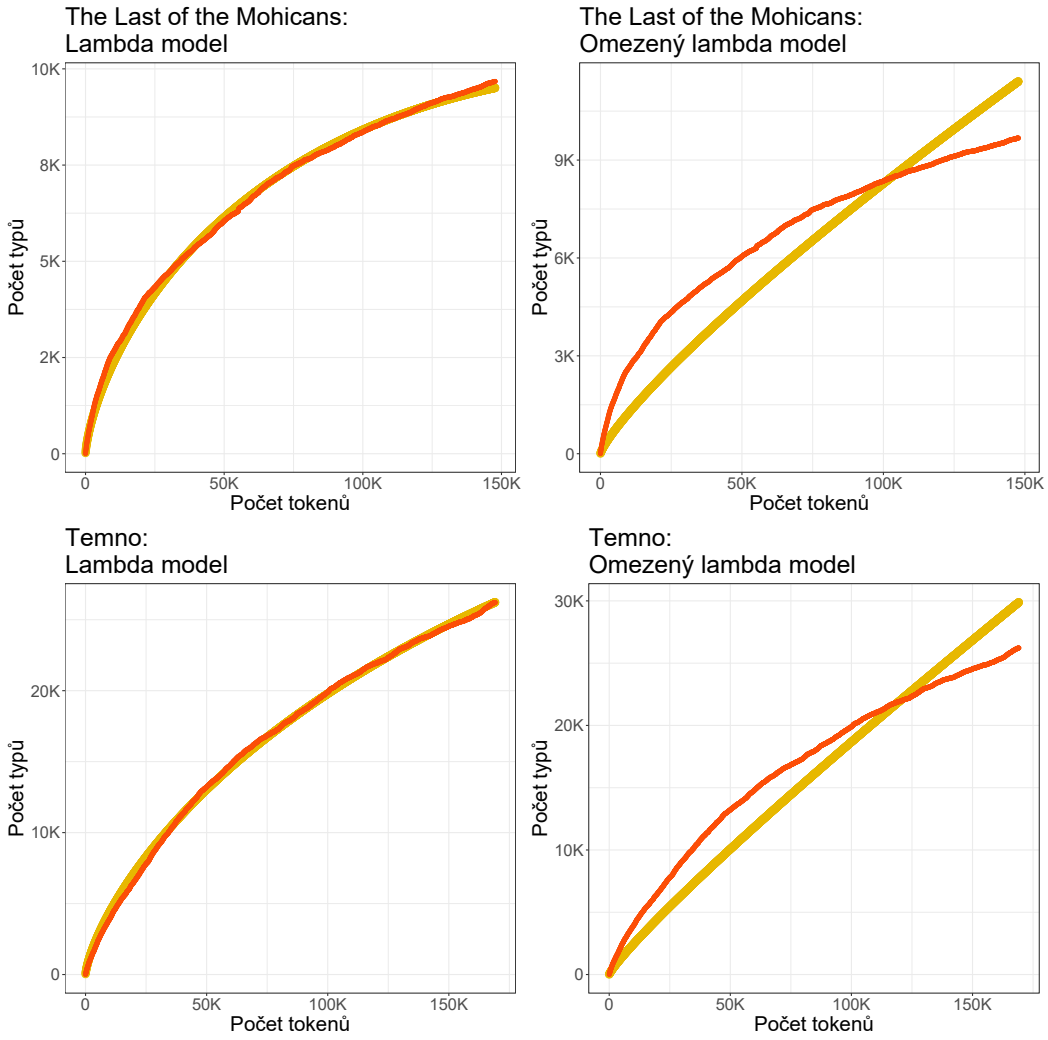
$$V = N\left(\frac{\lambda}{\log N} - p(t_1)\right) \tag{1.35}$$

Pokud budeme chápat λ a $p(t_1)$ jako volné parametry a tento model nafitujeme na data, pak model může překvapivě dobře empiricky sedět — viz obrázek 1.20 vlevo, červená křivka reprezentuje empirická data, žlutá pak model 1.35 s parametrem $\lambda = 7{,}06$ a s parametrem $p(t_1) = 0{,}52$, determinační koeficient je pak poměrně vysoký ($R^2 = 0{,}998$).[52] Vzhledem k tomu, jak obrovské množství modelů pro TTR již bylo prozkoumáno, je s podivem, že takto empiricky slušně sedící model už dávno nevzbudil pozornost. Soupis nejznámějších historických i současných modelů pro TTR naleznete například ve studii Davida Mitchella (2015).

Problém indexu lambda tkví v tom, že při jeho počítání je parametr p_1 pevně dán frekvencí nejčastějšího slovního typu. Tím pádem se snaží nafitovat TTR pomocí jednoho parametru a bohužel v tom selhává — nejčastější slovo v románu *The Last of the Mohicans* má relativní frekvenci rovnu 0,0872 (je to určitý člen *the*), a pokud se pokusíme nafitovat model 1.35 s takto pevně daným parametrem $p(t_1)$, tak dostaneme parametr λ roven 1,91, což ovšem sníží determinační koeficient takto omezeného modelu na $R^2 = 0{,}785$ (což je hezky ilustrováno na obrázku1.20 vpravo).

[51] Dekadický logaritmus zde měním na přirozený, protože se s ním lépe počítá. V původním vzorci nemá žádné teoretické opodstatnění a báze logaritmu nemá vliv na parametr p_1.

[52] Nafitováno pomocí programu Eureqa (Schmidt – Lipson, 2009).



Obrázek 1.20: *The Last of the Mohicans* a *Temno*. Type-token relation a jeho modely odvozené od indexu lambda.

Mohli bychom se pokusit index nějak zachránit. Upravit ho tak, aby parametr $p(t_1)$ lépe seděl, klidně ho určit pro každý jazyk zvlášť… Jenomže jak si ukážeme v následujících kapitolách, nezávislost indexu na délce textu je vlastně umělý problém, který není nutné řešit pro každý index zvlášť, ale je možné jej vyřešit systematickou skutečnou normalizací, která bude obecná a bude fungovat pro všechny indexy nezávisle na způsobu jejich počítání.

Kapitola 2

Délka textu

Posledních sedmnáct stránek jsem věnoval metrikám, které vznikly ne proto, že by byly inherentně lepší než metriky popsané v předchozích kapitolách, tedy že by lépe seděly na naše konceptualizace lexikální diverzity, nebo že by byly lépe interpretovatelné. Spíše naopak. Vznikly proto, že si jejich autoři mysleli, že jejich hodnoty jsou *nezávislé na délce textu*. Respektive na délce vzorku, který z onoho textu vyjmeme.

Metriky lexikální diverzity totiž více či méně korelují s délkou textu a ona korelace není lineární, tedy nemůžeme se s ní nějak jednoduše vypořádat. Vlastně není ani jednoduše modelovatelná nějakou teoreticky zdůvodněnou a empiricky dobře nafitovatelnou funkcí — třeba David Mitchell (2015) vyjmenovává hezkou řádku modelů pro vztah typů a tokenů a soupis není ani zdaleka kompletní. Jediné, čím si můžeme být jistí, je, že počet typů nekonverguje k nějaké nepřekročitelné hranici (Milička, 2013), neboť jazyk je dynamický systém. Tedy že nás nijak nespasí užívání dlouhých textů nebo celých korpusů. Totéž platí i pro ostatní metriky odvozené od frekvenční distribuce typů v textu, jako je perplexita či pravděpodobnost opakování. Jsme totiž limitováni tím, že v současné době nemáme ani uspokojivý model distribuce slovních frekvencí,[1] natož pak model její dynamiky.

Dřív si lidé mysleli, že najdou nějakou *ideální* metriku, která bude na délce textu nezávislá. Co víc, panovalo obecné přesvědčení, že *ideální metrika se pozná právě podle toho, že bude nezávislá na délce textu*. Slovy autorů *Lambda Structures* (jedné ze studií, která o sobě tvrdí, že problém vyřešila) „[...] history [of vocabulary richness] is almost an epos describing the battle against the influence of text length N“ (Popescu et al., 2011, str. 1). Tento výrok autoři podpořili výčtem několika desítek publikací od 40. let dvacátého století až po současnost.

[1] Čímž nechci říct, že bychom měli málo těchto modelů, naopak, modelů distribuce frekvencí slov (respektive zipfovsky pojaté rank-frequency relation) máme obrovské množství, ovšem už samo toto množství signalizuje naši určitou bezradnost a nejistotu v tom, co vlastně modelujeme, jaké jsou hranice této entity a jak zjistíme, který model je nejlepší. Otázka je, pokud bychom na správný model narazili, jestli bychom ho vůbec uměli poznat.

Úkolem této kapitoly bude vás přesvědčit, že touha po metrice nezávislé na délce textu sice možná není marná, nicméně je zbytečná, neboť existuje několik způsobů, jak prakticky libovolnou metriku lexikální diverzity normovat a oné závislosti ji zbavit.

2.1 Rozsah problému

Jak bylo naznačeno v první kapitole, kdesi v hloubi sedmnáctého století byl barometr a teploměr jedním přístrojem. Prvním krokem k jeho rozdělení do dvou bylo zjištění, že míchá dohromady dvě různé veličiny a že je nejen možno, ale hlavně že je záhodno je rozdělit. Ani my teď neopomeneme tento důležitý krok a podíváme se, jestli je ono rozdělení vlastně vůbec nutné, jestli se za určitých podmínek (například v dlouhých vzorcích) neděje samo od sebe.

Pojďme se tedy nejprve podívat, jak moc popsané metriky na délce textu závisí. Pro jistotu ještě jednou zdůrazním, že to rozhodně neděláme proto, abychom zjistili, která metrika je *lepší*, neboť nezávislost na délce textu nepovažuji za známku ideální metriky. Jednoduše chceme jen zjistit, jestli je normování vlastně vůbec za běžných podmínek potřeba.

2.1.1 Metodika

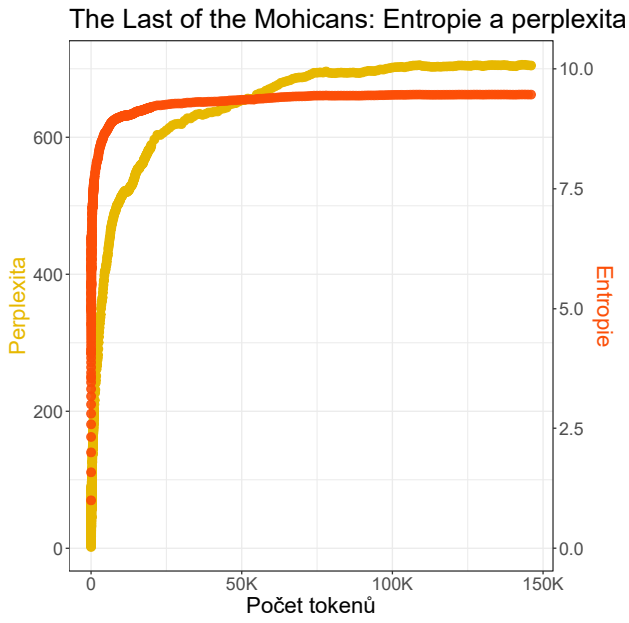
Najít obecnou metodu, jak porovnávat systematické chyby u všech různých metrik, je docela náročné, neboť různé metriky mají různé měřítko a škálování (lineární versus logaritmické).

Užívanou, byť podle mě nevhodnou metodou je sledovat, jestli metrika se zvyšující se délkou textu konverguje k nějaké hodnotě.[2] Bez nějakých teoretických předpokladů je ovšem empiricky těžké určit, jestli se jedná o konvergenci, nebo ne, a abychom dosáhli „optické“ konvergence, stačí nám výsledek metriky třeba i jen zlogaritmovat. Například o perplexitě, jak jsme ji definovali v kapitole 1.4, bychom určitě neřekli, že její hodnota s rostoucí délkou vzorku konverguje k nějaké konstantě, zatímco shannonovská entropie tak může docela dobře vypadat (obr. 2.1). Přitom obě metriky se liší jen škálováním.

Jak tedy nezávisle na měřítku určit, jak moc velkou chybu udělá ten, kdo by danou metriku lexikální diverzity považoval za invariantní vůči délce textu? Pokud otázku formulujeme právě takhle, nabízí se přímočará metodologie, která přímo pracuje s pojmem systematické chybovosti — budeme srovnávat sekvence z korpusu stejné a nestejné délky a spočítáme podíl těch, kde různost délky způsobila chybu.

Formálně řečeno postupujeme následovně. Z korpusu vybereme dvě náhodné sekvence o určitém počtu tokenů N (stejně jako v ostatních případech i zde dbáme na

[2] Například Shi – Lei (2022) se pomocí této metody snaží dokázat, že Zhagův estimátor entropie je metrika lexikální diverzity od určité délky textu na délce textu nezávislá.



Obrázek 2.1: Závislost perplexity (žlutá linie) a entropie (červená linie) na počtu tokenů ve vzorku (*The Last of the Mohicans*, od začátku do konce).

to, aby sekvence nekřížily hranice textů), nazvěme si je $S_1(N)$ a $S_2(N)$. Následně ze stejného místa, z jakého jsme vybrali sekvenci $S_2(N)$, vybereme sekvenci o něco kratší, dejme tomu o čtvrtinu, takže máme další vzorek $S_2(0{,}75N)$, obecně $S_2(kN)$. Na všech třech sekvencích změříme všechny potřebné metriky lexikální diverzity, nazvěme si hodnotu jedné takové metriky jako $D(S)$. Takovýchto trojic vybereme z korpusu velké množství (v našem případě to bylo deset milionů) a do vzorku zařadíme ty, ve kterých se $D(S_1(N))$ liší od $D(S_2(N))$(formálně ve vzorci 2.1). Jako chybnou počítáme takovou trojici, kde $D(S_1(N))$ je větší než $D(S_2(N))$ a $D(S_1(N))$ je menší než $D(S_2(kN))$ nebo obráceně (formálně ve vzorci 2.2).

Naší metrikou chybovosti $E_D(k, N)$ je pak podíl chybných trojic ve vzorku (vzorec 2.3).

$$c_D(k, n) = |\mathrm{sgn}(D(S_1(N)) - D(S_2(N)))| \tag{2.1}$$

$$e_D(k, n) = \left\lfloor \frac{\mathrm{sgn}(D(S_1(N)) - D(S_2(N))) - \mathrm{sgn}(D(S_1(N)) - D(S_2(kN)))}{2} \right\rfloor \tag{2.2}$$

$$E(k, N) = \frac{\sum e_D(k, N)}{\sum c_D(k, N)} \tag{2.3}$$

Mohli bychom použít i jiné statistiky, nicméně tady tato nám dává jistotu, že jsme nic

nezanedbali co se týče škálování a měřítka, navíc se dá docela snadno interpretovat.[3] Ovšem ani tato metoda nás nezbaví nutnosti nalézt nějakou mez, od které můžeme chybovost považovat za příliš velkou, potřebujeme vědět, jakých hodnot chybovosti dosahují metriky na délce textu skutečně nezávislé. Využívám toho, že mezi metrikami lexikální diverzity máme zařazeny dvě, které jsou nezávislé na délce sekvence už z definice — průměrná délka tokenů a podíl autosémantik. Empiricky naměřené hodnoty chybovosti těchto metrik tedy budeme považovat za ideál, ke kterému se ostatní metriky mohou (ale spíše nebudou) blížit.

2.1.2 Výsledky

Zajímá nás tedy, jakou chybovost můžeme očekávat od různých metrik lexikální diverzity, různých velikostí sekvence a různých rozdílů mezi sekvencemi. Musel jsem nutně udělat jenom malý výběr mezi všemi možnostmi, které se naskýtají, avšak myslím, že oněch deset grafů bude dostatečně výmluvných. Zařadil jsem dva rozdíly ve velikosti sekvencí — menší (grafy 2.2, 2.4, 2.6, 2.9 a 2.11) a větší (grafy 2.3, 2.5, 2.7, 2.10 a 2.12). Menší rozdíl je dvacetiprocentní, tedy kratší sekvence je o pětinu menší než ta delší. Větší rozdíl je padesátiprocentní, tedy kratší sekvence je vůči delší sekvenci poloviční. Rozdíly byly vybrány tak, aby odpovídaly scénáři, kdy texty v korpusu považujeme za prakticky stejně dlouhé a kdy je považujeme za srovnatelně dlouhé.

Nelemmatizovaný text

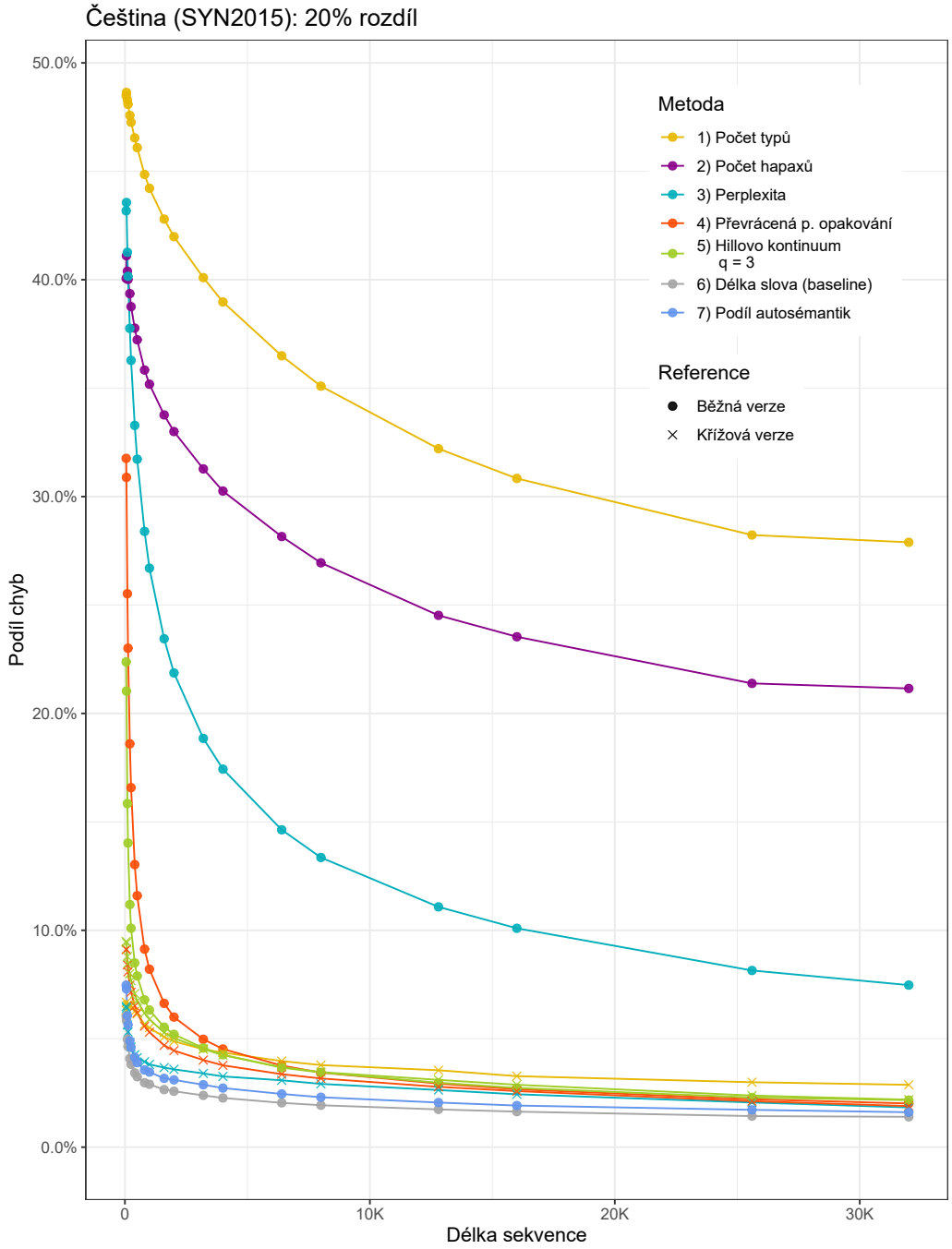
Pojďme se nejprve podívat na výsledky naměřené na prostých nelemmatizovaných textech.

Nepřekvapí, že počet typů a počet hapaxů má ve všech třech jazycích obrovskou chybovost, u sekvencí s větším rozdílem se u krátkých sekvencí blíží padesáti procentům, čili maximu, kterého tato metrika může dosáhnout. Perplexita[4] také začíná na vysokých číslech, nicméně celkem rychle padá díky tomu, že dává větší váhu na slova s vyšší frekvencí — z pohledu Hillova kontinua má vyšší koeficient q. Můžeme předpokládat, že čím vyšší je Hillovo q, tím menší dopad délka textu má, neboť relativní frekvence častých slov jsou s přibývající délkou sekvence ovlivněny méně a méně,

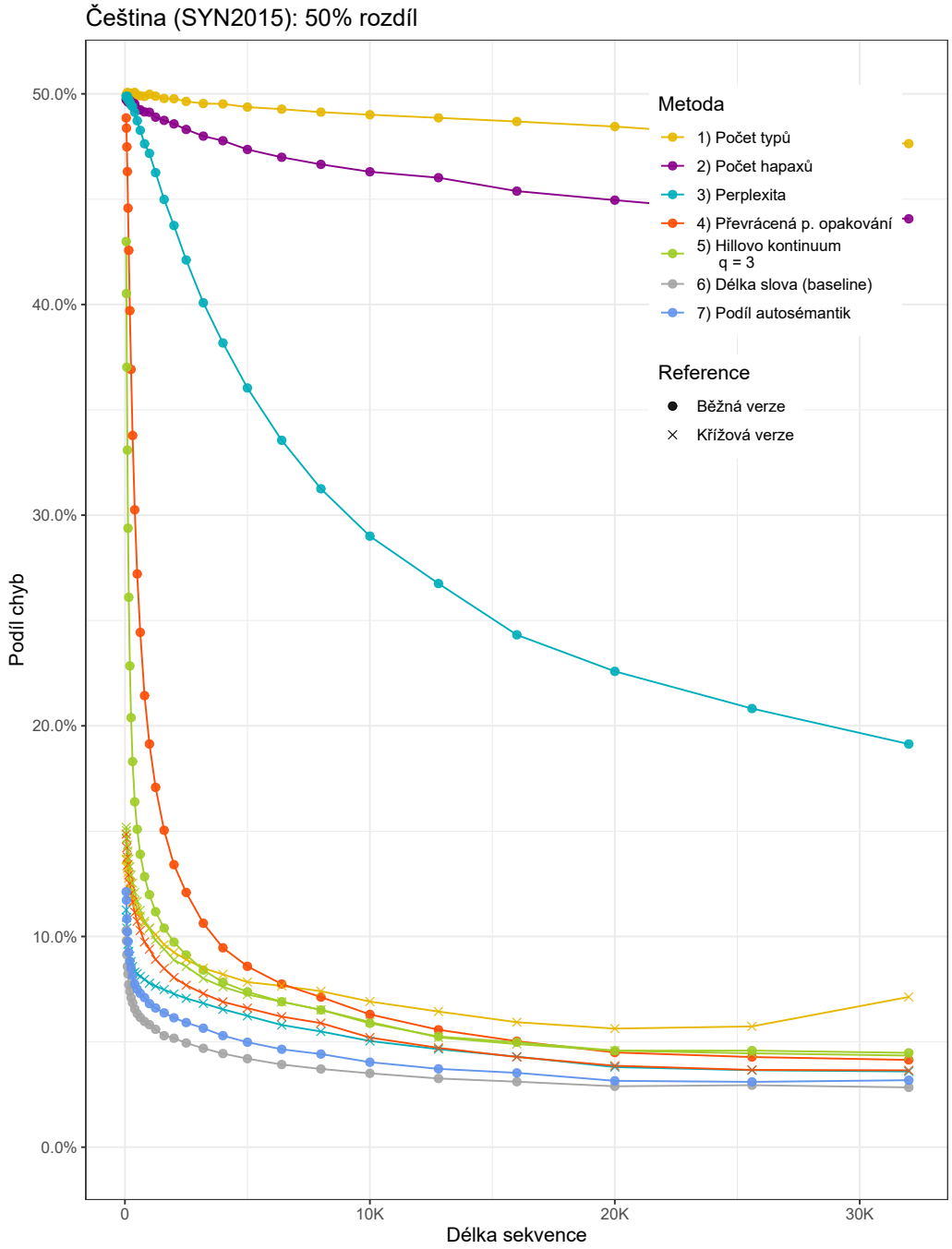
[3] Z jiných metrik jsem zkoušel průměrný rozdíl kratšího a delšího vzorku normovat průměrnou absolutní odchylkou rozdílu vzorků stejné délky (2.4; average absolute deviation, AAD, konkrétně aritmetický průměr odchylky od aritmetického průměru), nicméně výsledky byly prakticky totožné, jen mírně transformované, proto je zde neuvádím.

$$e'_D(k, N) = \frac{D(S_1(N)) - D(S_2(kN))}{AAD(D(S_1(N))) - D(S_2(N))} \tag{2.4}$$

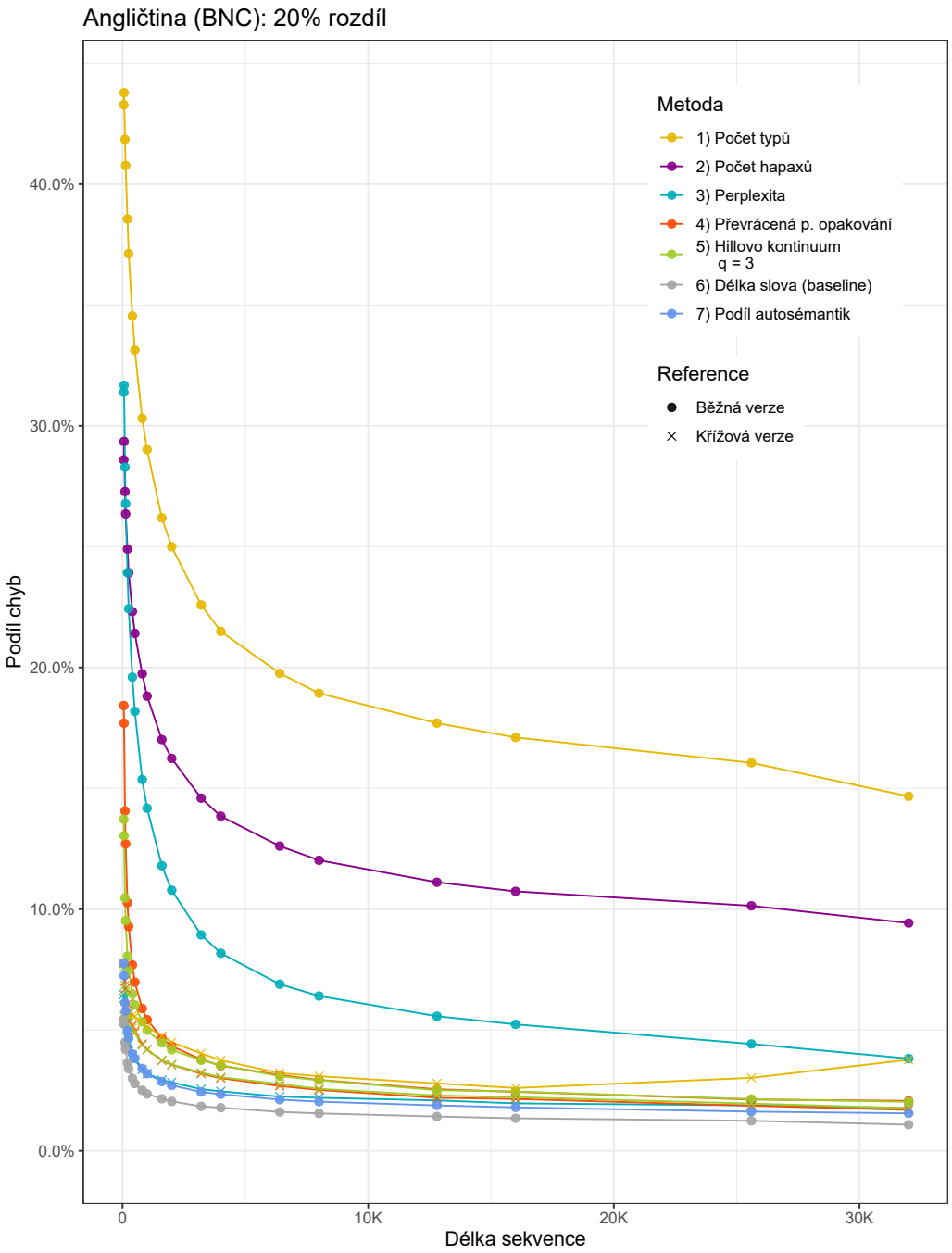
[4] Výsledky chybovosti pro perplexitu jsou stejné jako výsledky pro entropii. Totéž platí pro křížovou perplexitu a křížovou entropii. U této metodologie to platí z definice, ovšem zkoušel jsem to i empiricky jako způsob ověření, že všechno funguje správně.



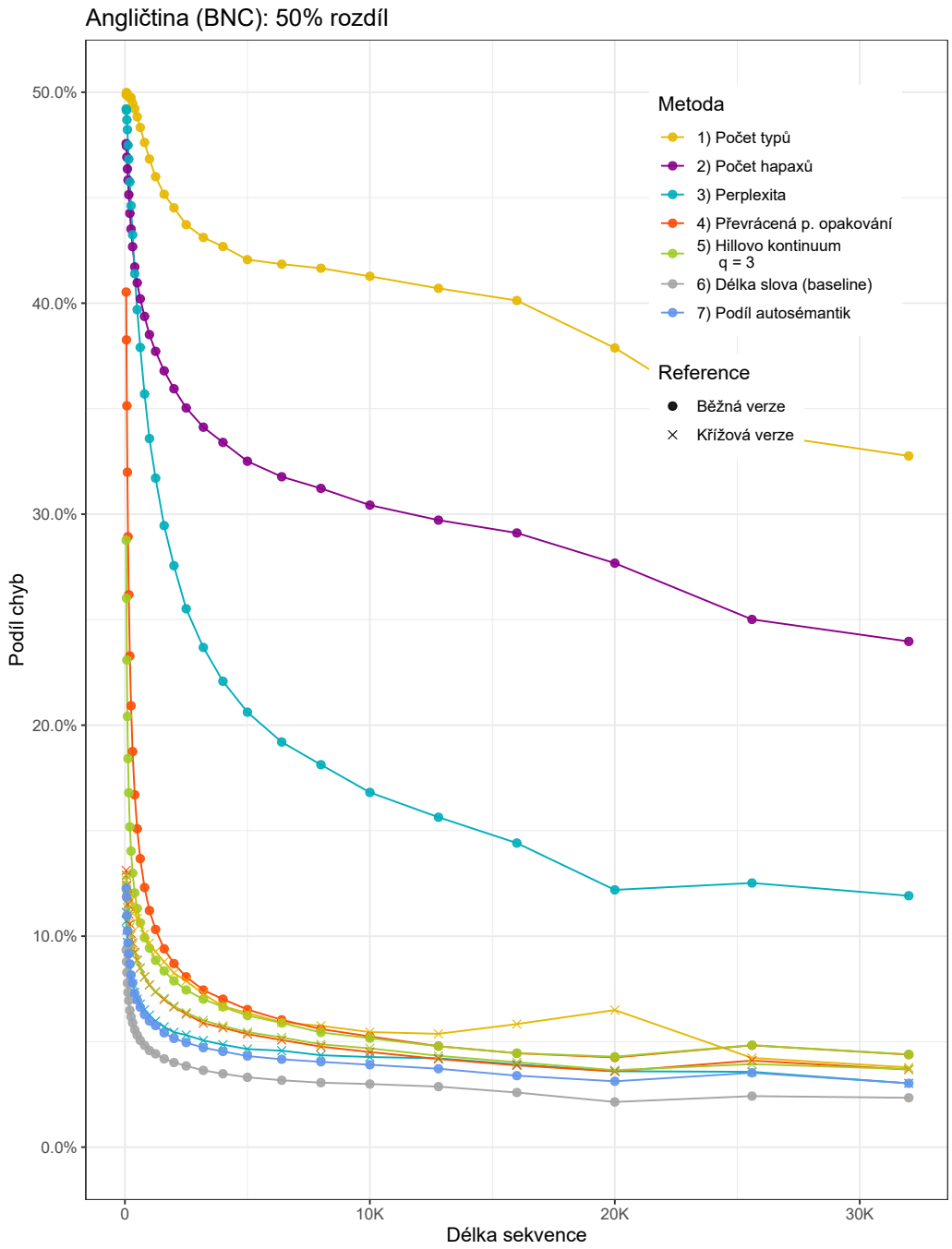
Obrázek 2.2: Srovnání, jak jednotlivé metriky ovlivňuje menší rozdíl v délce sekvence (české texty).



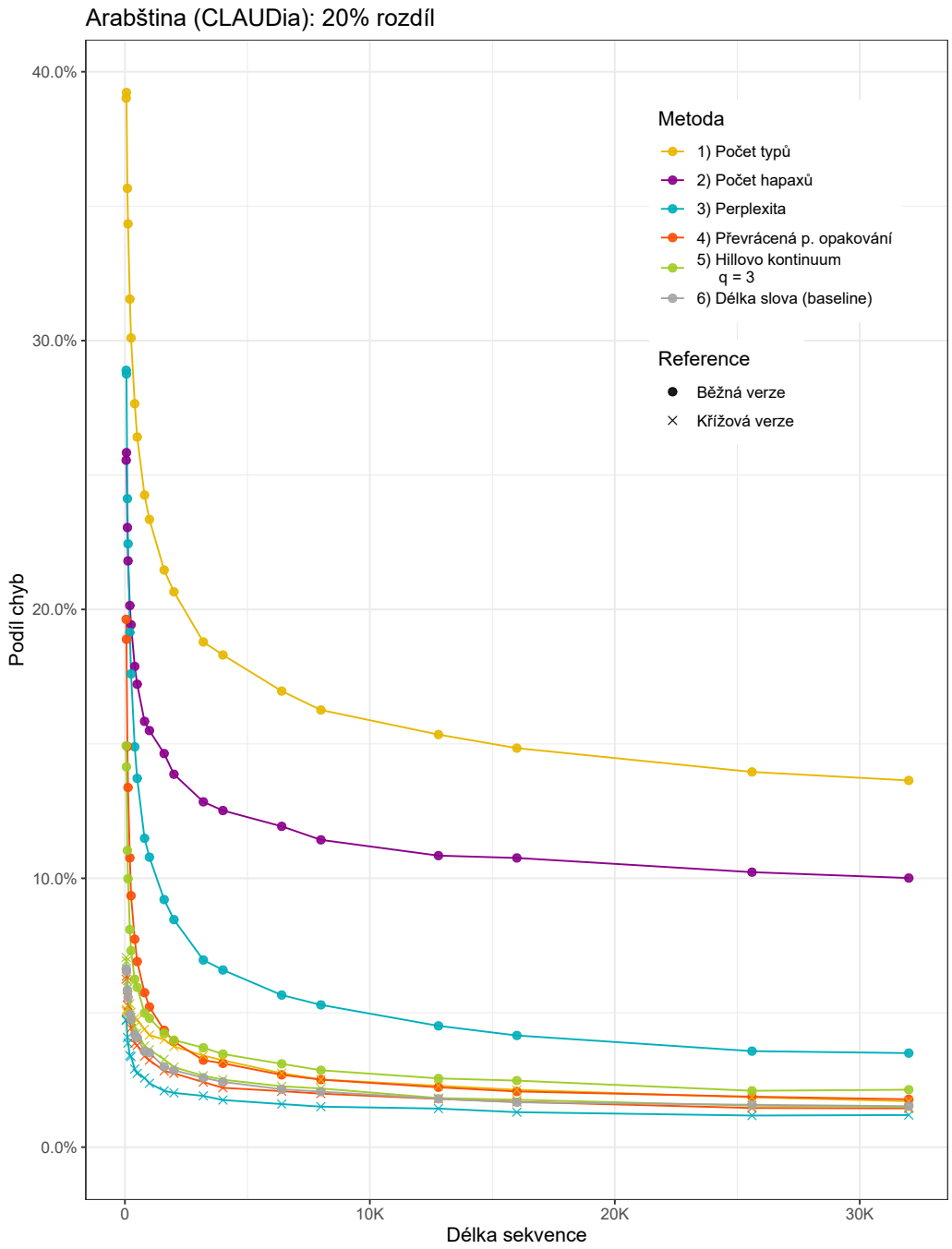
Obrázek 2.3: Srovnání, jak jednotlivé metriky ovlivňuje vetší rozdíl v délce sekvence (české texty).



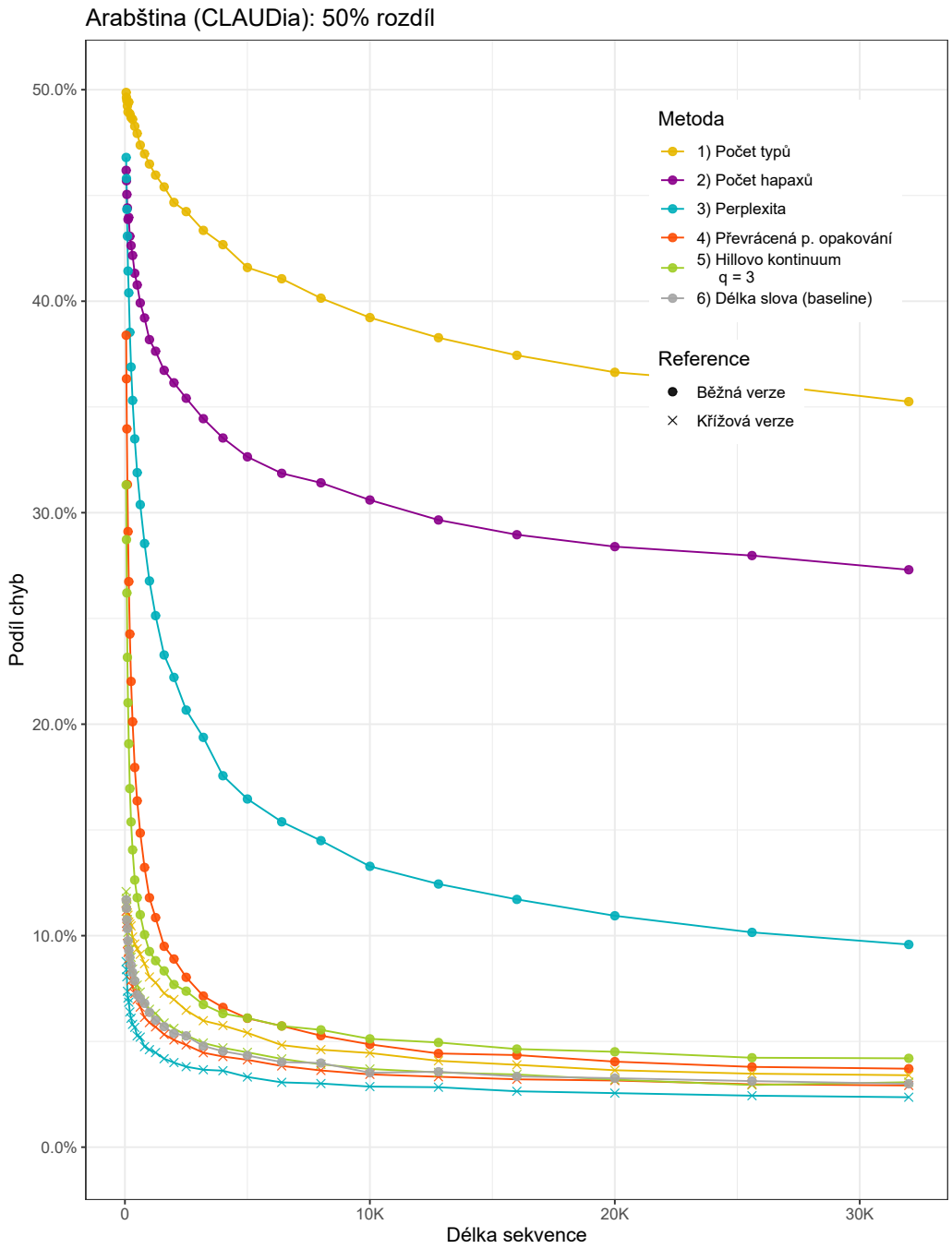
Obrázek 2.4: Srovnání, jak jednotlivé metriky ovlivňuje menší rozdíl v délce sekvence (anglické texty).



Obrázek 2.5: Srovnání, jak jednotlivé metriky ovlivňuje větší rozdíl v délce sekvence (anglické texty).



Obrázek 2.6: Srovnání, jak jednotlivé metriky ovlivňuje menší rozdíl v délce sekvence (arabské texty).



Obrázek 2.7: Srovnání, jak jednotlivé metriky ovlivňuje větší rozdíl v délce sekvence (arabské texty).

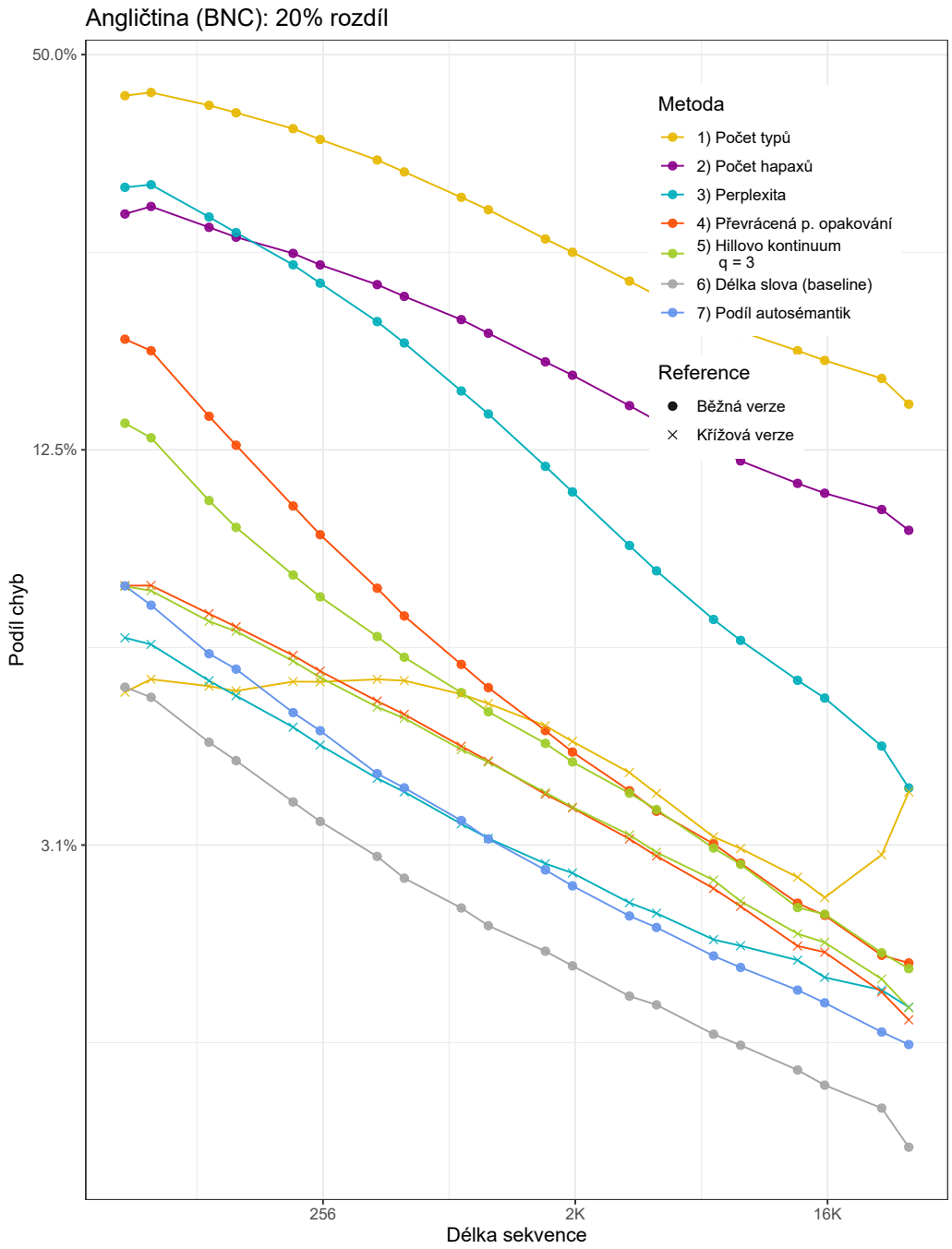
a pro extrémně vysoké hodnoty q bychom mohli očekávat invarianci, neboť nejčastější slovo má relativní frekvenci na délce textu od určitého bodu nezávislou. A skutečně, převrácená pravděpodobnost opakování, kterážto metrika má q rovno dvěma, si vede líp než počet typů a perplexita, a to už od začátku. Ovšem pokud q zvedneme na tři, tak křivka chybovosti již příliš neklesá a v některých případech je dokonce vyšší, ovšem i tak s přibývající délkou okna klesne zhruba do prostoru, kde se pohybují průměrná délka tokenu a podíl autosémantik, tedy metriky z definice na délce textu nezávislé.

Ve stejném prostoru se také, ovšem už od začátku, pohybuje křížová perplexita (potažmo křížová entropie). To také není překvapivé, neboť od křížové entropie můžeme očekávat vlastnosti podobné délce tokenu (za podmínek, že jazyk kóduje slova efektivně), jak jsem psal v podkapitole 1.6.1. Podobně ostatní křížové metriky (křížový počet typů, křížová pravděpodobnost distinkce a křížové Hillovo číslo pro vyšší q) se pohybují nízko a zhruba platí, že čím vyšší je q, tím víc se metrika blíží naší baseline v podobě průměrné délky tokenů. Referenční korpus zde také hraje svou úlohu a podobné výsledky bychom mohli zřejmě očekávat od všech „křížových“ variant ostatních popsaných metrik.

Aby byly grafy aspoň trochu přehledné, nezahrnuji do nich Kullback-Leiblerovy divergence popsaných metrik. Tyto divergence (například relativní perplexita) si systematicky vedou hůř než čistě křížové varianty, což dává smysl — pokud metriku do značné míry invariantní (např. křížová perplexita) vydělíte metrikou závislou na délce sekvence (perplexita), tak jednoduše získáte metriku na délce sekvence závislou.

Tato pozorování platí pro všechny tři jazyky, pouze u češtiny klesají křivky neochotněji a u angličtiny vidíme jakési hrby, které jsou dány rozmanitostí korpusu (BNC obsahuje i vzorky mluveného jazyka, které jsou kratší, a tedy u delších sekvencí se statistik neúčastní). Je podivuhodné, že systematická chybovost u arabštiny je spíš podobná analytické angličtině než syntetické češtině. Vzhledem k tomu, že lexikální diverzita je extrémně ovlivněna morfologií daného jazyka, bych nic takového nepředpokládal a nemám pro to žádné vysvětlení. Bylo by tedy zajímavé rozšířit výzkum na více jazyků a udělat typologickou studii.

Také by stálo za to se podívat, jestli je i v jiných jazycích tento vztah zhruba modelovatelný mocninným modelem. To si demonstrujeme na obrázku 2.8, kde jsou obě osy zlogaritmovány, díky čemuž křivky vypadají opticky lineárně. Nejen opticky, mocninný model jde nafitovat úspěšně na všechny křivky: R^2 je nad 0,95, přičemž zhruba platí, že čím má daná metrika celkově menší chybovost, tím lépe sedí i model, nejlepší fit nalezneme vždy pro metriky, které jsou z definice invariantní vůči délce textu, tedy pro průměrnou délku slova a podíl autosémantik. Výjimkou je křížový počet typů, který se chová poněkud chaoticky, což ovšem může být dáno tím, že metoda výpočtu užívající add-k smoothing není ideální. Na anglických textech je tento vztah viditelný asi nejlíp, ovšem i čeština a arabština vypadají podobně. Netvrdím, že mocninný vztah je nejlepší model, pouze to, že můžeme očekávat, že s přibývající délkou textu bude systematická chybovost způsobená délkou textu klesat zhruba mocninně.

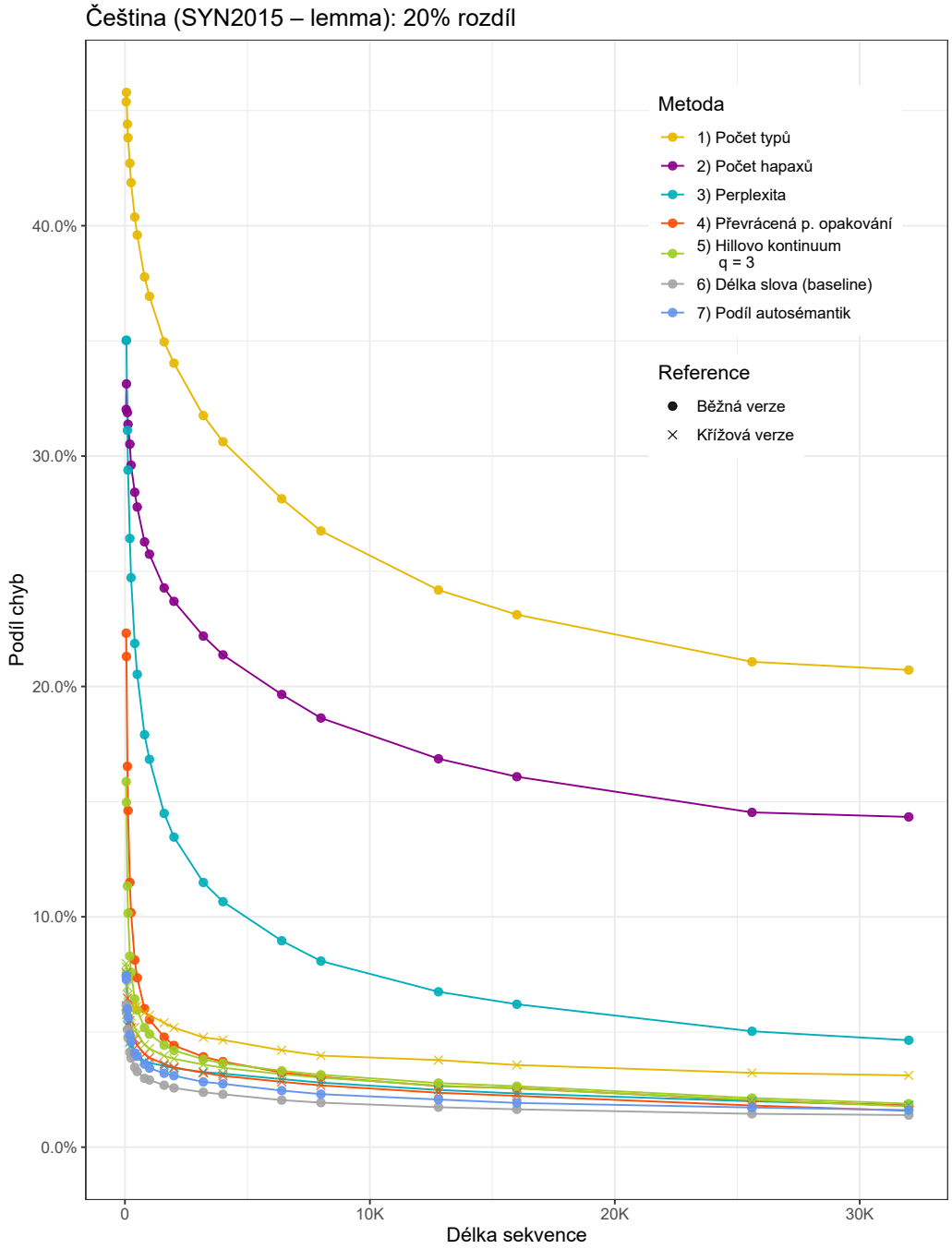


Obrázek 2.8: Srovnání, jak jednotlivé metriky ovlivňuje menší rozdíl v délce sekvence (anglické texty, stejná data jako 2.4, log-log zobrazení).

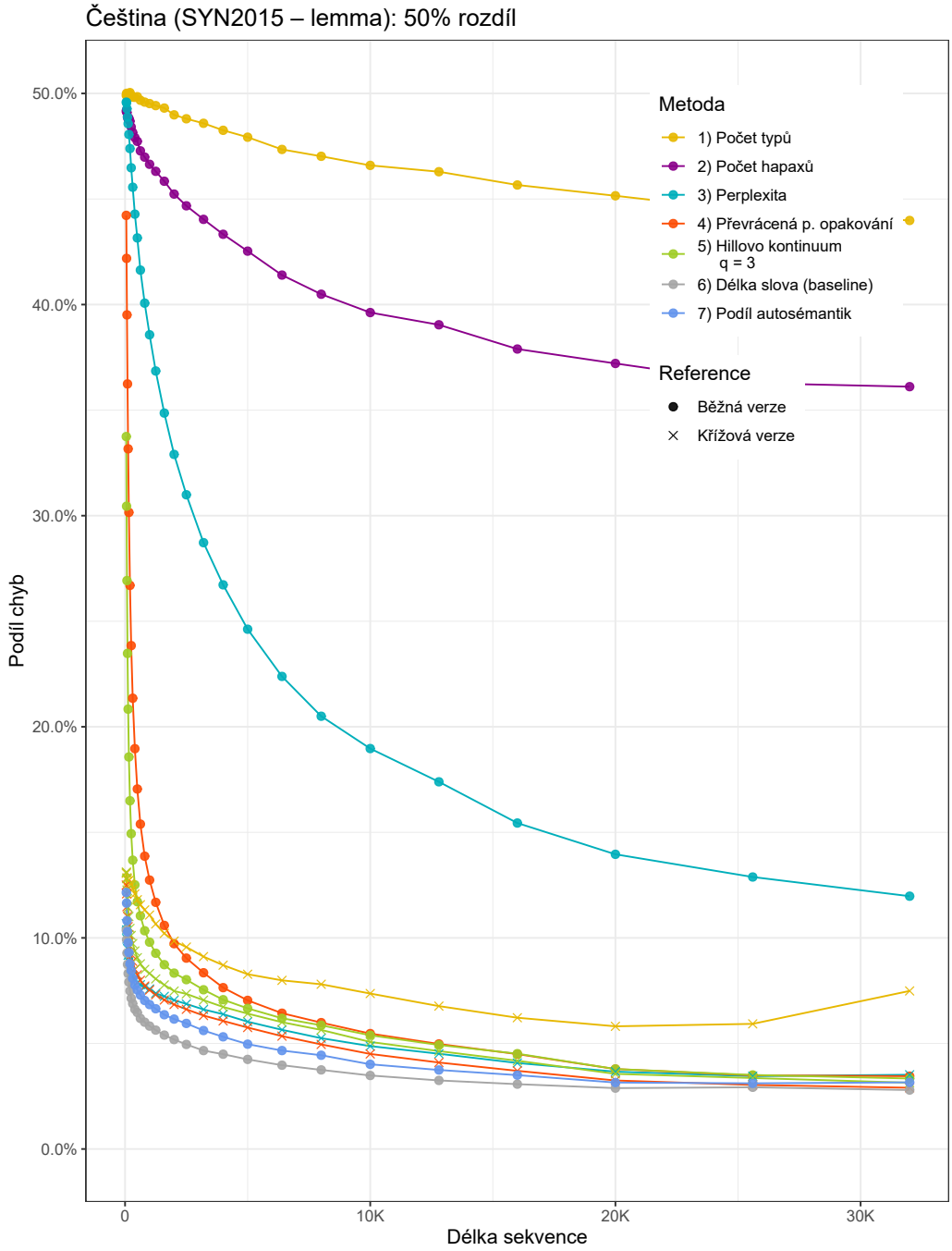
Lemmatizovaný text

Osobně jsem čekal, že po lemmatizaci budou výsledky pro češtinu připomínat víc výsledky pro angličtinu než pro nelemmatizovanou češtinu. Není tomu tak, křivky systematické chybovosti pro češtinu sice klesly mnohem víc než pro angličtinu, jak se na morfologicky bohatý jazyk sluší, nicméně stále jsou velmi podobné křivkám naměřeným na prostém nelemmatizovaném textu. Asi nejdůležitější závěr z grafů 2.9–2.12 je, že lemmatizací si nijak nepomůžeme a metriky lexikální diverzity budou *dále závislé na délce textu či sekvence*.

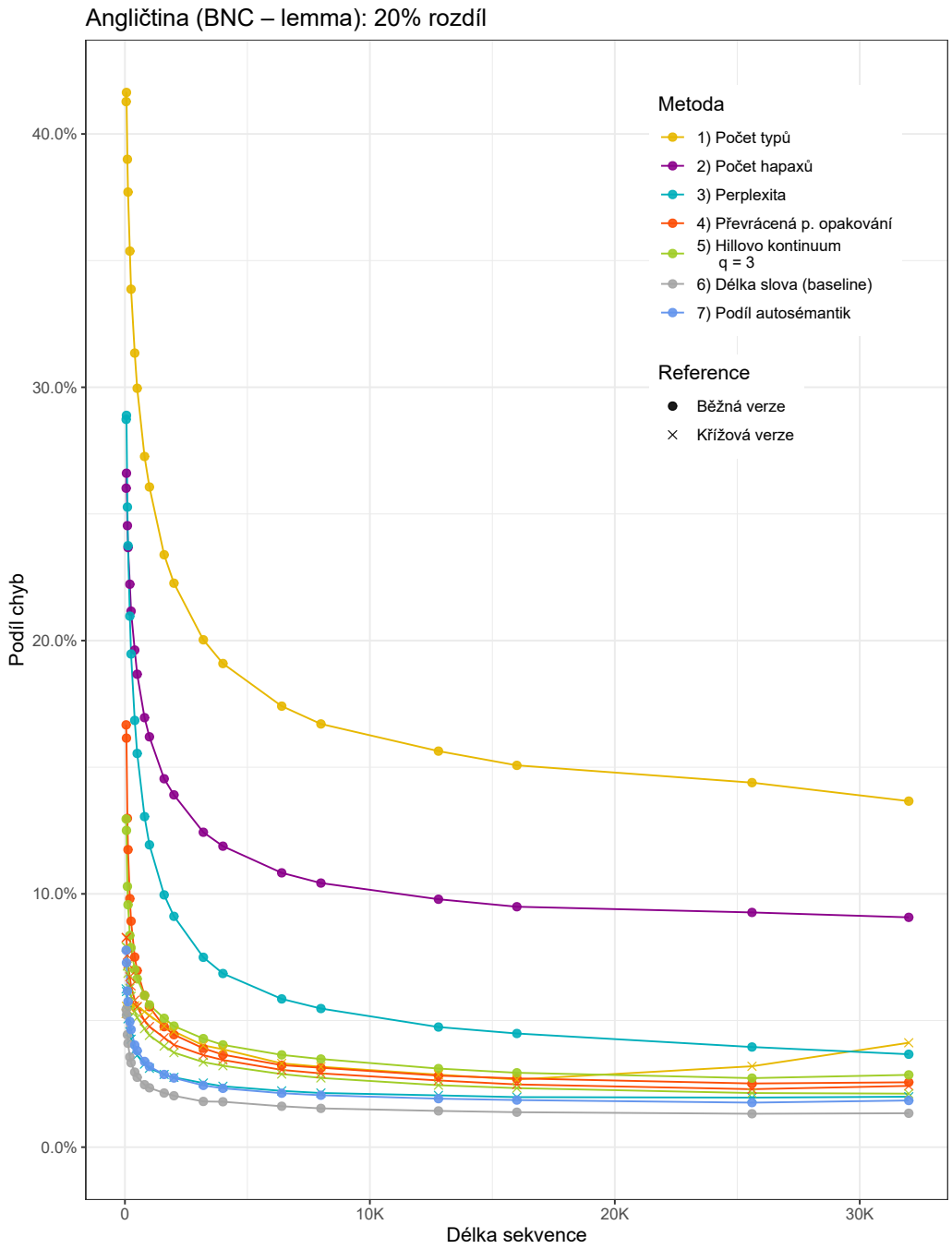
Lemmatizace nás zbavuje spousty informace, obzvlášť ve flektivních jazycích. Trochu se tím zredukuje onen pověstný zipfovský dlouhý ocas v distribuci frekvencí slov — to ovšem platí u dokonale provedené ruční lemmatizace, automatická lemmatizace různé fragmenty, chyby v OCR, nářeční a pravopisné varianty a podobný „nepořádek v datech“ jednoduše považuje za další lemma, čímž jsou výsledky poněkud zkresleny. Ovšem tak jako tak, pokud by někdo předpokládal, že lemmatizace problém délky textu vyřeší, neboť metriky lexikální diverzity začnou na lemmatizovaném textu konvergovat k nějaké hodnotě, mýlil by se.



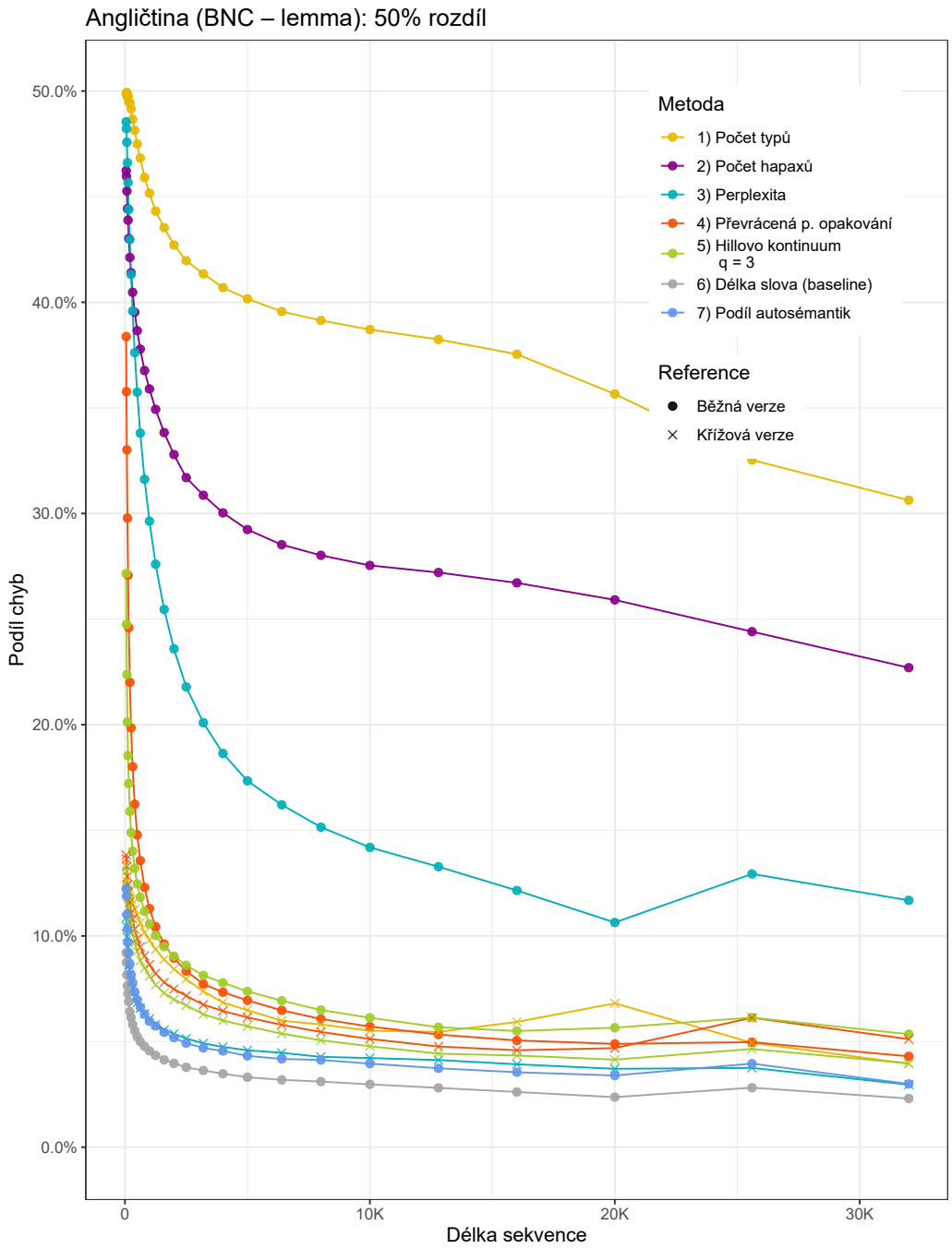
Obrázek 2.9: Srovnání, jak jednotlivé metriky ovlivňuje menší rozdíl v délce sekvence (lemmatizované české texty).



Obrázek 2.10: Srovnání, jak jednotlivé metriky ovlivňuje větší rozdíl v délce sekvence (lemmatizované české texty).



Obrázek 2.11: Srovnání, jak jednotlivé metriky ovlivňuje menší rozdíl v délce sekvence (lemmatizované anglické texty).



Obrázek 2.12: Srovnání, jak jednotlivé metriky ovlivňuje větší rozdíl v délce sekvence (lemmatizované anglické texty).

2.2 Metody normování

Jak je vidět, všechny představené metriky jsou závislé na délce textu a je nutné je nějak normovat, tedy s výjimkou délky slova, podílu autosémantik a s přimhouřením oka křížových metrik.

2.2.1 Segmentace na kratší sekvence

Asi nejjednodušší řešení je vybrat z textu několik vzorků stejné velikosti a ty zprůměrovat. Nemůžeme ovšem vybírat náhodně z textu jednotlivé tokeny, jako bychom vybírali karty z balíčku, neboť, na rozdíl od dobře rozmíchaného balíčku karet, slova jsou za sebou řazena se složitou strukturou. Když mluvčí použije nějaké slovo, zvýší se tím pravděpodobnost, že ho v příštích větách použije znovu. Přirozená sekvence padesáti po sobě jdoucích tokenů tak bude mít jiné vlastnosti, než když z textu vytáhneme padesát tokenů z náhodných míst. Pět set tokenů vytažených z náhodných míst tisícitokenového textu tedy bude mít jiné vlastnosti než pět set obdobně vytažených tokenů z textu násobně většího. Budeme-li se chovat k textu jako k balíčku karet, závislosti na velikosti textu, ze kterého tokeny vybíráme, se tedy paradoxně nezbavíme.[5] Proto je potřeba jako vzorek vybírat *kontinuální sekvence*.

Normování podle tohoto základního principu v sobě obsahuje prvek arbitrárnosti: je nutné zvolit velikost vzorku. Vzhledem k oné zmiňované struktuře pak ovšem velikost vzorku určuje, co vlastně měříme — různé velikosti nám o textu říkají různé věci. Krátké vzorky měří lexikální diverzitu na úrovni vět či odstavců a vlastně zkoumají schopnost či motivaci autora neopakovat stejná slova bezprostředně po sobě a nahrazovat je synonymy, dlouhé vzorky pak měří lexikální diverzitu na úrovni kapitol či delších úseků a spíše ukazují na tematickou rozrůzněnost daného díla. Osobně si ovšem myslím, že toto není nevýhoda, naopak, není třeba hledat nějakou ideální velikost vzorku, ale využít toho, že pomocí různých velikostí můžeme změřit různé kvality textu. Tomuto tématu ostatně dedikuji celou kapitolu 5.

Z bohaté literatury na toto téma jsem si odnesl dojem, že pokud někomu tento způsob normování přinesl nevyhovující výsledky, bylo to primárně kvůli špatně zvolené velikosti vzorku, například Kyle et al. (2021) používali pro normování počtu typů vzorek zvíci padesáti slov, což je asi tak jedna věta, na které se samozřejmě neprojeví prakticky nic.

[5] Metrika založená na takovýchto vzorcích by totiž byla vlastně jen jakousi transformací původního balíčku karet, respektive frekvenční distribuce typů v původním textu. Podobně jako pravděpodobnost opakování nebo perplexita by tak byla závislá na délce textu, pokud by ovšem nebyla znovu nějak normována, což je tak trochu drbání se levou rukou za pravým uchem. Ovšemže to neznamená, že se o to nikdo nepokoušel, slavná a docela používaná je metrika *vocd* (MacWhinney, 2000). Pokud rádi sledujete vědu tak trochu jako sportovní zápas, pak vám pro pobavení doporučuji článek *vocd: A theoretical and empirical evaluation*, ve kterém si McCarthy a Jarvis (2007) autora *vocd* a její proponenty namažou na chleba.

Velikost vzorku je sice arbitrární, nicméně bylo by skvělé, kdyby se stalo zvykem měřit pomocí několika standardizovaných velikostí, například 100–300–1000–3000… Pokud by totiž každá studie používala stejnou sadu velikostí, bylo by možné výsledky mezi jednotlivými studiemi porovnávat (samozřejmě by nebylo nutné se omezovat *pouze* na onu standardizovanou sadu). A také proto, aby si člověk mohl vyvinout intuici, jaká hodnota dané metriky je ve vzorku o dané velikosti obvyklá, co je moc a co je málo. Podobně jako si profesionální potápěč během života utvoří intuici, jak studená je asi tak voda, která má 12 °C, mohla by si pak zasloužilá stylometrička představit text, který obsahuje zhruba dejme tomu 500 slovních typů v tisíci tokenech.

Nepřekrývající se sekvence

Průkopníkem segmentace na kratší sekvence byl už v roce 1944 Wendell Johnson, který definoval MSTTR, tedy *mean segmental type token ratio*.[6] Jde vlastně o text rozdělený na několik stejně velkých částí, přičemž na každé z nich je samostatně změřen počet typů.[7] Ovšemže není důvod, proč jej nepoužít i na ostatní metriky lexikální diverzity — mean segmental perplexity, mean segmental dissimilarity atd.

Tento způsob normování není ekvivalentní vybírání náhodných vzorků z textu, a pokud je délka textu srovnatelná s délkou vzorku, tak vzorků může být jen velmi málo, což ovšem znamená, že výběr je značně ovlivněn náhodou. Rozsáhlejší kritiku této metody podávají Malvern a jeho spoluautoři (2004, str. 95–98). Prakticky všechny body jejich kritiky jsou řešitelné tak, že provedeme skutečně náhodný výběr. Což je prakticky ekvivalentní tomu, když zprůměrujeme všechny překrývající se sekvence.

Překrývající se sekvence

Míst, odkud můžeme náhodně vybrat sekvenci délky G z textu o délce N, je pouze $N - G + 1$, není tedy problém zprůměrovat *všechny možné vzorky o dané délce*. To je vlastně ekvivalentní situaci, kdy danou metriku změříme v postupně se posouvajícím okně o velikosti G, tedy například pokud má být vzorek velikosti 500 tokenů, tak metriku změříme nejprve na prvních pěti stech tokenech, pak okno posuneme o token dál a změříme ji na druhém až pětistém prvním tokenu atd. Výsledky pro všechna okna následně zprůměrujeme.

[6] Sborník, ve kterém byla tato metodologie poprvé prezentována, obsahuje empirické studie tří různých autorek a autorů, kteří MSTTR používají (Fairbanks, 1944; Mann, 1944; Chotlos, 1944) a Johnson má pouze úvodní slovo (Johnson, 1944), není tak úplně jasné, jestli je autorem právě on. Každopádně i kdyby byl, stejně zůstane slavnější jako autor experimentu, při kterém se pokusil několik nekoktajících dějí naučit koktat, k jejich smůle úspěšně (Goldfarb, 2005). Mimochodem, terminologickou diverzitu indexů lexikální diverzity ilustruje, že tato metrika je v nástroji WordSmith nazvána jako nicneříkající *standardised type/token ratio* (STTR) (Scott, 2010).

[7] Tento počet typů je následně z magických důvodů podělen počtem tokenů, k tomuto fenoménu viz kapitolu 1.10.1. Nicméně na tom nesejde, hlavní je princip normování pomocí rozdělení textu na vzájemně se nepřekrývající sekvence.

Troufám si tvrdit, že tato metoda normování nejenže jednoduše, přesvědčivě a trvale řeší problém závislosti na délce textu,[8] ale hlavně je prakticky intuitivní. Proto je fascinující, že byla poprvé zavedena teprve až v jednadvacátém století (Covington – McFall, 2010). Podle klouzavého okna je pojmenována jako *moving average type-token ratio* (MATTR),[9] ovšem, podobně jako v předchozím případě, neexistuje důvod, proč by se nedala použít na ostatní metriky, a vytvořit tak třeba *moving average entropy* nebo *moving average Kullback-Leibler divergence*. To je vlastně druhá fascinující věc, tato metoda měla deset let na to, aby se rozšířila i na ostatní metriky, to se ovšem nestalo a v literatuře dále najdeme další a další neúspěšné a především zbytečné pokusy o hledání metriky lexikální diverzity přirozeně nezávislé na délce textu (například Shi – Lei (2022)).

Je asi spíš otázka pro historika či sociologa vědy, proč trvalo 70 let, než byla nedokonalá Johnsonova metoda nepřekrývajících se segmentů nahrazena, a proč je literatura plná desítky let trvajícího honu na metriku přirozeně nezávislou na délce textu, honu nesmyslného tím evidentněji, čím déle trval. Přitom mnozí nebyli daleko: Köhler – Galle (1993) zkoumali počet typů v sériích překrývajících se oken, ale poslední krok ke zprůměrování již neučinili. Durán et al. (2004) navrhli používat náhodný výběr (což je ekvivalentní pohyblivému oknu), ale okamžitě tento nápad bez argumentů zase zavrhli. Mám podezření, že zoologická zahrada špatně interpretovatelných metrik lexikální diverzity výzkumníkům vlastně vyhovovala, neboť se zvyšovala šance, že aspoň jeden z indexů bude mít pozitivní výsledky — tento obecně metodologický problém je ve statistice dobře popsán a je znám pod názvem *garden of forking paths* (Gelman – Loken, 2013).

Na tomto místě je třeba upozornit, že jednotlivé náhodné vzorky na sobě nejsou nezávislé, není tedy možné na nich dělat přímo statistické operace, které vyžadují nezávislost vzorků. Pokud tedy například chceme získat konfidenční intervaly pomocí resamplingu, je třeba s tím počítat a resamplovat vzorky tak, aby se nepřekrývaly. Takže vlastně resamplujeme z nepřekrývajících se sekvencí, jak je popisuji v předchozí kapitole.[10] Abychom vyrovnali nevýhody z toho plynoucí (například možný malý počet vzorků, ze kterého resamplujeme), můžeme resampling provést na všech možných konfiguracích nepřekrývajících se sekvencí. Tedy například napřed bootstrappujeme konfidenční intervaly resamplováním nepřekrývajících se sekvencí obsahujících tokeny 1–500, 501–1000, 1001–1500 atd., následně resamplujeme sekvence obsahující

[8] Metrika normovaná pomocí klouzavého okna je nezávislá na délce textu už z definice, pokud by však přeci jen někdo pochyboval, onu nezávislost přesvědčivě ukázal na datech Kubát (2014).

[9] Všimněte si, že nejde o moving average number of types, ale že bylo opět, z magických a numerologických důvodů, nutné podělit počet typů počtem tokenů, přestože je počet tokenů specifikován délkou okna.

[10] K tomuto tématu doporučuji Evert et al. (2017).

tokeny 2–501, 502–1001 atd., následně totéž provedeme se sekvencemi 3–502, 503–1002… a nakonec konfidenční intervaly získané pro jednotlivé konfigurace zprůměrujeme.

2.2.2 Srovnání s referenčním korpusem

Když zjistíme, že v česky psaném textu o dejme tomu 1258 tokenech je 875 slovních typů, obvykle nám tohle číslo nic neřekne. Je to moc, nebo málo? Moc v porovnání s čím? Rádi bychom měli nějaký referenční bod. Zjištění, že tento text má slovní bohatství větší než 86 % textů stejné délky a stejného žánru v korpusu SYN2015, je mnohem informativnější.

A právě tento způsob relativizace je možno používat i na normování podle velikosti. Můžeme totiž z korpusu vytáhnout velké množství vzorků dané velikosti (respektive překrývajících se sekvencí dané velikosti, jako v předchozí kapitole) a náš text vůči těmto vzorkům vztáhnout. Jediné, na co si musíme dávat pozor, je, aby naše vzorky nešly přes hranice textu. Můžeme pak porovnávat i texty různé délky: pokud má text A větší perplexitu než 88 % vzorků stejné délky, pak je lexikálně diverzifikovanější než text B, který má perplexitu větší než 43 % vzorků stejné délky.

Problematické je, že musíme nějak zvolit referenční korpus, což je nutně arbitrární a arbitrárnost je při měření nevítaná, neboť snižuje dojem objektivnosti, ač intersubjektivitu si taková metrika zachovává. Nejspíš by bylo nutné ustanovit nějaké standardní referenční korpusy, aby se daly vzájemně porovnávat výsledky v různých studiích.

Možná to je důvod, proč tento jednoduchý způsob normování nikdo nepoužívá a proč ještě nebyl popsán v literatuře — tedy pokud vím. Ona nejistota je na místě, neboť spousta metrik a originálních metodologických přístupů týkajících se lexikální diverzity je pohřbená v nedostupných sovětských sbornících nebo v jazycích jako estonština (Tuldava, 1997) nebo čeština (Milička, 2022).[11] V souladu s touto úctyhodnou tradicí je důmyslně ukryt i článek Cvrček – Chlumská (2015), ve kterém autorská dvojice navrhuje něco podobného: porovnávat slovní bohatství textu s referenčním korpusem pomocí standardní odchylky.[12] Díky publikování v časopise, který se tématu nevěnuje, v článku, jehož primárním cílem bylo něco úplně jiného, nadto

[11] Díky tomu ovšem může každá generace pocítit onu radost z objevování a pojmenovávat stále stejné indexy po dalších a dalších mužích. Všimněte si, že je to už podruhé, co jsem vyjádřil tuto myšlenku a explicitně při tom zmínil pojmenovávání po *mužích*. Skutečně, během celé (a věřte mi, že dlouhé a nudné) heuristiky k této knize jsem nenarazil na jediný index pojmenovaný po ženě. Přitom žen věnujících se stylometrii byla a je celá řada.

[12] Metriku pojmenovali jako zTTR, neboť, hádáte správně, před porovnáním s referenčním korpusem je počet typů vydělen počtem tokenů. I když se tato operace při následujících úpravách neutralizuje, předpokládám, že magický účinek přetrvá, podobně jako sladká voda, která se kdysi dotkla molekul pocházejících z jater pižmovky, má mnohem větší účinek proti rýmě než sladká voda, která to štěstí neměla, můžeme-li ovšem věřit výrobcům přípravku Oscillococcinum.

opatřeného názvem, který jen naznačuje, co za poklady vlastně ukrývá, tato myšlenka vzbudila jen malý ohlas a metodologie prakticky nebyla použita znovu. Jsem poněkud skeptický k užívání standardní odchylky, neboť distribuce, kterou s její pomocí charakterizujeme, není normální (jak ostatně autoři přiznávají). Nicméně samotná myšlenka užití referenčního korpusu si podle mě zaslouží mnohem větší pozornost.

2.2.3 Měření podle parametrů modelu

Konečně se dostáváme k metodě normalizace, která byla v dějinách lexikální diverzity nejoblíbenější a kterou zároveň považuji, ve světle normování pomocí klouzavého okna, za překonanou, totiž že namísto vlastnosti, kterou chceme zkoumat (například počet typů v textu) a která závisí na délce textu, zkoumáme model oné závislosti, respektive charakterizujeme lexikální diverzitu daného textu pomocí parametrů onoho modelu.

Když to napíšu takto, tak to asi zní poněkud komplikovaně, ale představte si běžné TTR (k němu podrobně 1.10.1), které není ničím jiným než parametrem lineárního modelu vztahu mezi typy a tokeny.

Tento přístup stojí za mnoha indexy lexikální diverzity, které byly svými tvůrci prezentovány jakožto nezávislé na délce textu (viz 1.10), byť model byl někdy implicitní, nepřiznaný (jako třeba v případě lambdy, 1.10.3). Obvykle se jednalo o model o jednom parametru, což samozřejmě znamenalo, že neseděl příliš dobře a že metrika se časem ukázala jako na délce textu závislá, což pro ni byl smrtící úder, neboť ona nezávislost byla její hlavní devízou a smyslem existence.[13] Nicméně není důvod, proč by nebylo možné charakterizovat lexikální diverzitu nějakého textu modelem o několika parametrech, taková metrika by naopak mohla být docela zajímavá — pokud by hodnoty parametrů mezi sebou nekorelovaly, ukazovaly by pak každá na jiný aspekt lexikální diverzity. Ovšem nevím o tom, že by někdo víceparametrový model takto používal.

Model o více parametrech by však vyžadoval, aby byl nafitován nikoli pomocí jednoho datového bodu (například počtu typů v celém textu), ale podle více datových bodů na více místech v textu (například celé křivky type-token relation), což je sice procedura, kterou Wimmer s Altmannem (1999) doporučují i pro modely o jednom parametru, ovšem reálně se snad nikdy nepoužívala.

Není náhoda, že jsem zmínil zrovna model pro počet typů. Drtivá většina metrik založených na tomto principu se nějak vztahuje k počtu typů, slovnímu bohatství. Teoreticky bychom sice mohli měřit slovní diverzitu podle parametrů distribuce typů, tedy například podle parametrů klasického zipfovského modelu, což navrhuje už starý dobrý Good (1953, s. 249), nicméně tento přístup se nikdy neujal.

Základní problém normování podle modelu je, že nám chybí spolehlivé modely.

[13] S čestnou výjimkou TTR, které nás přežije všechny.

A vlastně i obecná metodologie, jak ony modely hledat. Mohli bychom samozřejmě spustit nějakou automatickou heuristiku pro hledání symbolické regrese s nejlepším fitem na rozmanitá textová data, problém je, že tento přístup má tendenci k overfittingu na prezentovaná data a modely jsou navíc jen stěží interpretovatelné. Raději bychom našli nějaký lingvisticky plauzibilní stochastický proces popisující, jak je text vytvářen, od něj nějaký model odvodili a teprve poté jej testovali na empirických datech. Tedy postupovali bychom od ověřených obecných principů deduktivně dolů, od teorie k modelu. Dokážeme se ale vůbec shodnout na oněch základních principech? Například podle již zmiňované dvojice Wimmer – Altmann (1999) se teoreticky vhodný model pro počet typů v textu pozná podle toho, že počet slov je asymptoticky omezený,[14] zatímco podle mě ve světle našich empirických poznatků o dynamice přirozeného jazyka je to přesně naopak (Milička, 2013) a počet typů by měl asymptoticky směřovat do nekonečna.

Dalším, ovšem možná ještě důležitějším problémem je interpretace výsledků. Pokud vám někdo řekne, že viděl text o tisíci slovech, který obsahoval 500 slovních typů, tak i když si nedokážete přesně představit, co to znamená, ona hodnota má nějakou rozumnou jednotku (počet typů a tokenů) a můžete najít texty, které mají podobné kvality a podívat se, jak vypadají. Pokud vám řekne, že TTR nějakého textu bylo 0,5, je to o poznání horší, neboť se jedná o bezrozměrné číslo, které se sice dá interpretovat tak, že po odebrání duplicit zbude v textu polovina slov, ovšem bez znalosti délky textu to může znamenat leccos. Když ale někdo napíše, že logTTR daného textu je 0,9, tak nevím jak vy, ale já si třeba nepředstavím vůbec nic.

Konečně posledním a nejhůře řešitelným problémem je, že lexikální diverzita krátkých sekvencí je vlastně úplně jiná veličina než lexikální diverzita sekvencí dlouhých. I kdybychom z hlubokých tůní kvantitativní lingvistiky vylovili žábu, která by se po políbení proměnila v dokonalou metriku lexikální diverzity, jež by jedním krásně interpretovatelným číslem charakterizovala celý text nezávisle na jeho délce, stejně bychom pomocí ní nemohli srovnávat román o dvou stech stranách s románem tisícistránkovým — už jen proto, že autor onoho tisícistránkového románu měl mnohem víc příležitostí změnit téma i styl. Vlastně bychom onu naprosto dokonalou metriku museli stejně nějak normovat. Osobně bych to udělal pomocí metody klouzavého okna.

[14] Doslova píšou „as a matter of fact, we need a curve with finite asymptote, good linguistic interpretability and meaningful parameters“ (Wimmer – Altmann, 1999, str. 6–7), s posledními dvěma body se dá souhlasit.

Kapitola 3

Škálování

Představte si, že chcete ozdobit vánoční stromeček tak, že na nejspodnější větve přijdou největší koule a na vrchol ty nejmenší.[1] Když je začnete řadit od největší po nejmenší, je úplně jedno, jestli jako metriku velikosti vezmete jejich poloměr, průměr, povrch, objem, či třeba logaritmus obvodu jejich řezu v nejširším místě. Všechny tyto metriky jsou tedy dobrým *indexem* velikosti koule. A protože všechny metriky řadí koule stejně, vlastně ani nemusíme řešit, která z nich je nejlepší metrikou velikosti.

V reálném životě nás ale většinou nezajímá jen, která koule je větší, ale jak moc je větší. V tu chvíli ovšem dává smysl otázka, co se tou velikostí vlastně myslí. Pokud někdo má kulovitý vodojem a chce si koupit dvojnásobně velký, nebude tím asi myslet dvojnásobný průměr, ale objem.

Podobně je to s metrikami lexikální diverzity. Dejme tomu, že se chceme podívat, jestli žáci v osmé třídě píší slohovky s menší lexikální diverzitou než jejich stejně staří kolegové na víceletých gymnáziích. Dokud chceme jenom vědět, jestli má rozdělení na devítiletky a gymnázia *vůbec nějaký efekt*, tak je nám jedno, jestli použijeme pravděpodobnost distinkce, či převrácenou pravděpodobnost opakování, jestli dáme přednost entropii, či perplexitě. Ale pokud nás zajímá *velikost efektu*, tak už to najednou vůbec jedno není.

Jost (2007) uvádí krásný příklad, kdy někdo vyvraždí polovinu všech živočišných a rostlinných druhů na kontinentu, načež Gini-Simpsonův index (t.j. pravděpodobnost distinkce) nonšalantně klesne z 0,99999997 na 0,99999993. Převrácená pravděpodobnost opakování, tedy tatáž metrika v jiném škálování, ovšem klesne ze zhruba třiatřiceti milionů efektivních druhů na čtrnáct milionů, což tak nějak intuitivně souzní s rozsahem katastrofy.[2]

Co vlastně považujeme za intuitivní škálování lexikální diverzity? Intuitivnost je

[1] Uznávám, že tohle není zrovna pravděpodobný scénář, ale jiný důvod, proč by někdo chtěl porovnávat velikost koulí, mě nenapadla. Zároveň jsem nucen moravské čtenáře informovat, že tady v Praze se na stromek skutečně místo baněk věší koule, jakkoli se to může zdát neuvěřitelné.

[2] Jost věnoval tématu diverzity (z pohledu ekologie) značnou energii a chystá se publikovat na toto

poněkud tenký led, neboť je to záležitost navýsost subjektivní. Což ovšem neznamená, že bychom se na nějakých takovýchto kritériích nemohli vespolek shodnout. Hill, ve stejném článku, ve kterém představuje své kontinuum metrik (Hill, 1973), navrhuje dvě taková základní kritéria, která by měla splňovat každá metrika lexikální diverzity: když všechny tokeny ve vzorku patřící ke stejným typům rovnoměrně rozdělíme na dva typy, měla by se hodnota metriky lexikální diverzity také zdvojnásobit. Tedy například pokud bychom měli text, každý druhý token obarvili červeně a následně se rozhodli, že červené tokeny patří k jinému typu než černé, tak se nám najednou zdvojnásobí počet typů. Také se zdvojnásobí perplexita (a entropie se zvýší o jeden bit). Zdvojnásobí se i převrácená pravděpodobnost opakování… a vůbec všechny metriky Hillova kontinua, které je definováno právě tak, aby vyhovovalo tomuto podle mě velmi rozumnému kritériu.

Druhým kritériem je, že pokud mají všechny typy ve vzorku stejnou četnost, pak by se hodnota dané metriky měla rovnat počtu typů. Metriky Hillova kontinua splňují také tuto podmínku, což znamená, že jsou vyjádřeny ve stejné jednotce: efektivních typech.

Pokud tedy považujete škálování obyčejného počtu typů (alias slovního bohatství) za intuitivní, budete také považovat za intuitivní škálování všech ostatních metrik Hillova kontinua — což považuji za velmi důležité pro běžný vědecký provoz. Nyní se pustíme do empirického ověření předpokladu, že metriky Hillova kontinua spolu škálují lineárně, což by z jeho definice mělo vyplývat. Zároveň se podíváme, jaké postavení zaujímají křížové verze Hillova kontinua, zejména křížová perplexita. A také nás bude zajímat, jestli s počtem typů lineárně škálují i ostatní metriky, které jsme představili v 1. kapitole.

3.1 Metodika

Stejně jako v následujících kapitolách, i zde jsem vycházel ze vzorku několika tisíc sekvencí o tisíci slovech náhodně vytažených z korpusu tak, že respektují hranice textů. Na těchto sekvencích jsem vždy změřil metriky, které nás zajímají, a následně výsledky porovnal.

Porovnávám vždy původní, nijak netransformovanou metriku, dále zlogaritmované hodnoty této metriky (tedy $\log_2(\text{metrika})$) a následně exponenciálu dané metriky (2^{metrika}). Základní je porovnání s počtem typů, neboť právě k němu se teoreticky vztahujeme všemi metrikami, které jako jednotku používají efektivní počet typů, ale

téma monografii. Pokud tyto řádky čtete ještě před jejím vydáním a nevlastníte stroj času, můžu prozatím doporučit jeho poněkud chaotické, nicméně inspirativní stránky `http://www.loujost.com/Statistics%20and%20Physics/Diversity%20and%20Similarity/DiversitySimilarityHome.htm`.

protože křížové metriky s počtem typů korelují velmi málo v jakékoliv variantě, podíváme se zejména na souvislosti křížových metrik s jejich nekřížovými variantami, kde je korelace o poznání lepší. Tedy, některé dvojice metrik spolu ani tak příliš nekorelují, což ovšem není žádné velké překvapení a je to vlastně důvod, proč vůbec víc metrik používáme. Nicméně i při těchto slabších korelacích zkouším zjistit, jestli je lineární korelace vhodnější pro přirozenou, logaritmizovanou, nebo exponenciovanou variantu.

Dále se díváme na distribuci výsledků pro danou metriku a měříme její koeficient šikmosti (skewness) s předpokladem, že příliš asymetrická distribuce může vzniknout vlivem nevhodného škálování. Obrázky distribucí pro jednotlivé metriky zde pro úsporu místa neuvádím, neboť jsou zobrazeny na okrajích grafů v následující kapitole (5), a to pro větší počet textových typů a různé délky sekvencí.

Korelaci si také můžete opticky ohodnotit na grafech, kde každý bod znázorňuje jednu sekvenci. Abych zdůraznil, jestli jsou dvě zobrazené metriky v lineárním vztahu, nechal jsem v grafech zobrazit lineární regresi (zelená linka) a lokální regresi vypočítanou metodou LOESS (modrá linka).

3.2 Metriky Hillova kontinua

Už při letmém pohledu na první řádek tabulky 3.1 nás napadne, že rozdíly mezi transformovanými metrikami nejsou nijak závratné: počet typů celkem nepřekvapivě stoprocentně koreluje se sebou samým, nicméně docela dobrá je i jeho lineární korelace s logaritmizovaným a exponenciovaným počtem typů (0,997). Rozdíly tedy můžeme očekávat docela subtilní.

Potěší, že u všech metrik Hillova kontinua (tedy perplexita, RRR a Hillovo číslo s vyšším koeficientem q, které nemá žádný tradiční název) je u všech tří jazyků (tabulky 3.1, 3.2 a 3.3) buď největší korelace u nijak netransformovaných hodnot, nebo jsou alespoň rozdíly mezi korelačními koeficienty mezi netransformovanou (přirozenou) a logaritmizovanou variantou velmi malé. Také distribuce jsou nejsouměrnější právě pro netransformované hodnoty, popřípadě jsou také rozdíly mezi koeficienty šikmosti velmi malé.

Podobně se chová i počet hapax legomena, který sice k Hillovu kontinuu neřadíme, ovšem s počtem typů lineárně koreluje velmi dobře a v netransformované variantě má hezky symetrickou distribuci.

Tato zjištění ilustrují grafy na obrázcích 3.1–3.4, kde opticky nejlineárnější mi přijdou vždycky grafy nejvíc vlevo, tedy aspoň tam, kde o nějaké korelaci můžeme vůbec mluvit.

Můžeme tedy uzavřít, že dané metriky spolu škálují velmi dobře bez nutnosti transformace, přesně podle teoretických předpokladů.

	Šikmost (skewness)			Pearsonův korelační koef.		
Metrika	přiroz.	log	exp	přiroz.	log	exp
Počet typů	-0.47	-0.956	25.6	1	0.997	0.997
Počet hapax legomena	-0.308	-1.02	45.3	0.984	0.981	0.978
Perplexita	-0.123	-0.743	53.5	0.953	0.955	0.945
RRR	0.251	-0.485	30.5	0.68	0.688	0.672
Hillovo číslo (q = 3)	0.328	-0.429	20.1	0.461	0.47	0.456
Křížový počet typů	4.5	0.0544	NaN	0.368	0.471	0.351
Křížová perplexita	14.8	0.884	NaN	0.416	0.707	0.396
Křížová RRR	0.967	0.345	54.7	0.408	0.424	0.387
Křížové HČ (q = 3)	0.535	0.231	25.8	0.178	0.19	0.162
Délka tokenu	0.243	0.057	0.328	0.753	0.758	0.745
Rozdílnost	0.375	0.353	0.377	0.212	0.213	0.209
Podíl autosémantik	-0.0713	-0.568	-0.053	0.629	0.625	0.611

Tabulka 3.1: Český korpus (SYN2015Fic): Šikmost distribucí jednotlivých metrik a jejich lineární korelace s počtem typů.

	Šikmost (skewness)			Pearsonův korelační koef.		
Metrika	přiroz.	log	exp	přiroz.	log	exp
Počet typů	-0,288	-0,909	24,6	1	0,995	0,995
Počet hapax legomena	-0,121	-1,16	33,1	0,983	0,974	0,974
Perplexita	-0,111	-0,997	25,1	0,889	0,885	0,884
RRR	-0,467	-1,18	7,6	0,17	0,171	0,188
Hillovo číslo (q = 3)	-0,312	-0,976	4,93	-0,084	-0,074	-0,062
Křížový počet typů	2,27	-0,252	NaN	0,524	0,547	0,494
Křížová perplexita	2,78	0,854	NaN	0,647	0,715	0,606
Křížová RRR	2,75	0,249	54,7	-0,061	-0,058	-0,054
Křížové HČ (q = 3)	0,806	0,035	54,7	-0,263	-0,257	-0,246
Délka tokenu	0,234	0,059	0,305	0,661	0,664	0,638
Rozdílnost	-0,349	-0,37	-0,347	-0,145	-0,144	-0,136
Podíl autosémantik	0,538	0,295	0,548	0,622	0,624	0,593

Tabulka 3.2: Anglický korpus (BNCWrittenFic): Šikmost distribucí jednotlivých metrik a jejich lineární korelace s počtem typů.

Metrika	Šikmost (skewness) přiroz.	log	exp	Pearsonův korelační koef. přiroz.	log	exp
Počet typů	0,071	-0,458	54,7	1	0,991	0,991
Počet hapax legomena	0,412	-0,41	54,1	0,981	0,973	0,959
Perplexita	0,14	-0,765	54,7	0,973	0,947	0,96
RRR	0,205	-1,05	54,7	0,851	0,813	0,844
Hillovo číslo (q = 3)	0,142	-1,12	54,7	0,718	0,702	0,724
Křížový počet typů	15,3	-0,070	NaN	0,439	0,691	0,403
Křížová perplexita	50,7	0,645	NaN	0,106	0,851	0,088
Křížová RRR	37,3	0,477	NaN	0,333	0,675	0,32
Křížové HČ (q = 3)	22,9	0,932	54,7	0,385	0,544	0,384
Délka tokenu	0,535	0,255	0,675	0,678	0,683	0,675
Rozdílnost	-1,32	-1,39	-1,32	0,359	0,36	0,386

Tabulka 3.3: Arabský korpus (CLAUDia): Šikmost distribucí jednotlivých metrik a jejich lineární korelace s počtem typů.

Metrika	Čeština přiroz.	log	Angličtina přiroz.	log	Arabština přiroz.	log
(Křížový) počet typů	0,368	0,471	0,524	0,547	0,439	0,691
(Křížová) perplexita	0,45	0,744	0,579	0,639	0,127	0,878
(Křížová) RRR	0,834	0,849	0,776	0,827	0,453	0,824
(Křížové) HČ (q = 3)	0,831	0,839	0,84	0,867	0,594	0,793

Tabulka 3.4: Pearsonův korelační koeficient (lineární korelace) metrik s jejich křížovými variantami.

3.3 Křížové metriky

Radikálně odlišná je však situace u křížových variant metrik Hillova kontinua. Abych se přiznal, bylo to pro mě největší překvapení celé studie, vůbec jsem s tím nepočítal a musel jsem kvůli tomu přepočítávat značnou část výsledků.

Pokud má referenční korpus stejné kvality jako měřená sekvence, tak jsou křížové metriky z definice rovny metrikám nekřížovým. Předpokládal bych tedy, že i když je referenční korpus trochu jiný než samotná testovaná sekvence (a to v našem případě je, konec konců vybraná sekvence byla samplována z onoho korpusu), tak to na věci nic moc nezmění, a i když se bude metrika chovat trochu jinak než její nekřížová varianta, bude s ní alespoň lineárně škálovat.

To se ovšem neděje. Už z tabulek 3.1, 3.2 a 3.3 je vidět, že soutěž o nejsymetričtější distribuci s přehledem vyhrávají logaritmizované varianty a že mají (o poznání méně přesvědčivě) i nejlepší lineární korelaci. Tento trend je naplno obnažen v tabulce 3.4, která porovnává přímo křížové metriky s jejich nekřížovými variantami. Zejména u křížového počtu typů a křížové shannonovské perplexity je rozdíl markantní.

Také grafy 3.5–3.8 naznačují totéž. Už fakt, že nejnižší hodnota křížového počtu typů je vzdálena od té nejvyšší pět řádů a že se po grafu potuluje spousta bezprizorních bodů, by nás měl upozornit, že něco není v pořádku. Ony body jsem původně považoval za outliery vzešlé z nepořádku v datech, ale při bližším pohledu zjistíme, že sekvence pocházejí z úplně normálních textů a že chyba je jen ve škálování.

Je tedy zřejmé, že jako primární je třeba vnímat nikoli křížovou perplexitu, ale křížovou entropii, která je s ní škálována logaritmicky. A jelikož i ostatní metriky křížového Hillova kontinua vycházejí z křížových Rényiho entropií, není nic jednoduššího, než se k těmto entropiím vrátit a pracovat právě s nimi. Ve všech dalších kapitolách tedy budu používat právě logaritmizované verze křížového počtu typů (tedy Rényiho entropie s parametrem $q = 0$), křížovou Shannonovu entropii ($q \to 1$), logaritmované RRR ($q = 2$) atd.

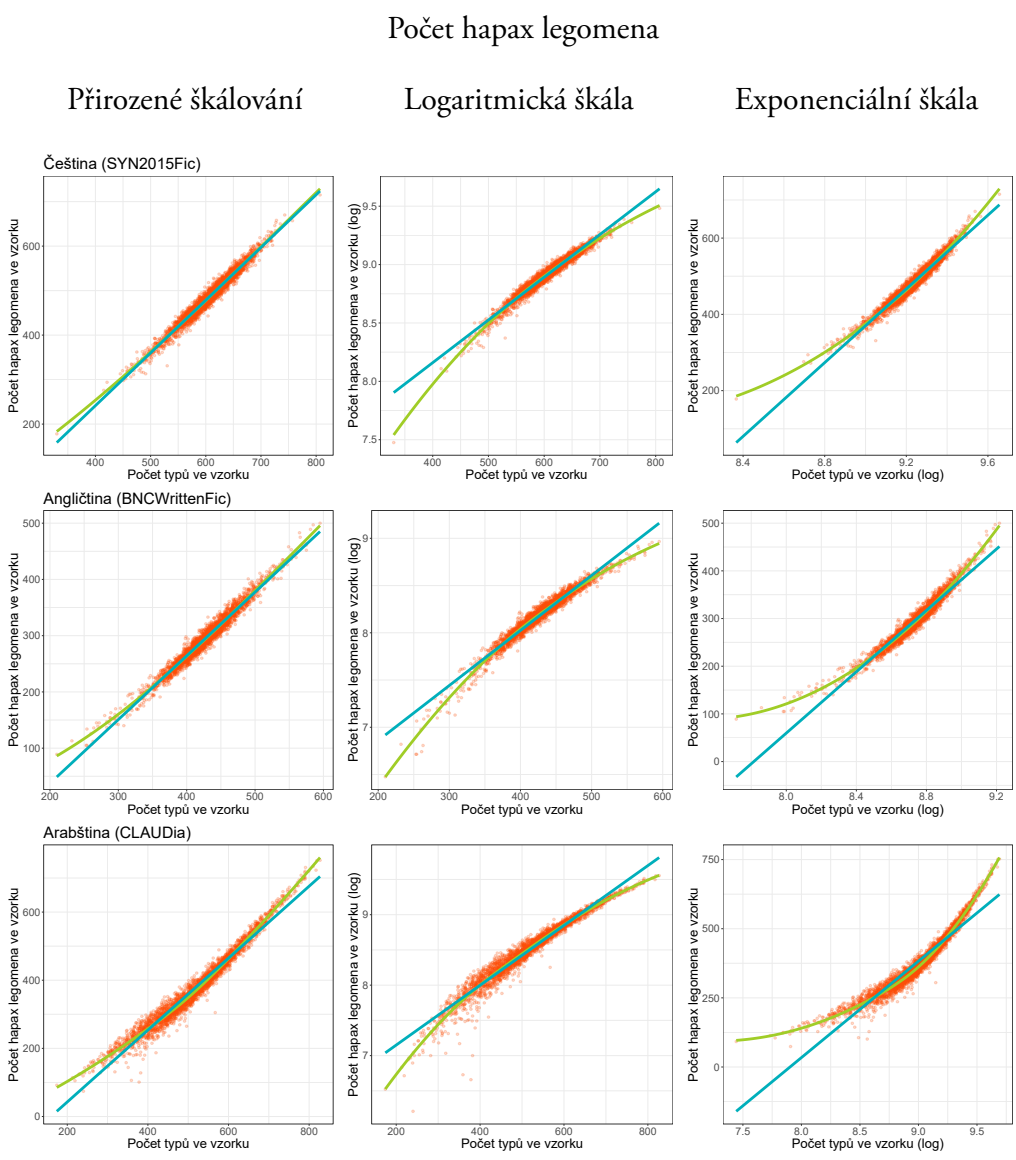
3.4 Ostatní metriky

Od průměrné délky slov očekávám, že bude dobře lineárně škálovat s křížovou entropií, z teoretických důvodů vyjádřených v podkapitole 1.7, což, jak vidíme na obrázku 3.12, empirická data podporují (a konec konců podobný výsledek jsme viděli už na obrázcích 1.8–1.9, které představují data naměřená na heterogennějším korpusu). Vzhledem k tomu, že křížová entropie (tedy log. křížové perplexity) zase lineárně škáluje s nekřížovou perplexitou, nebudeme se příliš vzrušovat tím, že logaritmus průměrné délky má symetričtější rozdělení a mírně líp lineárně koreluje s počtem typů než samotná průměrná délka.

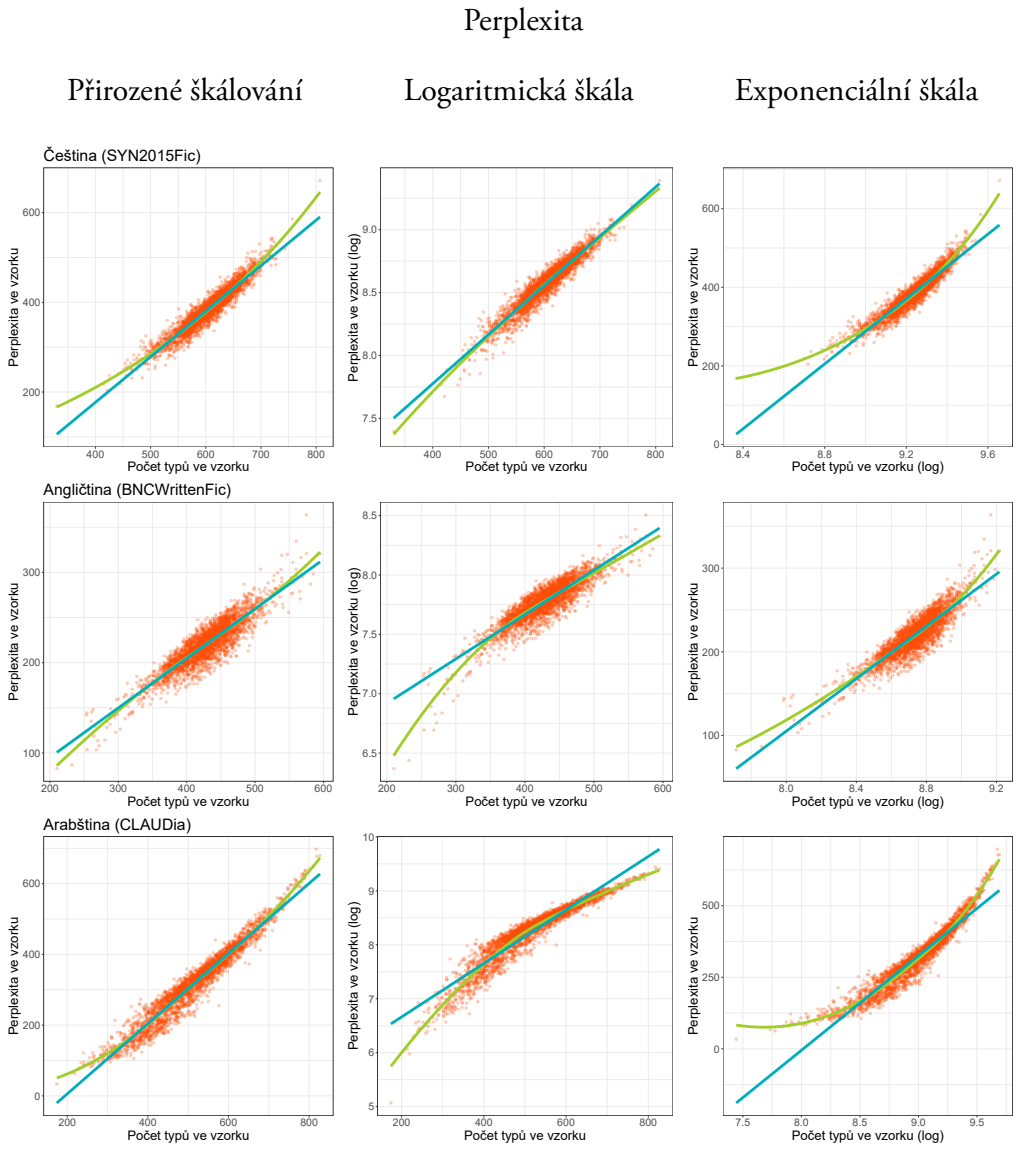
Pokud něco průzkum v této kapitole odkryl o rozdílnosti, je to fakt, že s ostatními

metrikami lexikální diverzity prakticky nekoreluje. Pokud jsem tedy při jejím představování sliboval, že vám může pomoci změřit lexikální diverzitu všude tam, kde si nejste jistí metodou typizace, byla to klamavá reklama, lexikální diverzita evidentně měří něco úplně jiného, k čemuž se ještě vrátíme.

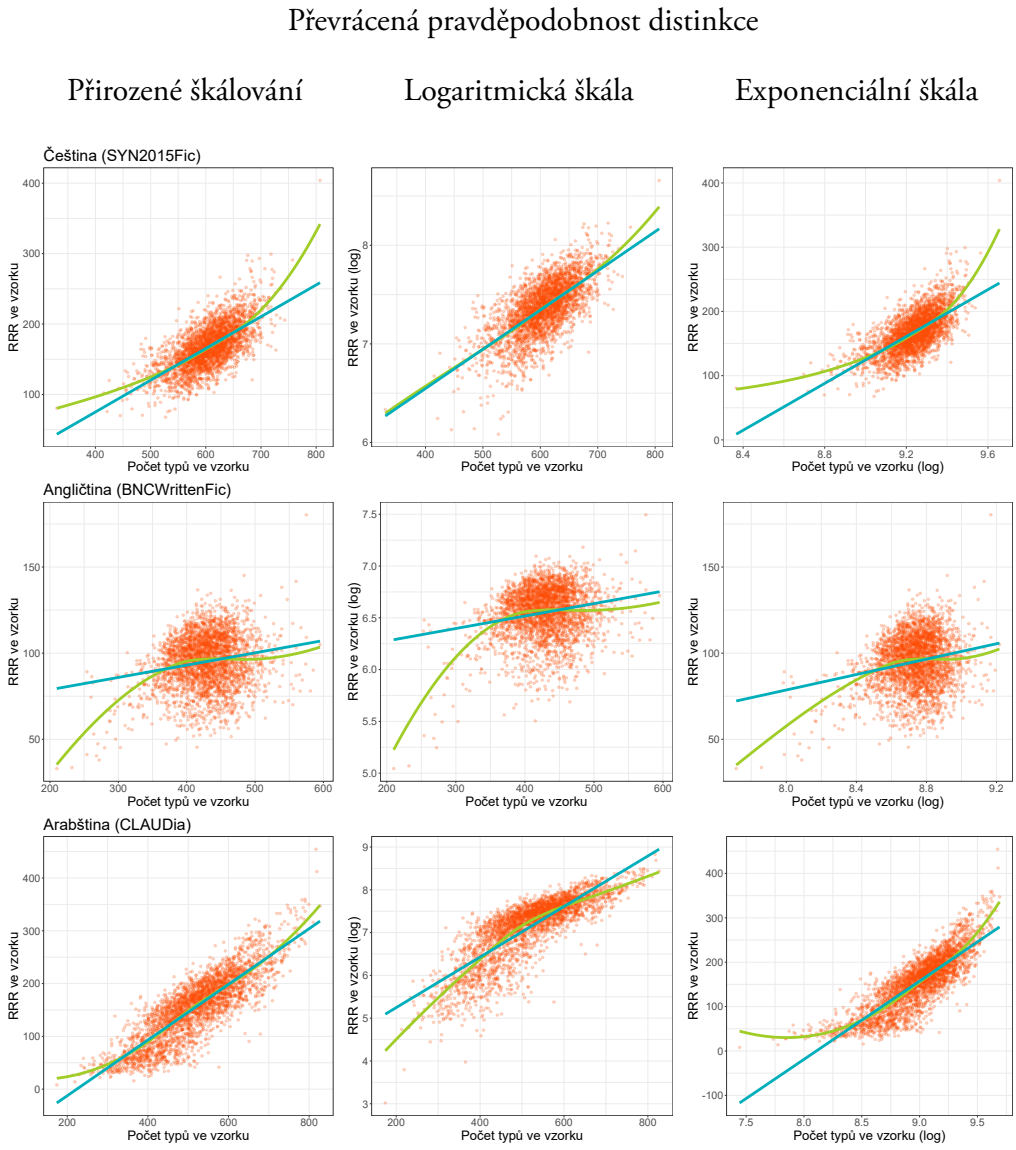
Výsledky u podílu autosémantik nejsou jednoznačné, vzhledem k malé korelaci s ostatními metrikami bychom se mohli přiklonit ke všem třem variantám, ale jelikož není důvod předpokládat jinou než lineární závislost, žádné transformace podstupovat nebudeme.



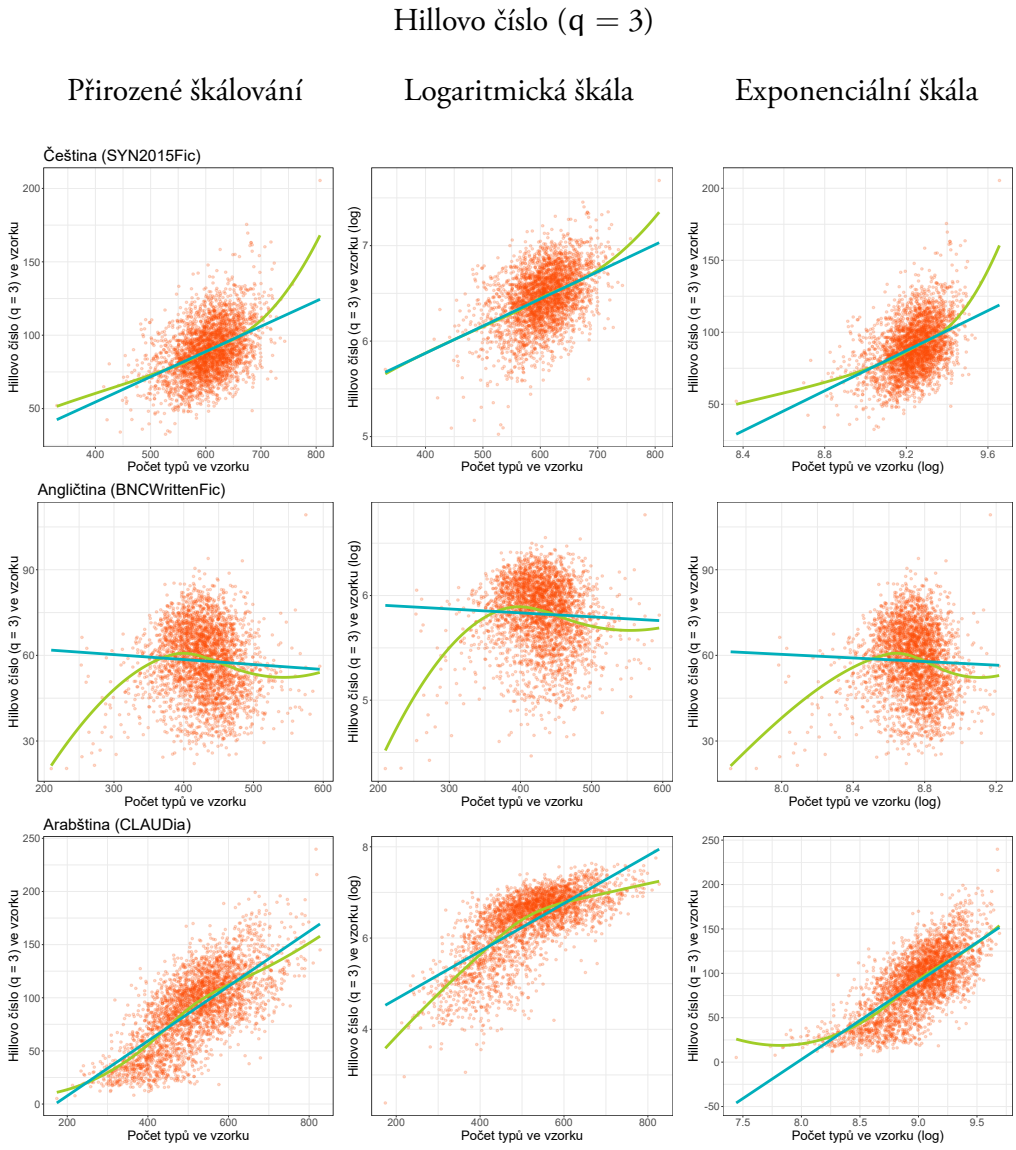
Obrázek 3.1: Korelace počtu typů s počtem hapax legomena.



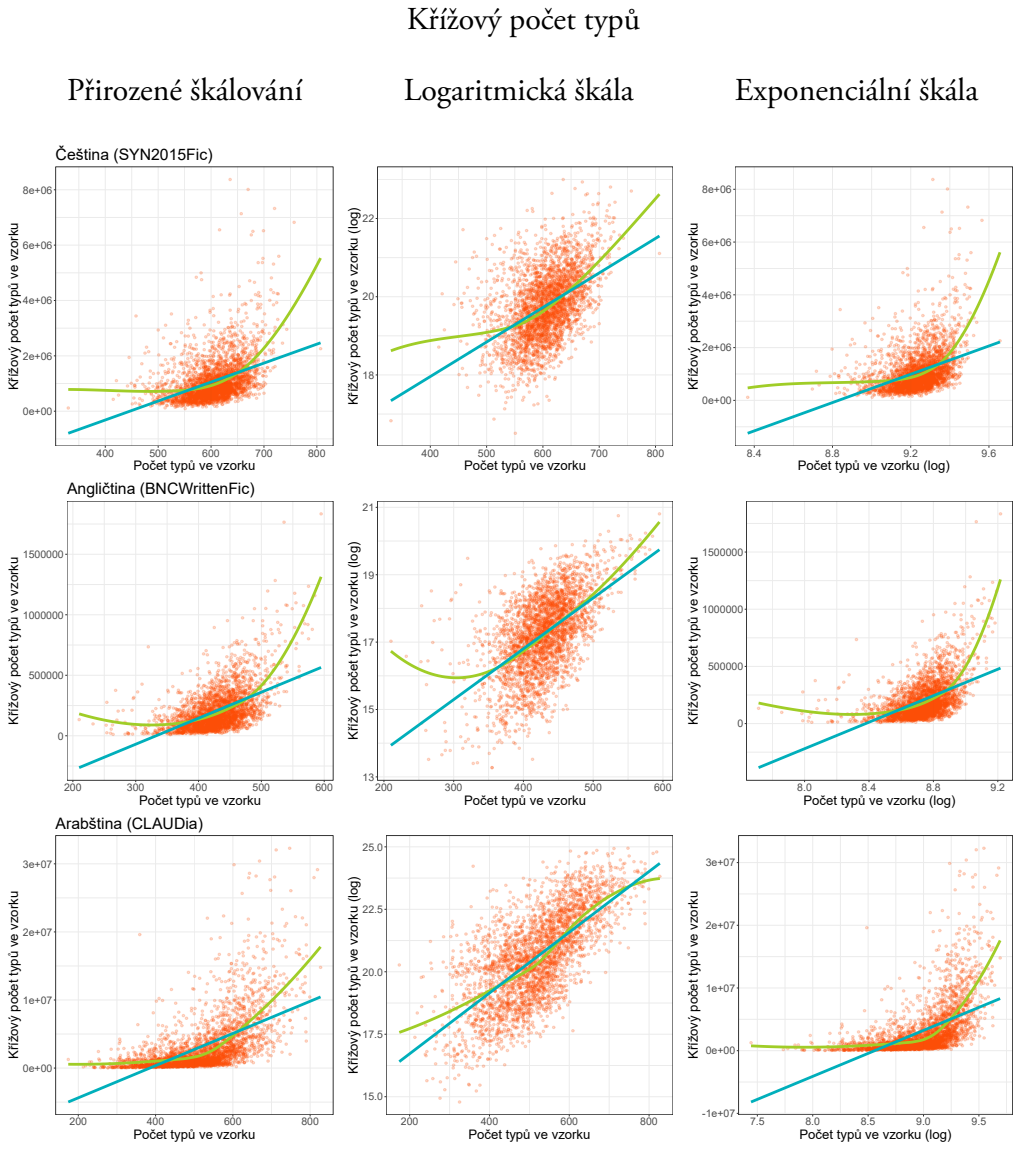
Obrázek 3.2: Korelace počtu typů s perplexitou.



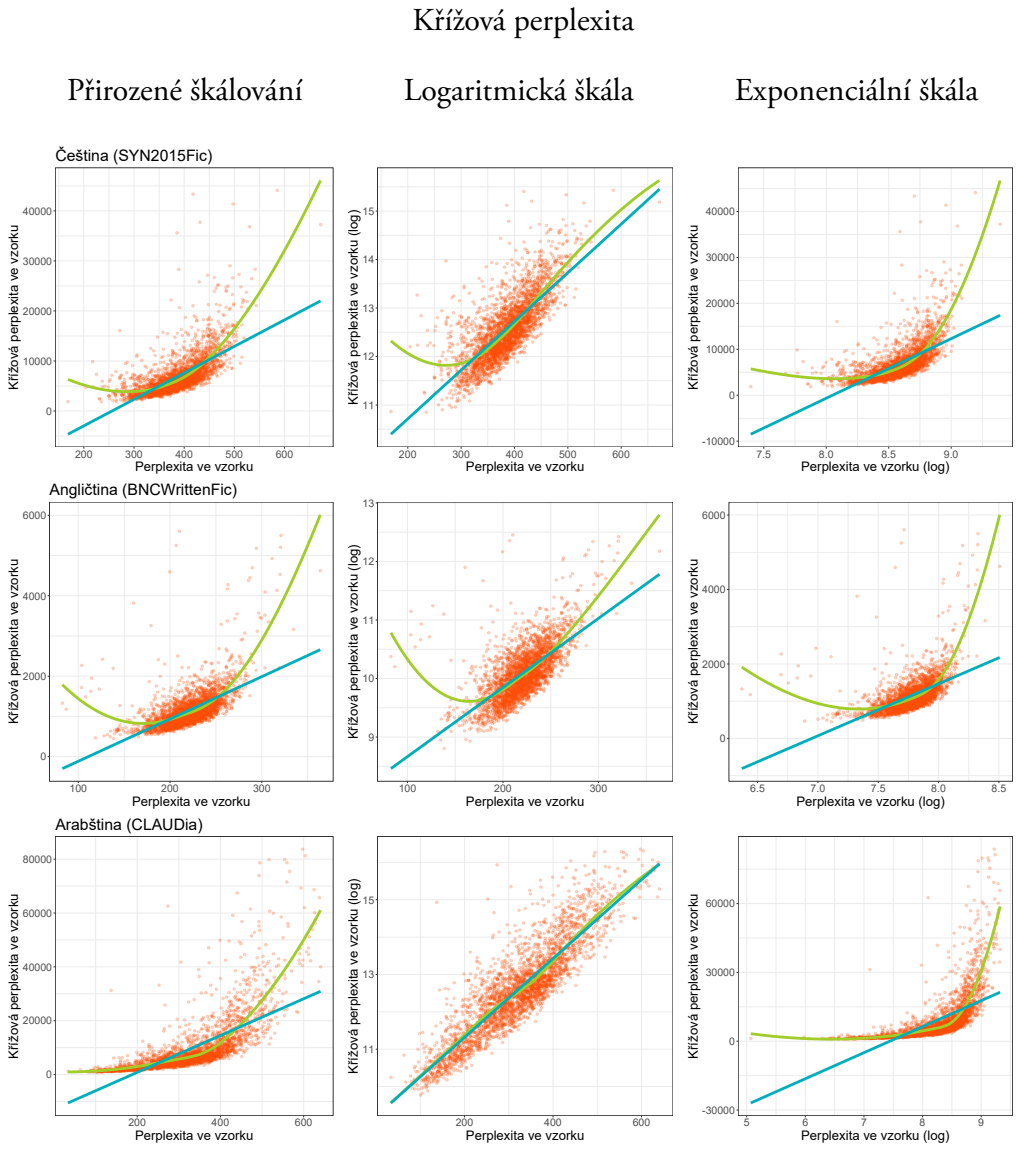
Obrázek 3.3: Korelace počtu typů s převrácenou pravděpodobností distinkce (RRR).



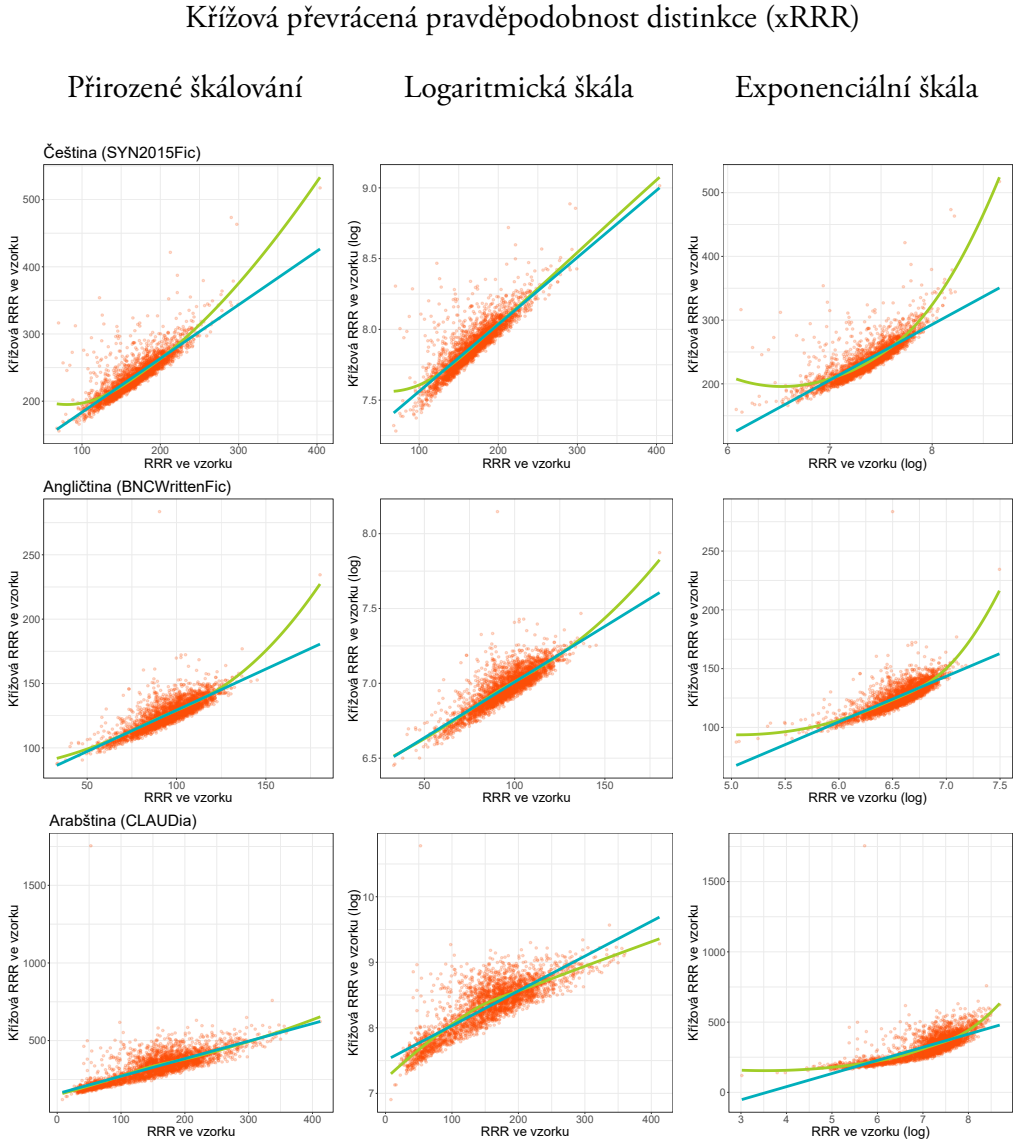
Obrázek 3.4: Korelace počtu typů s Hillovým číslem (q = 3).



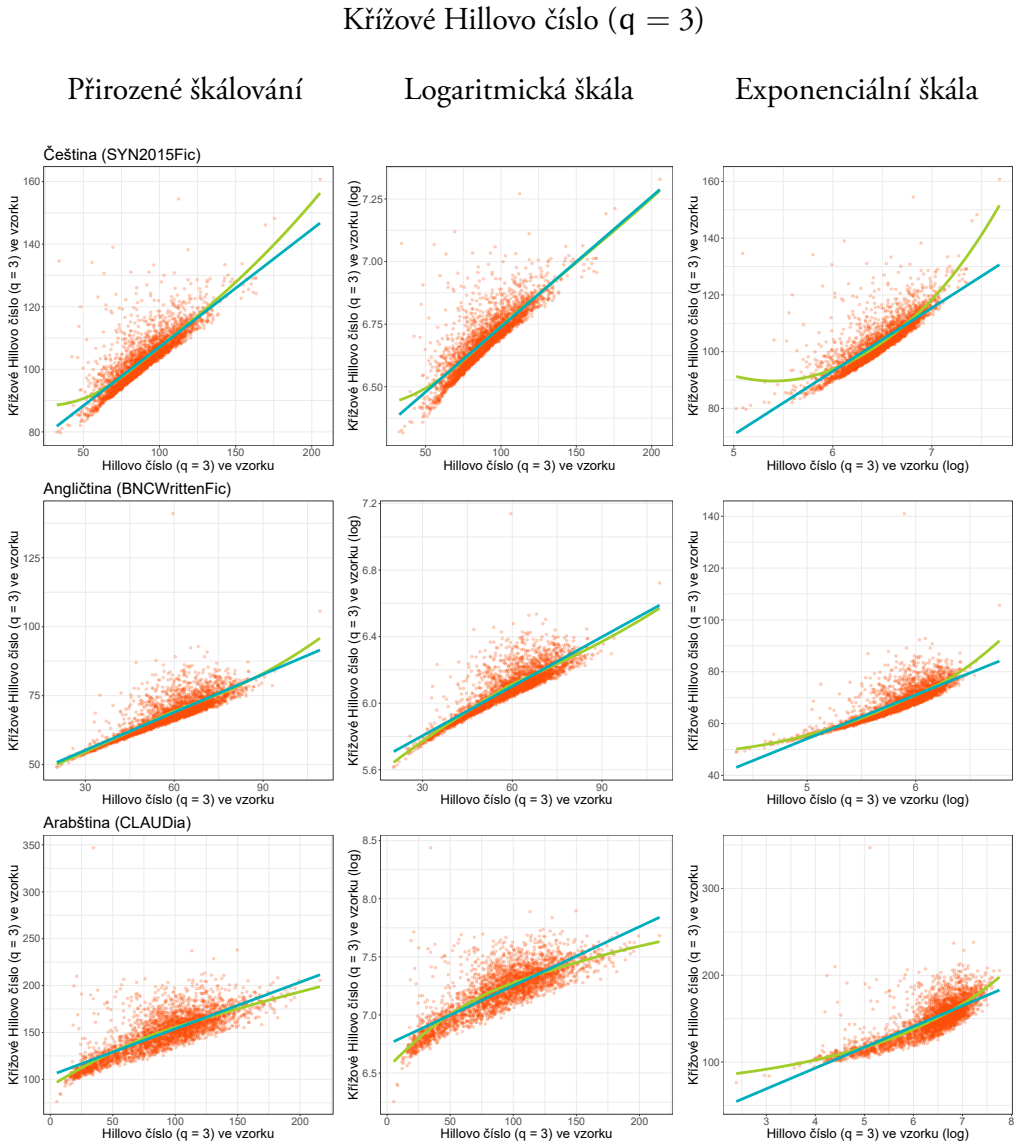
Obrázek 3.5: Korelace počtu typů s křížovým počtem typů.



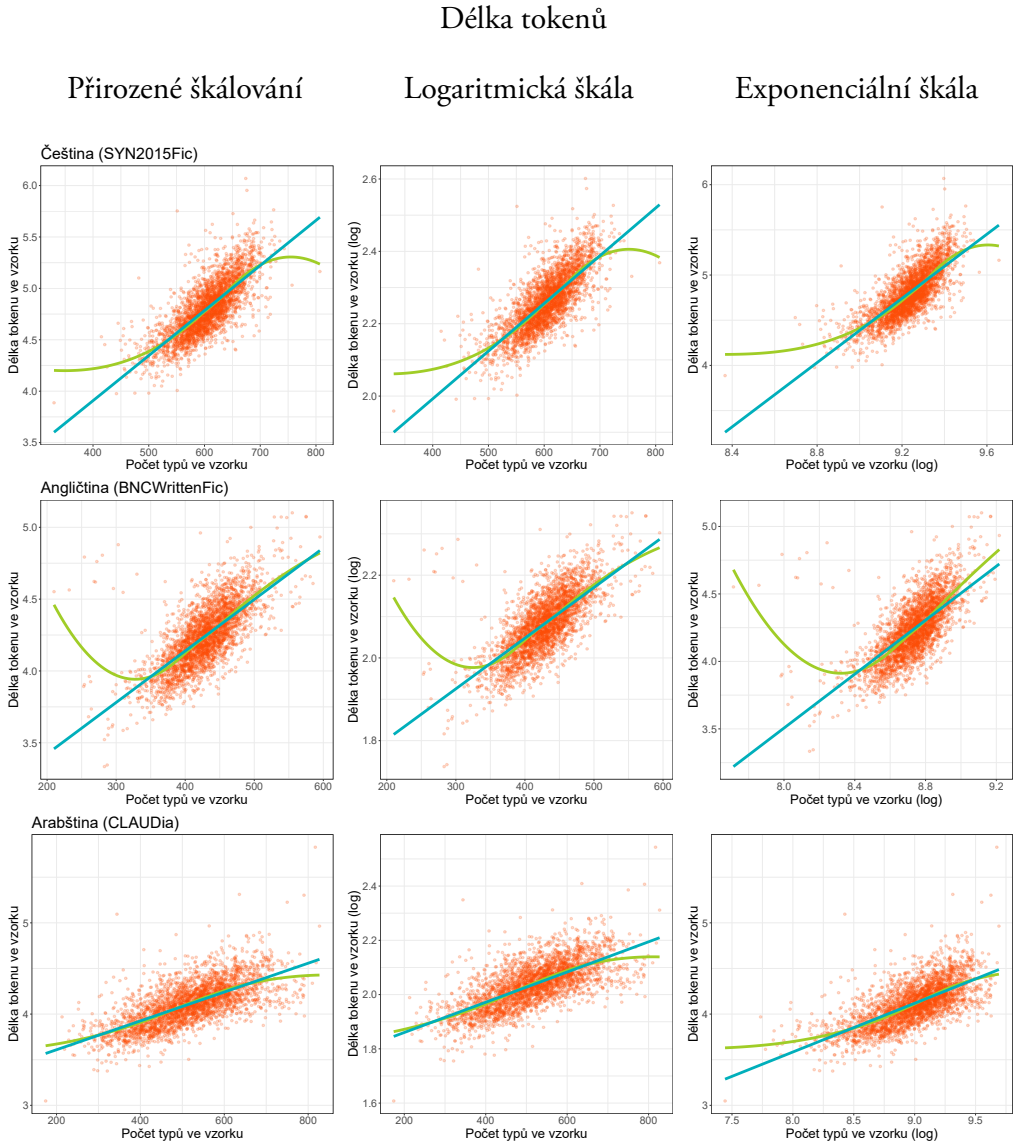
Obrázek 3.6: Korelace perplexity s křížovou perplexitou.



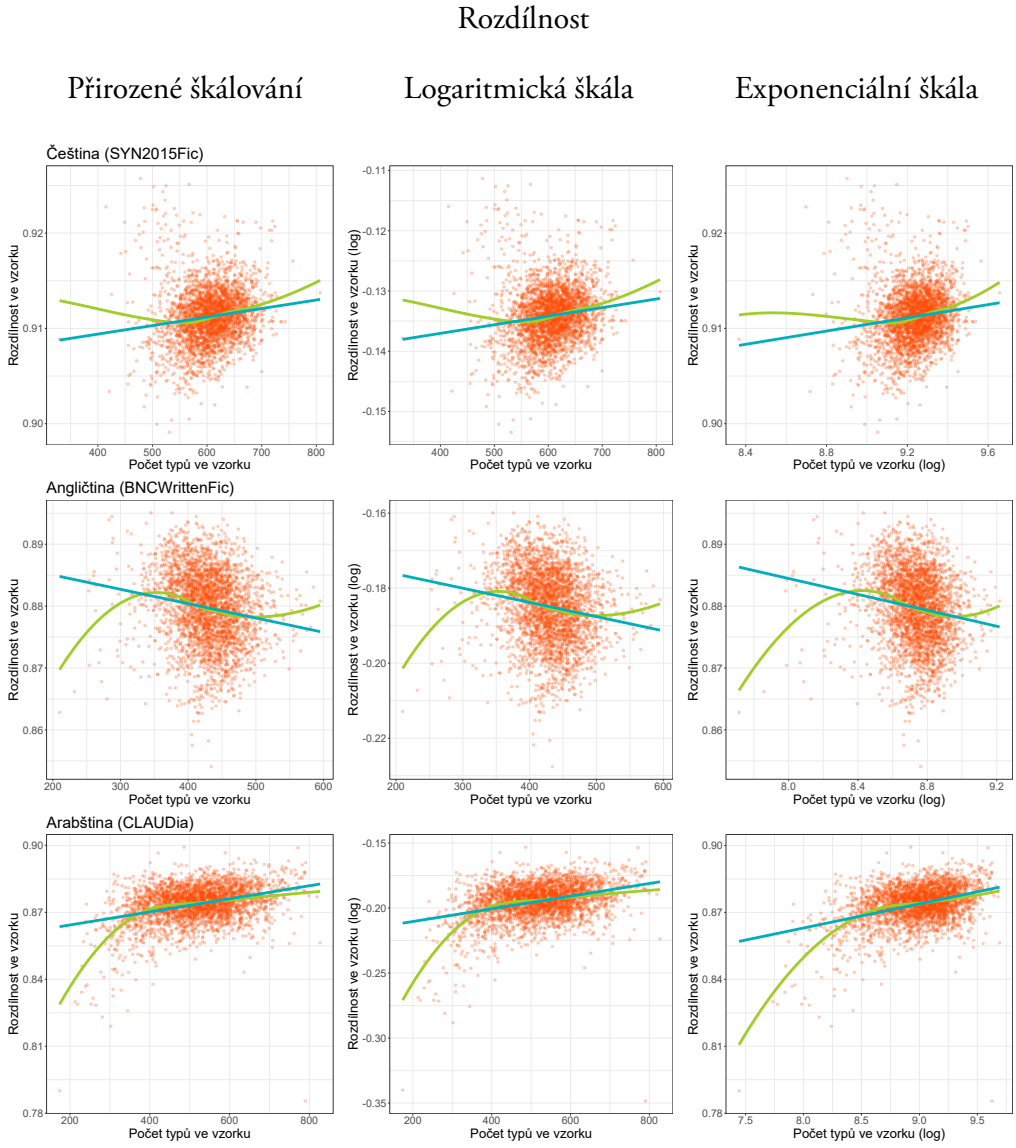
Obrázek 3.7: Korelace převrácené pravděpodobnosti distinkce (RRR) s křížovou převrácenou pravděpodobností distinkce (xRRR).



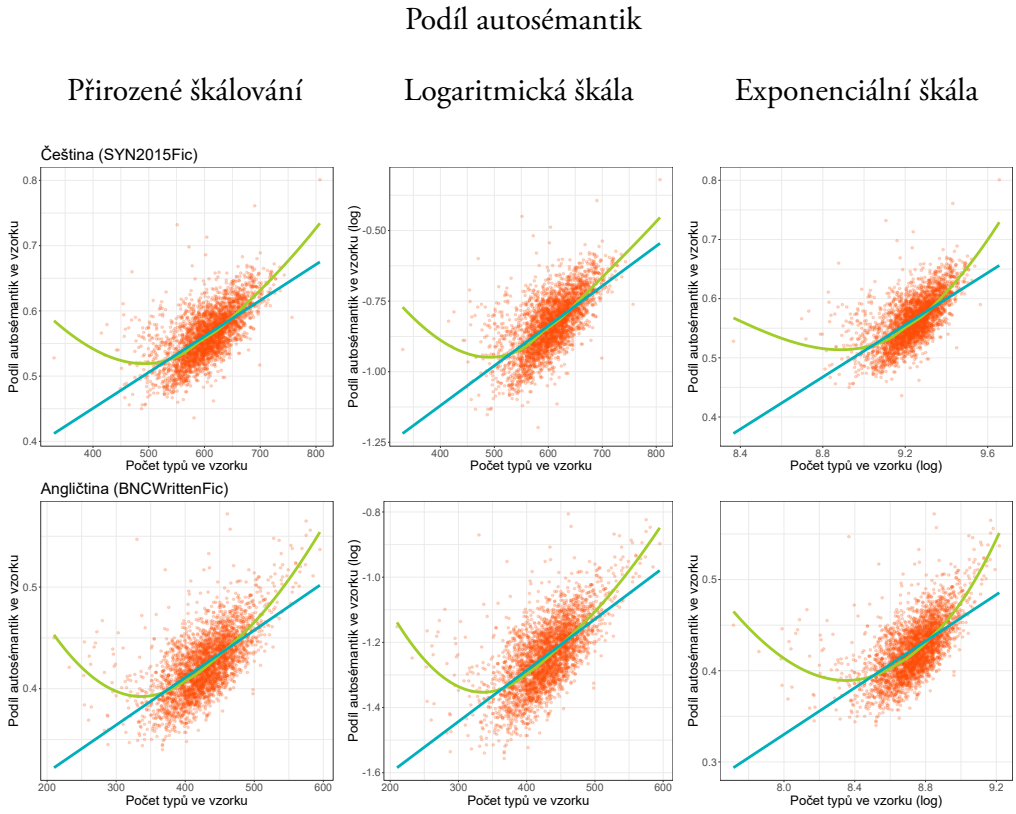
Obrázek 3.8: Korelace Hillova čísla s křížovým Hillovým číslem (v obou případech q = 3).



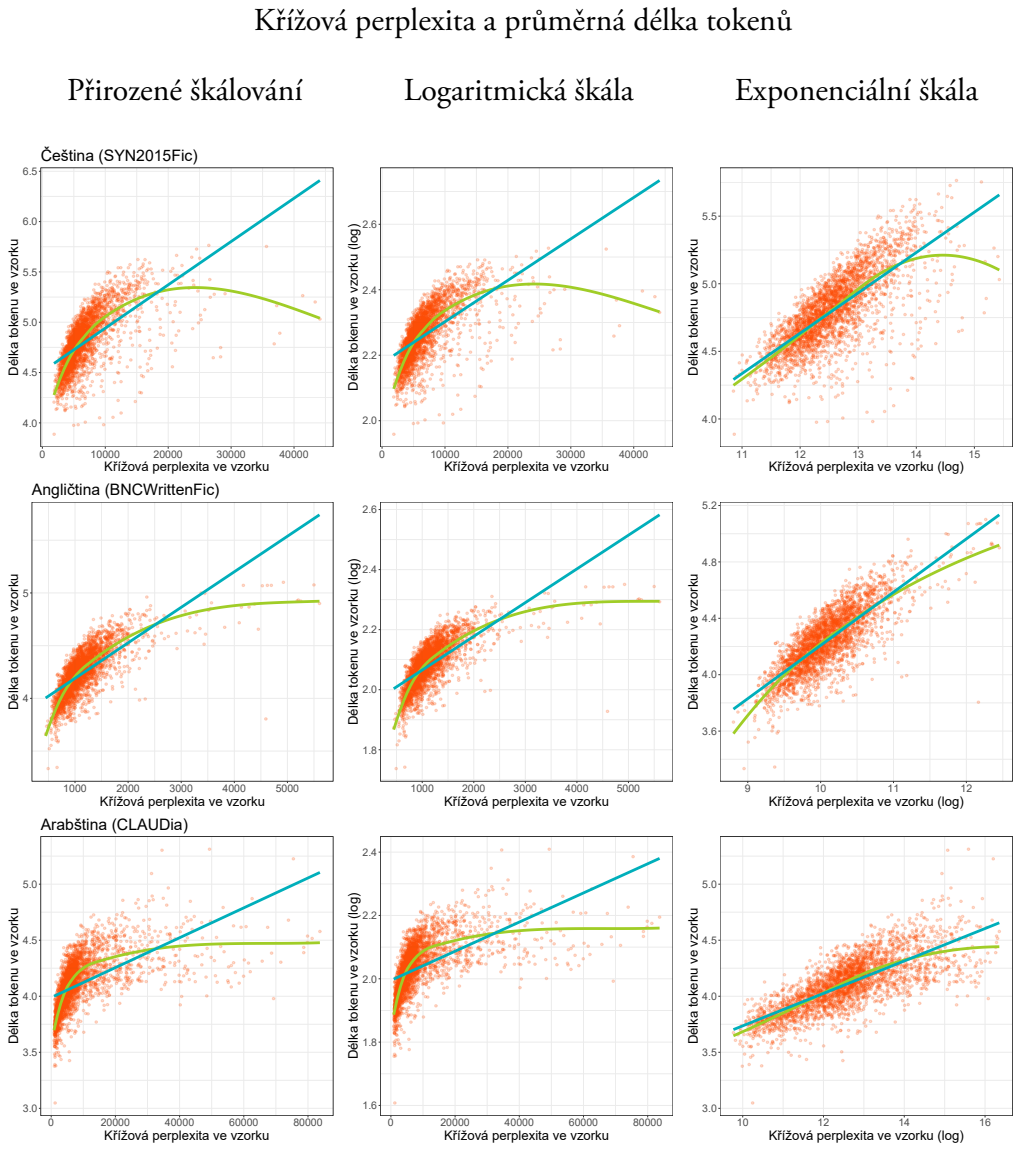
Obrázek 3.9: Korelace počtu typů s délkou tokenů.



Obrázek 3.10: Korelace počtu typů s rozdílností.



Obrázek 3.11: Korelace počtu typů s podílem autosémantik.



Obrázek 3.12: Korelace křížové perplexity s průměrnou délkou tokenů.

Kapitola 4

Vliv lemmatizace

Jak jsem už nakousl v kapitole 1.1, můžeme si vymyslet obrovské množství způsobů, jakými kategorizovat jednotlivá slova do typů, typizovat je. Funkce, která nám říká, jestli jsou dvě slova stejná, nemusí sice nutně dávat binární výsledky, může to být nějaké skóre podobnosti, jak jsem zmínil v kapitole 1.8, ovšem prostá binární typizace je jednoduše vysvětlitelná a prakticky celá lingvistická tradice s ní v nějaké formě pracuje.

Na jedné straně stojí striktní typizace podle slovních forem, kdy distinkci mezi typy zakládá libovolná distinkce v grafické podobě slova (v případě mluveného korpusu pak obvykle podle principů psaného jazyka), na druhé pak více či méně velkorysá lemmatizace, kdy distinkce mezi typy je uznána pouze tam, kde se typy liší více než v pouhé koncovce, popřípadě se může uplatnit nějaký složitější sémantický princip. Zjednodušeně: pokud se dvě slova liší alespoň jedním písmenem, pak patří k rozdílnému slovnímu typu, pokud byste dvě slova čekali pod jiným heslem ve slovníku, pak patří k jinému lemmatu.

Typizace podle slovních forem a typizace pomocí lemmatizace jsou klasikou v hlavním proudu korpusové lingvistiky, a byť nejsou úplně krajními body na škále, jsou od sebe docela vzdáleny, obzvláště v morfologicky bohatších jazycích. Proto má smysl se na ně podívat jako na vhodné prototypické představitele typizace a prozkoumat, jaký vliv mají na výsledky jednotlivých metrik.

Předesílám, že není možné apriori říct, že některá metoda typizace je lepší než jiná — a nějakých jednoznačných výsledků jsme se zatím nedobrali ani empiricky: Jarvis – Hazhangmoto (2021) zkoušeli, jak různé metriky měřené na textech různě typizovaných korelují se subjektivními hodnoceními lexikální diverzity, ovšem výsledky jejich experimentů jsou nejednoznačné, spíš matoucí. Osobně chápu způsob typizace jako jednu z proměnných, která nám může pomoci zvětšit diverzitu v metodách měření lexikální diverzity, a podobně jako nemůžeme, a vlastně ani nechceme říct, která konkrétní metrika je obecně nejlepší, neboť každá přináší jiný vhled, není možné ani nutné zaujmout nějaké všeobecně platné hodnotící stanovisko ke způsobu typizace.

Stejně jako v případě vlivu délky textu, i tady je jasné, že typizace bude mít nějaký

vliv na výsledky dané metriky, ovšem i zde vyvstává otázka, jak velký vliv to je. Není možné jej, alespoň za určitých okolností, zanedbat? Jsou všechny metriky stejně citlivé na metodu typizace, nebo se v tom nějak liší?

4.1 Korelace mezi výsledky pro lemmatizovaný a nelemmatizovaný text

Onu citlivost nejlépe vidíme na tom, jak spolu korelují výsledky jednotlivých metrik změřené nejprve na náhodně vybrané nelemmatizované sekvenci a následně na stejné sekvenci, ovšem lemmatizované. Na obrázku 4.1 nahoře vidíme 3000 takových sekvencí o délce 1000 tokenů, každá je reprezentována jedním datovým bodem. Jak je vidět, lineární korelace je vysoká, přičemž pro anglický text je vyšší než pro český, což je vzhledem k typologii obou jazyků zcela očekávatelné (Pearsonův korelační koeficient je pro anglické vzorky 0,99, což je opravdu vysoká hodnota). Když jsem graf poprvé uviděl, byl jsem v pokušení napsat, že lexikální diverzita je vůči metodě typizace překvapivě robustní, ale výsledky pro další metriky mě vyvedly z omylu. Ještě počet hapaxů (obr. 4.3) a perplexita (obr. 4.4) si zachovávají poměrně vysokou korelaci, ovšem jak stoupáme po Hillově kontinuu výše, korelace čím dál více slábne (obr. 4.6), až klesne k 0,72 pro Hillovo číslo s $q = 3$ (obr. 4.7).

Explanaci tohoto jevu můžeme dedukovat z vlastností Hillova kontinua, které tak konečně využijeme i jinak než jako pouhou systematizaci zdánlivě nesouvisejících metrik. Metriky s vyšším parametrem q jsou více závislé na typech s vyšší frekvencí, což implikuje, že lemmatizace má větší příležitost se projevit. Pokud je nějaké lemma hapaxem, pak je v nelemmatizovaném textu též hapaxem. Ovšem lemmata s vysokou frekvencí mohou být v nelemmatizovaném textu reprezentována různým počtem typů.

Naopak výsledky pro Rényiho entropie, tedy křížové varianty Hillových čísel, které jsem v souladu se závěry předchozí kapitoly zlogaritmoval, jsou poněkud chaotické. Pořád sice platí, že vzorky anglicky psaných textů mají mnohem větší korelaci než vzorky textů českých, ovšem vliv stoupajícího parametru q není tak jednoznačný: nejlépe koreluje křížová entropie (4.9), která má vyšší korelační koeficient než logaritmovaný křížový počet typů a také vyšší korelaci než čistá perplexita, nicméně dále se stoupajícím parametrem q korelace rychle klesá (viz $q = 2$ na obrázku 4.10) a spadne až k padesáti procentům pro křížovou Rényiho entropii s parametrem $q = 3$ (4.11). Tedy platí to tak pro češtinu, v angličtině dál zůstává korelační koeficient vysoký, a s rostoucím koeficientem tak nejspíš roste vliv morfologického bohatství jazyka.

Nemá sice smysl měřit délku tokenů na lemmatizovaném textu (lemma je arbitrárně zvolený „slovníkový“ tvar), ovšem pokud by někdo takovou chybu udělal, na výsledek by to nejspíš nemělo moc vliv (obrázek 4.12). Naopak rozdílnost (dissimilarity), kterou také nedává smysl měřit na lemmatizovaném textu, neboť „základní

tvary“, které lemma reprezentují, jsou vybrané arbitrárně, na lemmatizaci záleží poměrně dost a hodnoty korelace jsou podobně vysoké jako u RRR.

4.2 Kolik chyb můžeme čekat, když místo lemmatizovaného textu použijeme nelemmatizovaný

Předpokládám, že běžný čtenář není úplně schopný si pod hodnotou Pearsonovy korelace představit nic konkrétního.[1] A právě pro něj nabízím ještě další způsob vizualizace, jak moc se liší daná metrika na lemmatizovaných a nelemmatizovaných textech — chybovost, jak ji známe z předchozí kapitoly. Tedy, pokud bychom dvě sekvence seřadili podle jejich lexikální diverzity, pak sekvence lemmatizovali, lexikální diverzitu změřili ještě jednou a znovu je seřadili, jak moc by se ona dvě řazení lišila.

Spodní grafy na obrázcích 4.1–4.13 ukazují, co se stane, když z korpusu náhodně vybereme dvě sekvence určité délky,[2] spočítáme rozdíl v lexikální diverzitě mezi těmito dvěma sekvencemi a následně je umístíme do prostoru stejně jako v předchozím případě — na ose x je rozdíl mezi sekvencemi v nelemmatizované formě, na ose y pak v lemmatizované. Červeně obarvené datové body pak reprezentují ty dvojice sekvencí, kde rozdíl diverzity nelemmatizované sekvence má jiné znaménko než rozdíl diverzity sekvence lemmatizované.

Jinými slovy ony dvě sekvence seřadíme podle hodnoty zvoleného indexu lexikální diverzity, a to jak podle lemmatizovaného, tak podle nelemmatizovaného textu. Pokud se ono pořadí liší, pak vzorek označíme jako chybný a obarvíme ho červeně.[3]

Opět se ukazuje, že rozdíl v typizaci zasahuje zejména flektivní češtinu, podíl chyb je v různých metrikách zhruba dvojnásobný oproti angličtině. Chybovost pak dosahuje až 20–30 procent, zejména pro křížové metriky s vyšším parametrem q. Ovšem na grafech vidíme, že chyby se koncentrují zejména tam, kde rozdíl v diverzitě mezi sekvencemi není příliš velký. Tedy že pokud bychom se soustředili na dvojice sekvencí, které se v dané metrice citelně liší, tak by byla chybovost mnohem menší. Záleží na vás, jak velký rozdíl považujete za „dost velký“, ovšem tak jako tak je reálný počet chyb, na kterých opravdu záleží, menší než počet chyb popsaný v grafech.

Můžeme se přímo podívat, jak chybovost závisí na velikosti rozdílu, a to na obrázcích 4.2 pro počet typů a 4.5 pro perplexitu. Na horních grafech ještě lépe vidíme, že

[1] To není žádná ostuda. Korelační koeficient se chová dost neintuitivně, rozdíl mezi hodnotou 0 a 0,5 je mnohem menší než mezi 0,5 a 1 (Taleb, 2019), pro lidské oko je mnohem příznivější podívat se na bodový graf, což je taky důvod, proč jich je v této knize tolik. Intuici je možné si vycvičit, ale i tak možná oceníte alternativní reprezentaci, kterou kromě korelace nabídnu.

[2] Proces výběru náhodných vzorků je přesně popsán v příloze C.4.

[3] Ve skutečnosti se o chyby v pravém slova smyslu nejedná, nemůžeme říct, že jedno nebo druhé řazení je špatně, pojem *chyba* používám spíš z terminologické nouze, možná lépe by bylo říkat *odlišnost* nebo *rozdíl*, ale to zní podivně.

chyby (červená výseč distribuce) se koncentrují ve vzorcích, které se od sebe v lexikální diverzitě příliš neliší. Pokud bychom brali v potaz pouze vzorky, kde se sekvence o tisíci tokenech liší o víc než padesát typů, byla by chybovost prakticky zanedbatelná.[4] Chybovost změřená na sekvencích různých délek se různí, jak si můžeme prohlédnout na obrázcích 4.14 pro češtinu a 4.15 pro angličtinu. Sekvence o délce tisíce tokenů jsou vlastně pořád ještě v nestabilním pásmu a teprve na delších sekvencích se metriky nějak stabilně seřadí podle úspěšnosti, nicméně právě v tomto pásmu se nacházejí délky sekvencí, které nás obvykle zajímají. O těchto grafech bude ještě řeč dále.

4.3 Klastrování pomocí lexikální diverzity lemmatizovaného a nelemmatizovaného textu

Největší překvapení této kapitoly nás ovšem čeká na grafech v prostřední řadě již představených obrázků, kde vidíme stejné datové body, tentokrát ovšem obarvené podle jejich příslušnosti k různým druhům textu. Původně jsem tato data generoval čistě proto, abych se podíval, jestli ona nízká korelace u některých indexů není dána prostě heterogenitou korpusu. Ovšem rozhodně jsem nečekal, že lexikální diverzita změřená na lemmatizovaném a nelemmatizovaném anglickém korpusu, tedy pouhé dvě dimenze, dokáže tak dobře klastrovat texty podle jejich typu a modality. Umí to dokonce i metriky, které vykazují korelaci nad 90 procent, jako je prostý počet typů — ovšemže čím menší korelace, tím více prostoru pro klastrování.

Asi není překvapení, že mluvené texty se nám vyseparovaly vpravo dole, protože prostě mají nižší lexikální diverzitu, zde by stačila pouhá jedna dimenze, jedno-li, zda týkající se lemmatizovaného či nelemmatizovaného textu. Ale zajímavé je, že beletrie má svou vlastní žlutou linii pod modrou linií odborných textů. Tedy sekvence vybrané z beletrie, které mají stejný počet lemmat jako sekvence vybrané z odborných publikací, mají vyšší počet slovních typů. Modrá linie odborných textů přitom částečně zasahuje i nad šedé texty mluvené, což by naznačovalo určitý orální charakter beletrie. Obdobně se textové typy chovají i u dalších indexů, i když vzájemné pozice oněch klastrů jsou porůznu posunuty či pootočeny.

Přitom, jak dokládají distribuce na okrajích, pokud bychom se dívali na lexikální diverzitu pouze lemmatizovaného nebo pouze nelemmatizovaného textu, k žádné výrazné separaci beletrie a odborných textů nedojde, neboť ony distribuce se více méně překrývají.

Obdobný vzorec pak můžeme vidět i na češtině, kde ovšem separace není tak zřetelná a projevuje se jenom u některých metrik (zejména RRR, 4.6), nicméně obecný

[4] Mimochodem, distribuce rozdílů mezi sekvencemi docela připomíná normální rozdělení, a to jak pro lemmatizovaný, tak pro nelemmatizovaný text (viz obrázky 4.2 a 4.5, prostřední a spodní řada, normální rozdělení je nafitováno a zobrazeno čárkovanou čarou). Přitom distribuce hodnot samotných metrik příliš normálně nevypadají, což je vidět na hranách grafů na obrázcích 4.6 až 4.13.

vzorec je vizuálně velmi podobný — tedy publicistika vpravo nahoře a odborné texty nad beletrií. U křížových metrik pak odborná literatura a publicistika splývají a vydělují se od beletrie, čímž vzniká jakási přirozená distinkce mezi fiction a non-fiction.

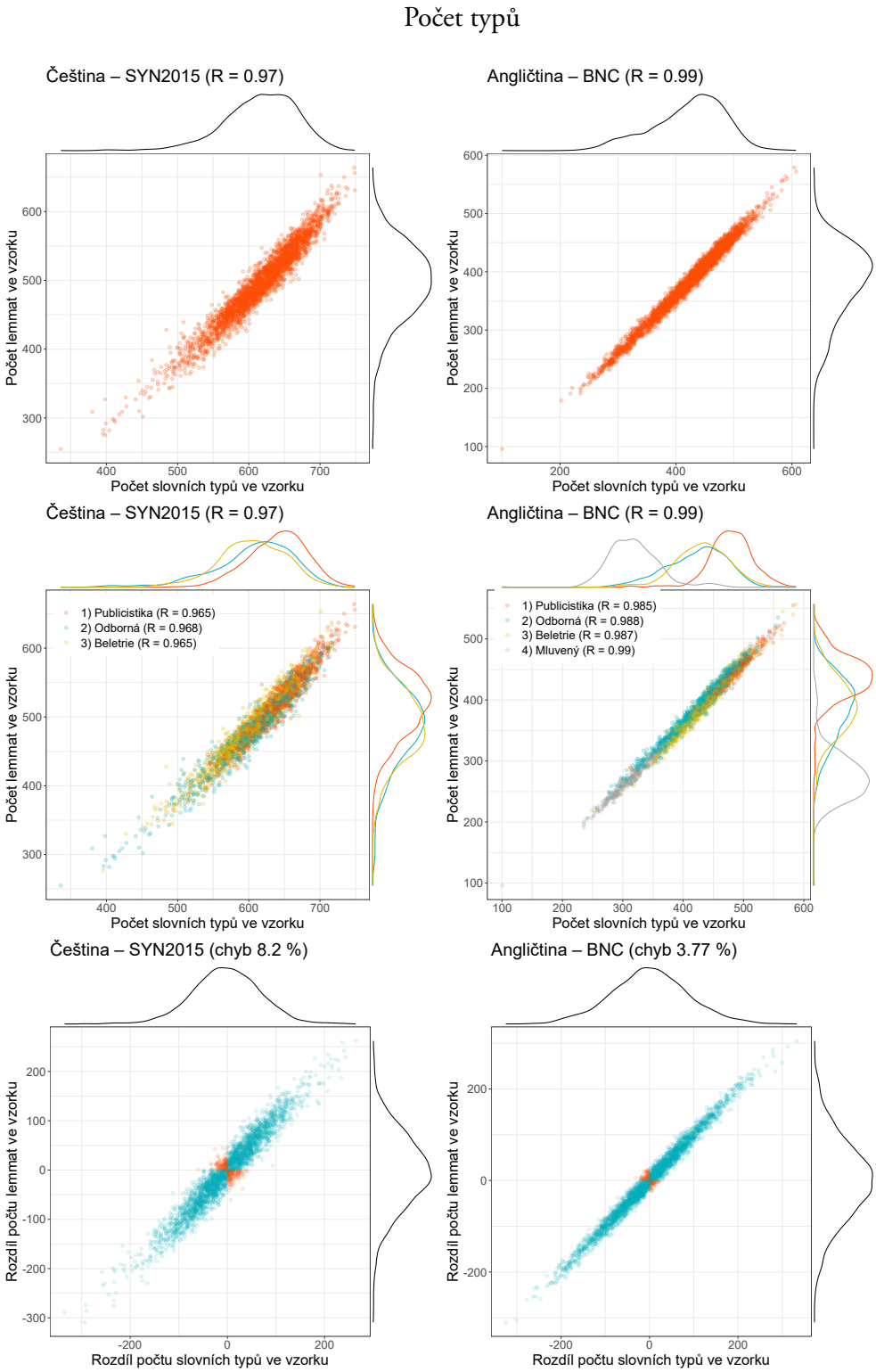
Tento výsledek je poněkud kontraintuitivní: pokud čeština kóduje v morfologii (respektive v koncovkách) více informace než angličtina, a tedy i korelace mezi lemmatizovanými a nelemmatizovanými texty je menší, očekávali bychom, že zde bude více prostoru pro clusterování textových typů. Při vysvětlení tedy budeme muset odhlédnout od obecných pravidel a uchýlit se k detailnějšímu pohledu na gramatiku obou jazyků.

V angličtině má každé lemma jen pár slovních tvarů (často doslova právě pár), s výjimkou osobních zájmen a slovesa *be*. A právě zájmena a instance slovesa *be* využívá beletrie s větší rozmanitostí než odborná literatura, která si vystačí se třetí osobou jednotného čísla. Tato gramatická bohatost také zvyšuje množství tvarů u ostatních sloves. To vysvětluje, proč texty s relativně vysokou bohatostí slovních tvarů a nižším množstvím lemmat jsou mnohem častěji beletristické než odborné.

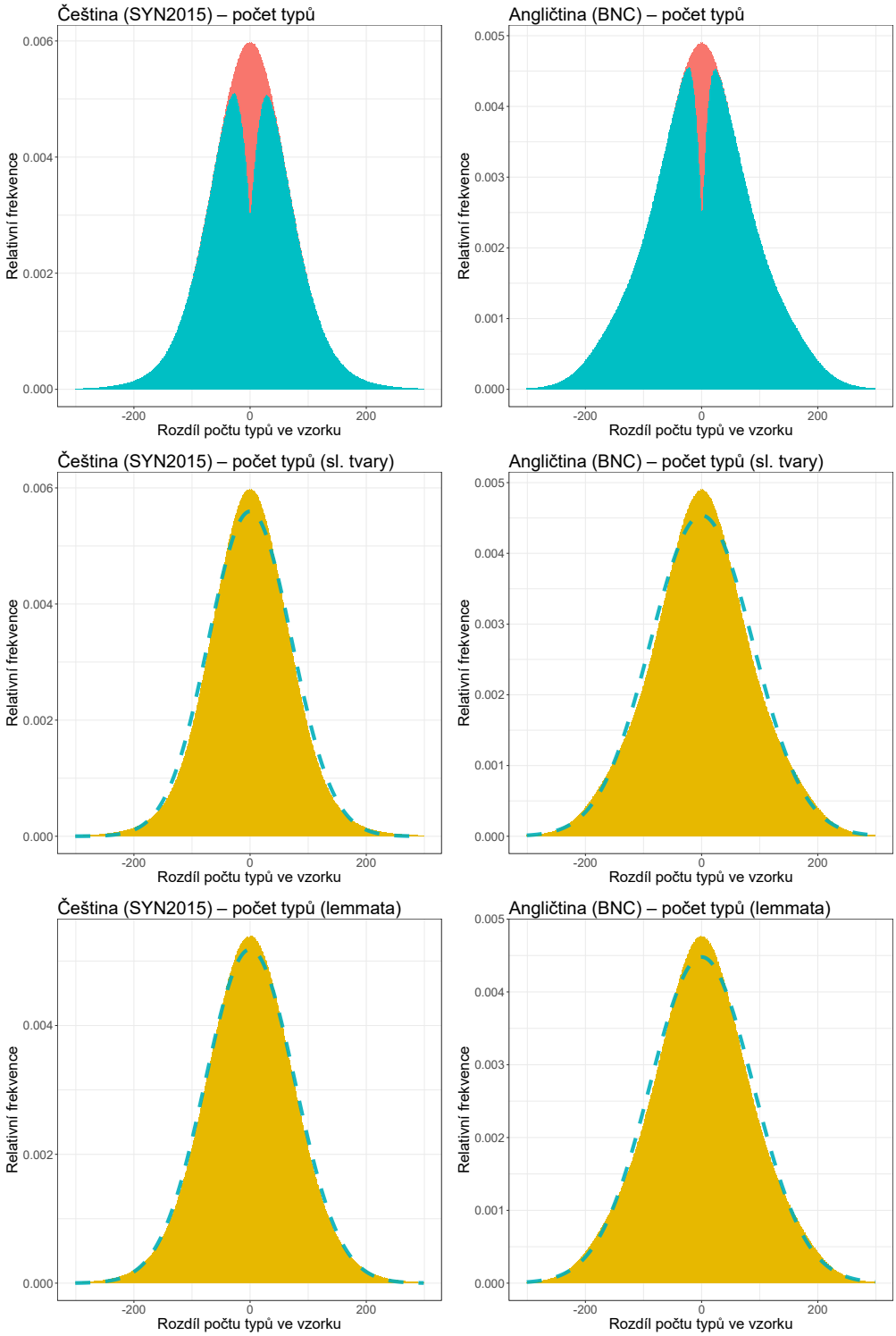
Tento princip se ovšem v češtině uplatňuje o poznání míň. Předpokládám, že je to proto, že vedle bohaté konjugace má čeština také bohatý deklinační systém, kde se vliv druhu textu projevuje míň nebo vůbec. Také se zde nejspíš ukazuje rozdílný přístup k lemmatizaci, neboť v BNC jsou všechny tvary osobních i přivlastňovacích zájmen lemmatizovány jako *I, you, he* etc., zatímco v korpusech ČNK se ponechává základní tvar *já* / *můj* / *ty* / *tvůj* atd.

Tohle je ale jen takový pokus o vysvětlení, celá záležitost by si zasloužila detailnější pohled.[5] Ostatně i uvedené grafy určitě obsahují spoustu zajímavých nebo podivných detailů, které se mi nepodařilo odhalit a které čekají na vás. Bohužel tato knížka mi nedovoluje uvést všechny grafy, které bych chtěl — přitom když změníme délku sekvence, na které metriky měříme, získáme odlišné výsledky, jak si ostatně ukážeme i v následující podkapitole.

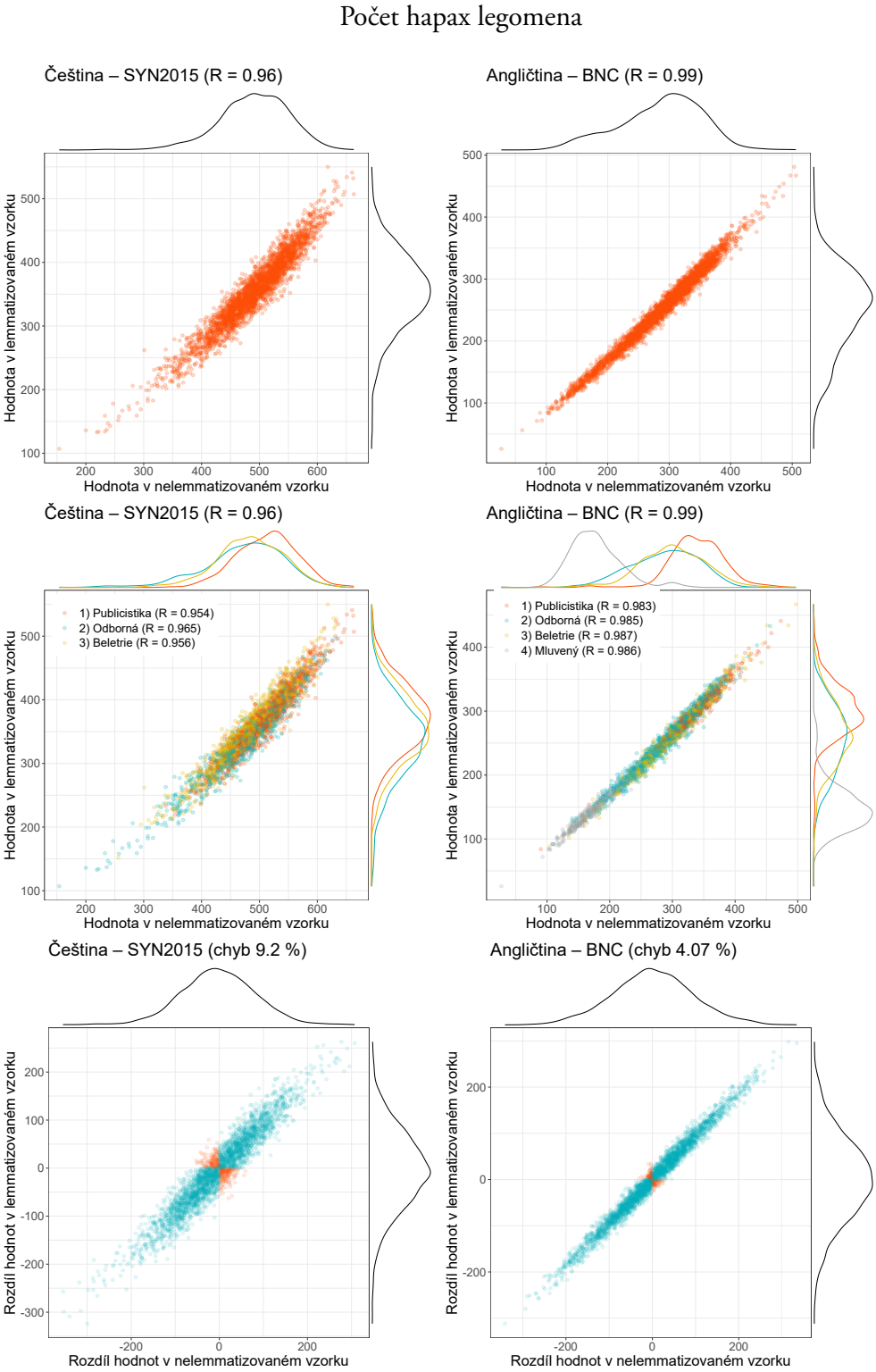
[5] Alternativní explanace, kterou mi poradil Václav Cvrček, je založena na tom, že publicistické texty obsahují mnohem více rozmanitých proprií než texty beletristické (Cvrček et al., 2020, viz rysy PROPA a PROPT). Tato vlastní jména pak jsou obvykle chybně automaticky lemmatizována, takže jednotlivé tvary proprií jsou vedeny jako lemmata, každý tvar zvlášť. To ovšem nevysvětluje rozdíl mezi češtinou a angličtinou, naopak, vlastní podstatná jména v angličtině by měla být obvykle jenom v jednom tvaru, neboť množné číslo od nich obvykle netvoříme.



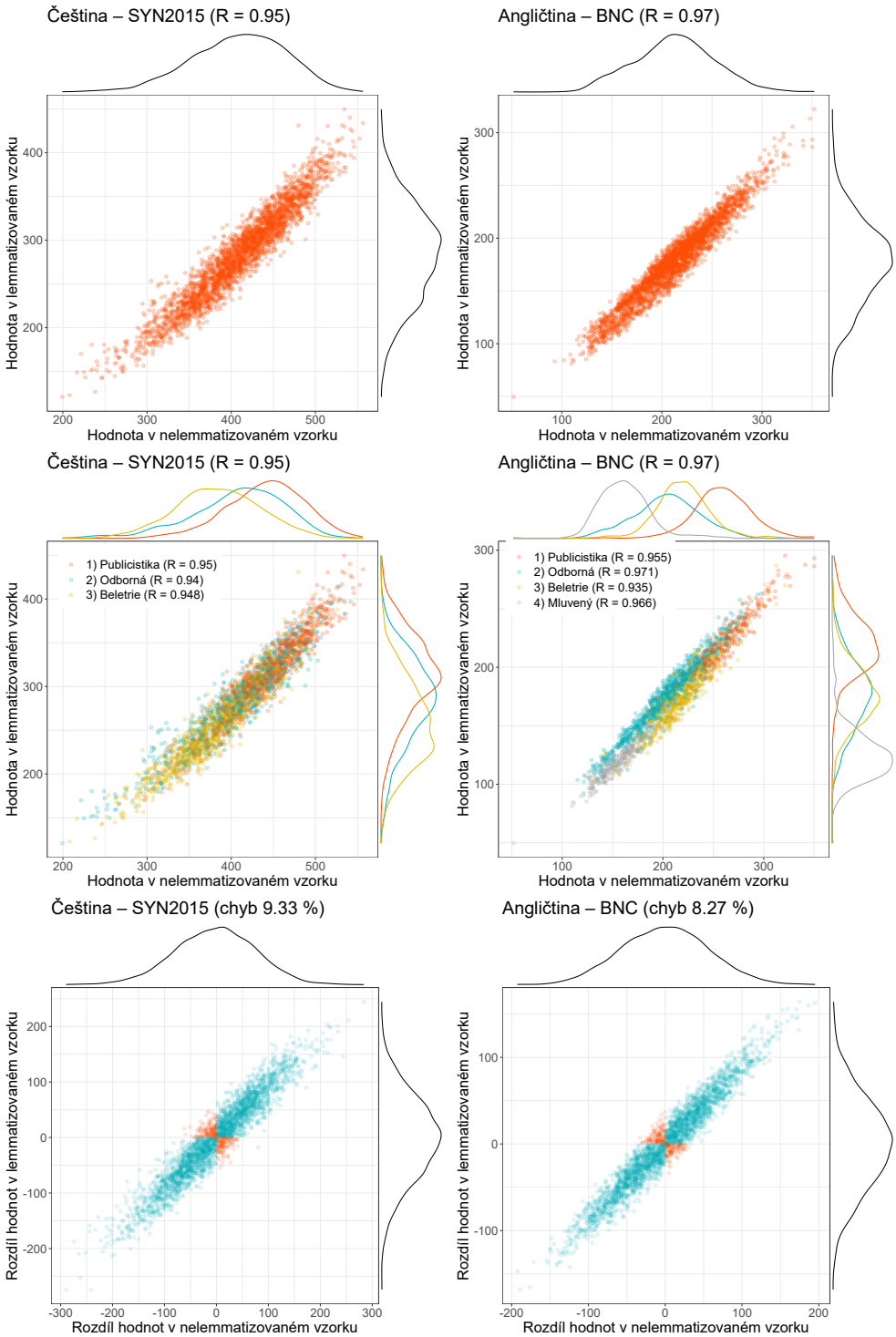
Obrázek 4.1: Korelace počtu typů nelemmatizovaného a lemmatizovaného textu.



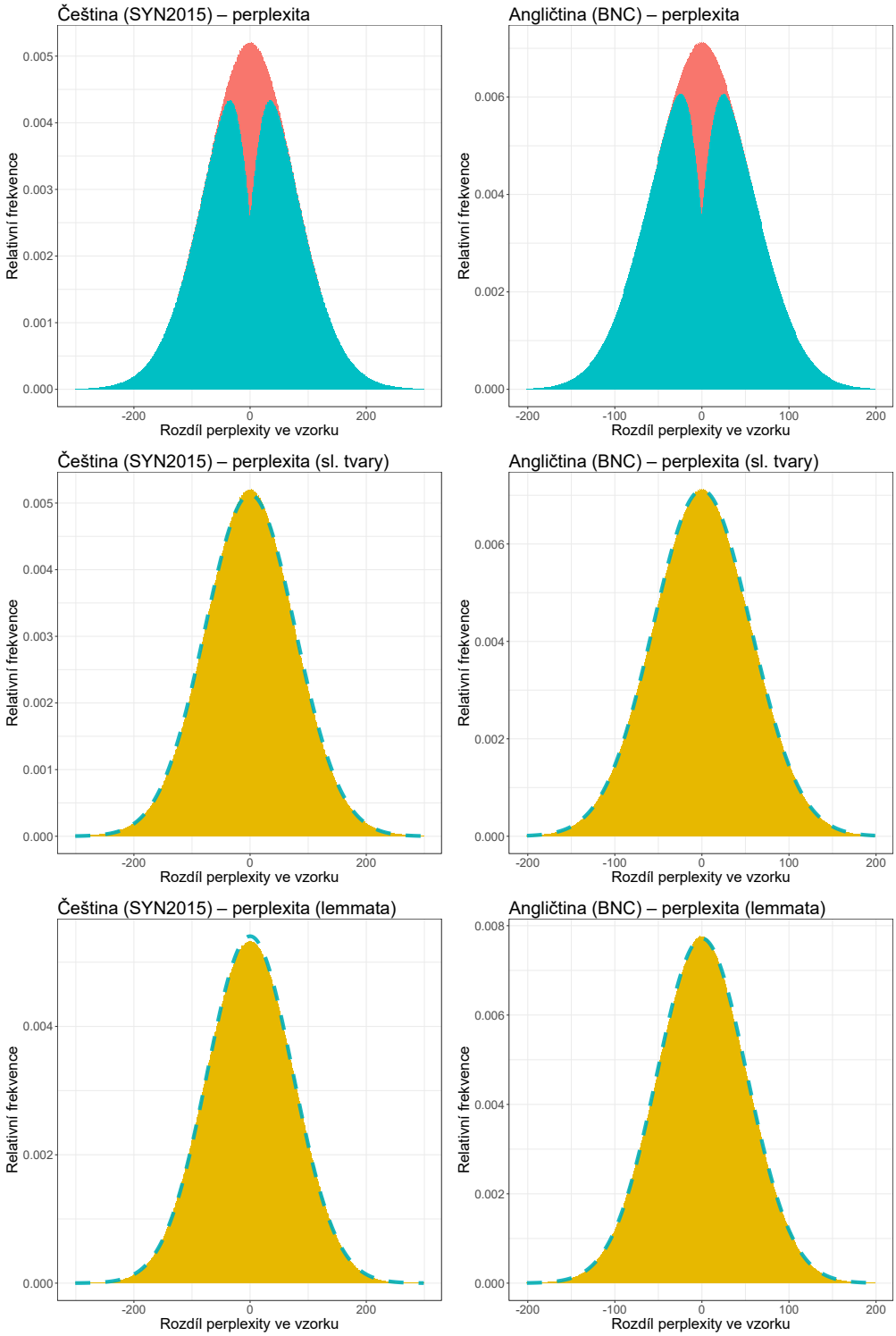
Obrázek 4.2: Distribuce rozdílů v počtech typů nelemmatizovaného a lemmatizovaného textu.



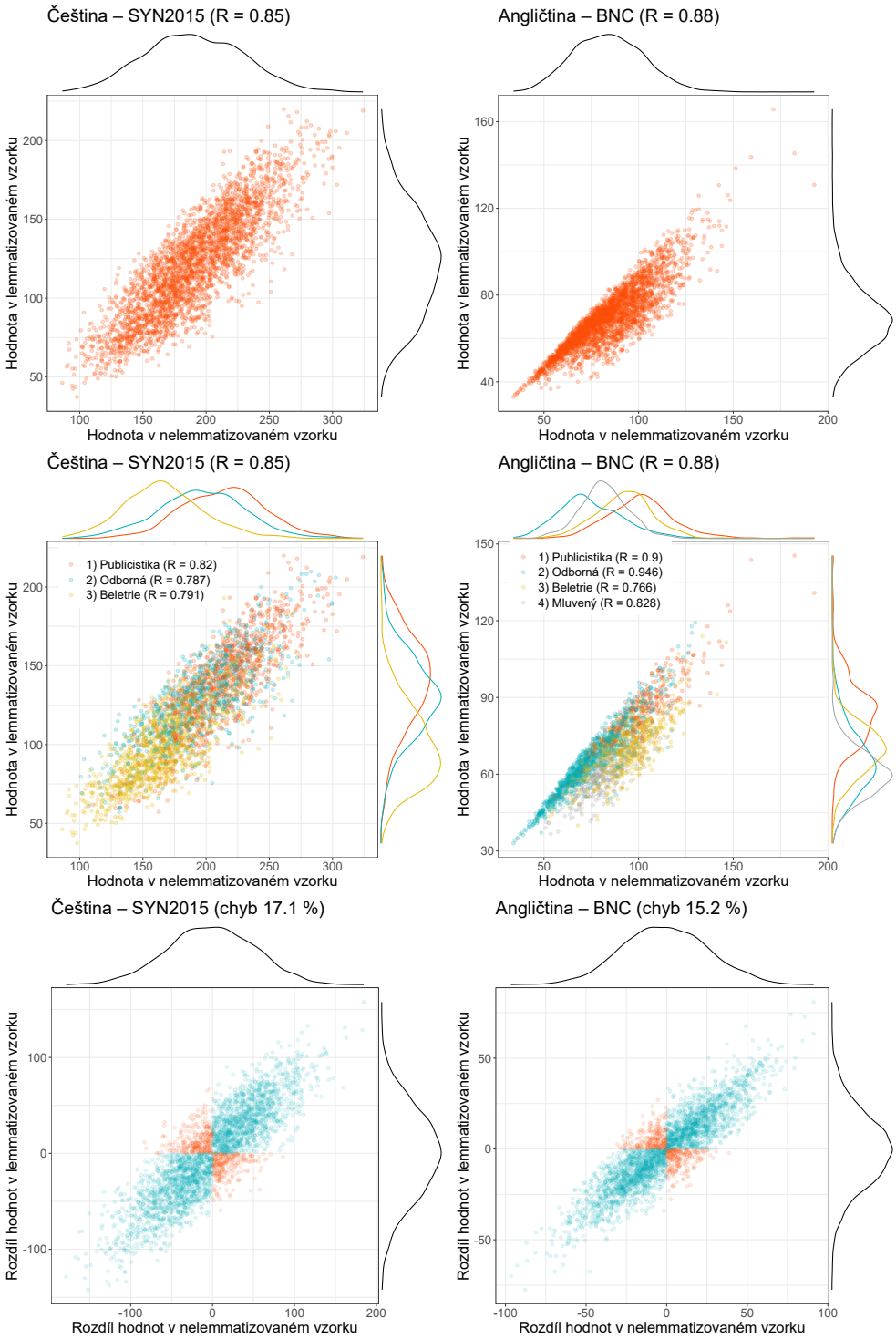
Obrázek 4.3: Korelace počtu hapax legomena nelemmatizovaného a lemmatizovaného textu.



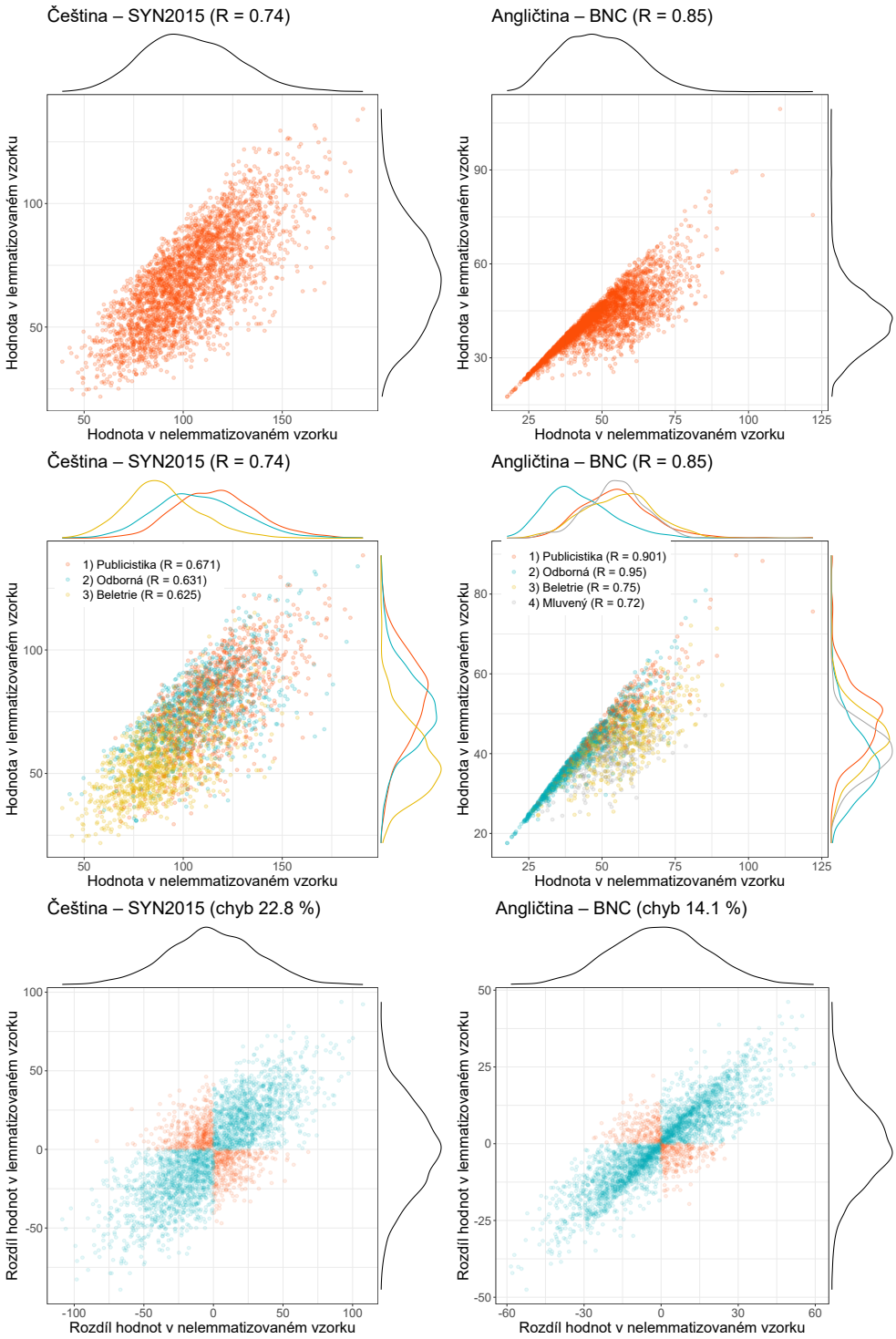
Obrázek 4.4: Korelace perplexity nelemmatizovaného a lemmatizovaného textu.



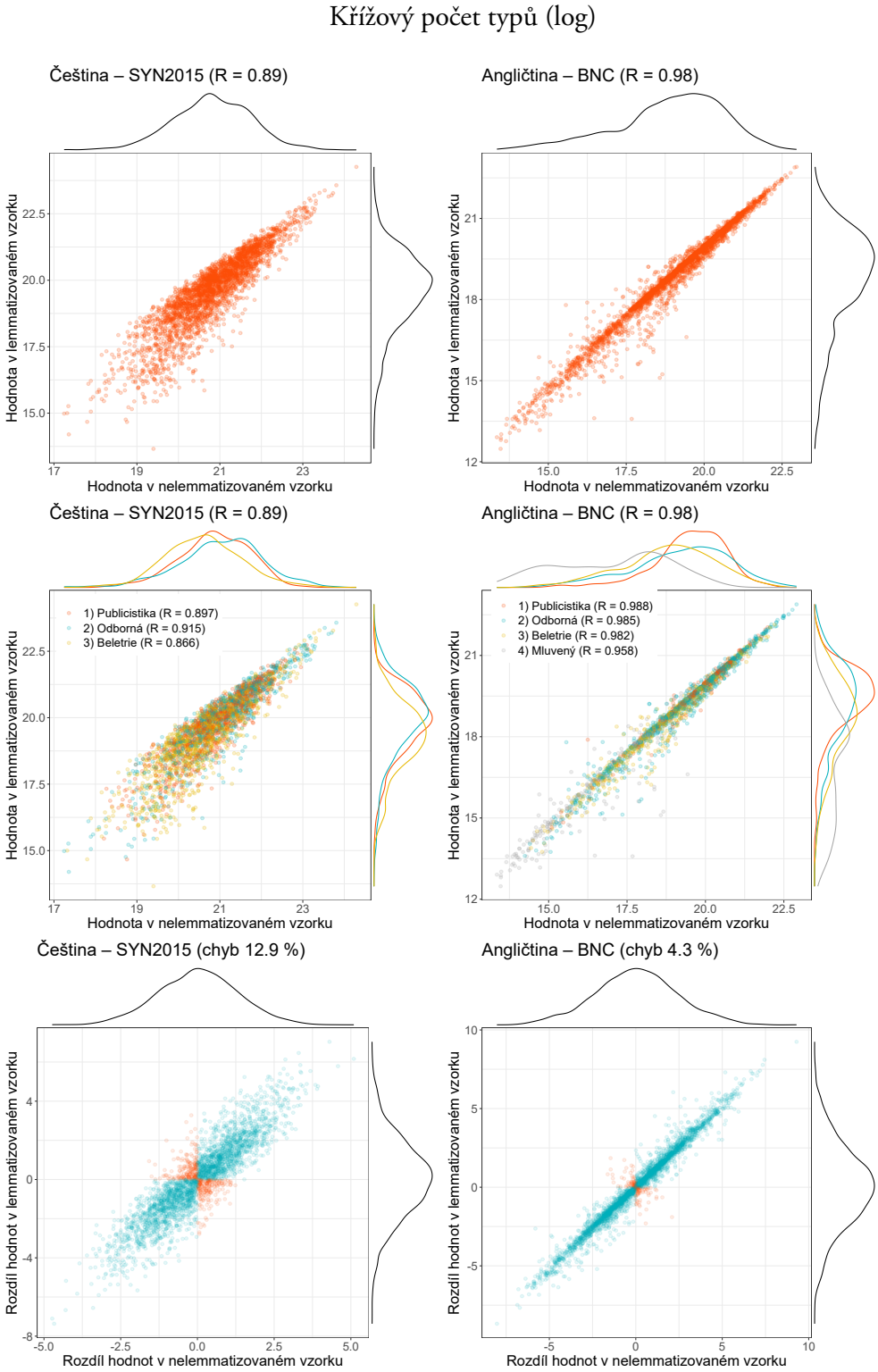
Obrázek 4.5: Distribuce rozdílů v perplexitě nelemmatizovaného a lemmatizovaného textu.



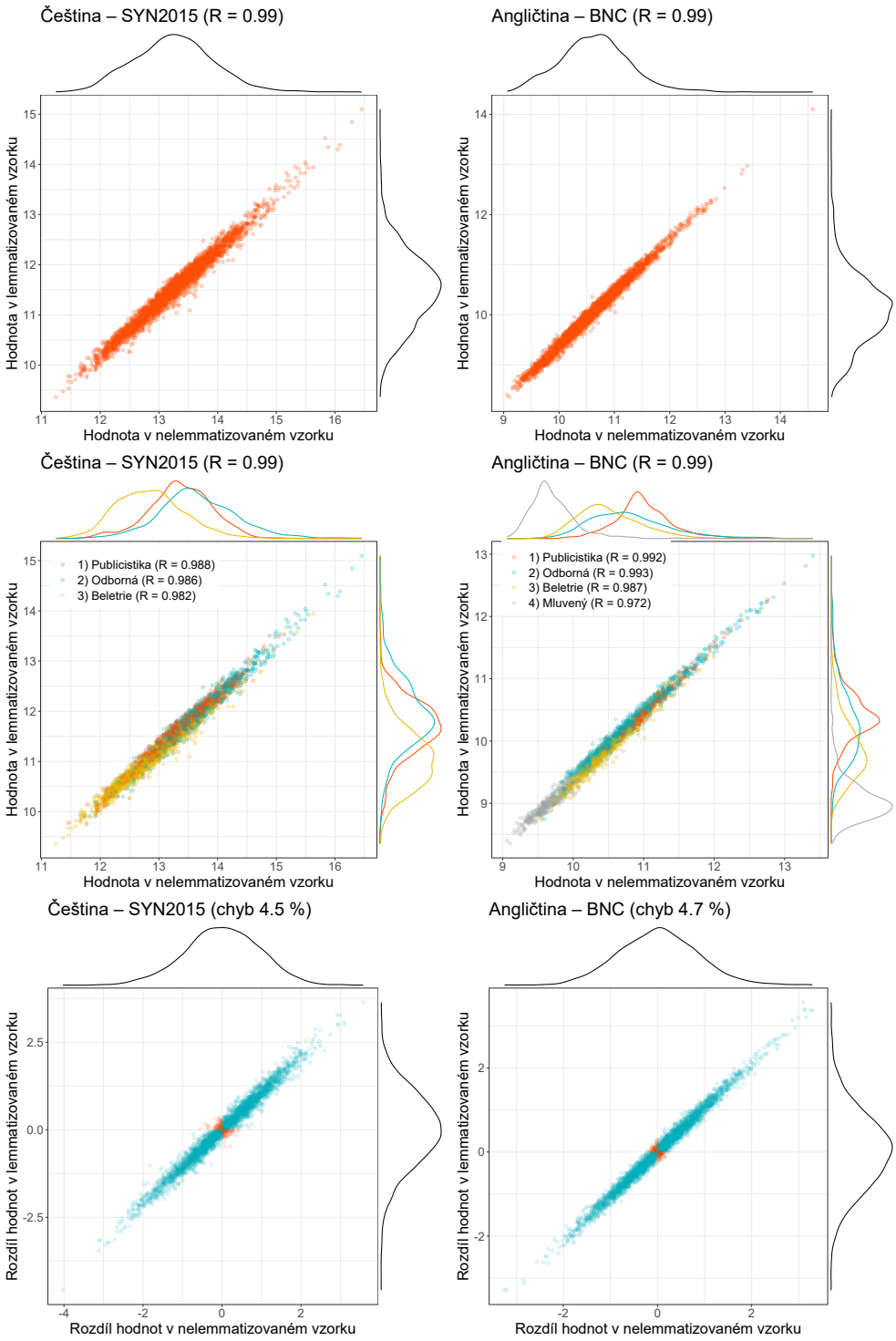
Obrázek 4.6: Korelace převrácené pravděpodobnosti opakování nelemmatizovaného a lemmatizovaného textu.



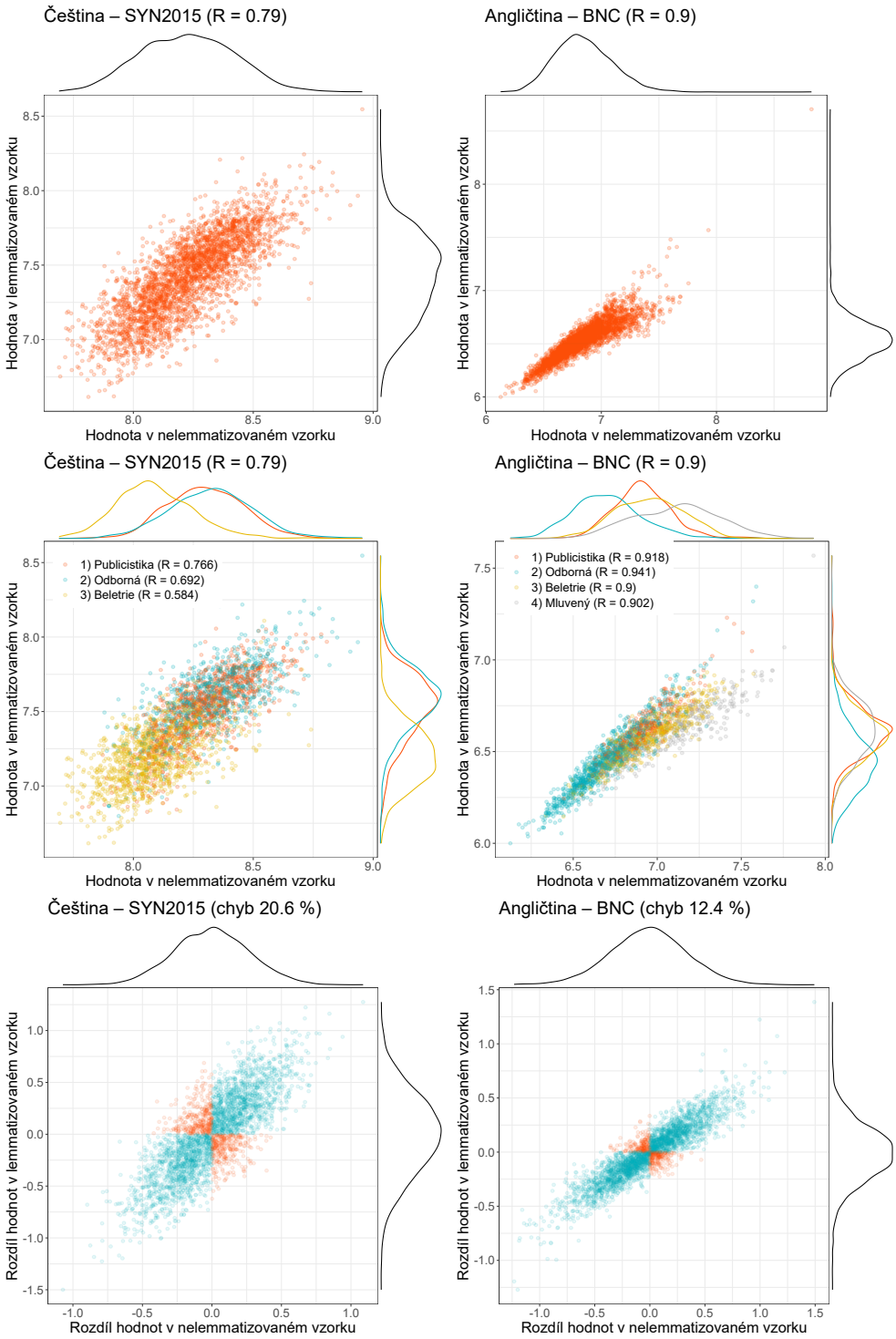
Obrázek 4.7: Korelace Hillova čísla (q = 3) nelemmatizovaného a lemmatizovaného textu.



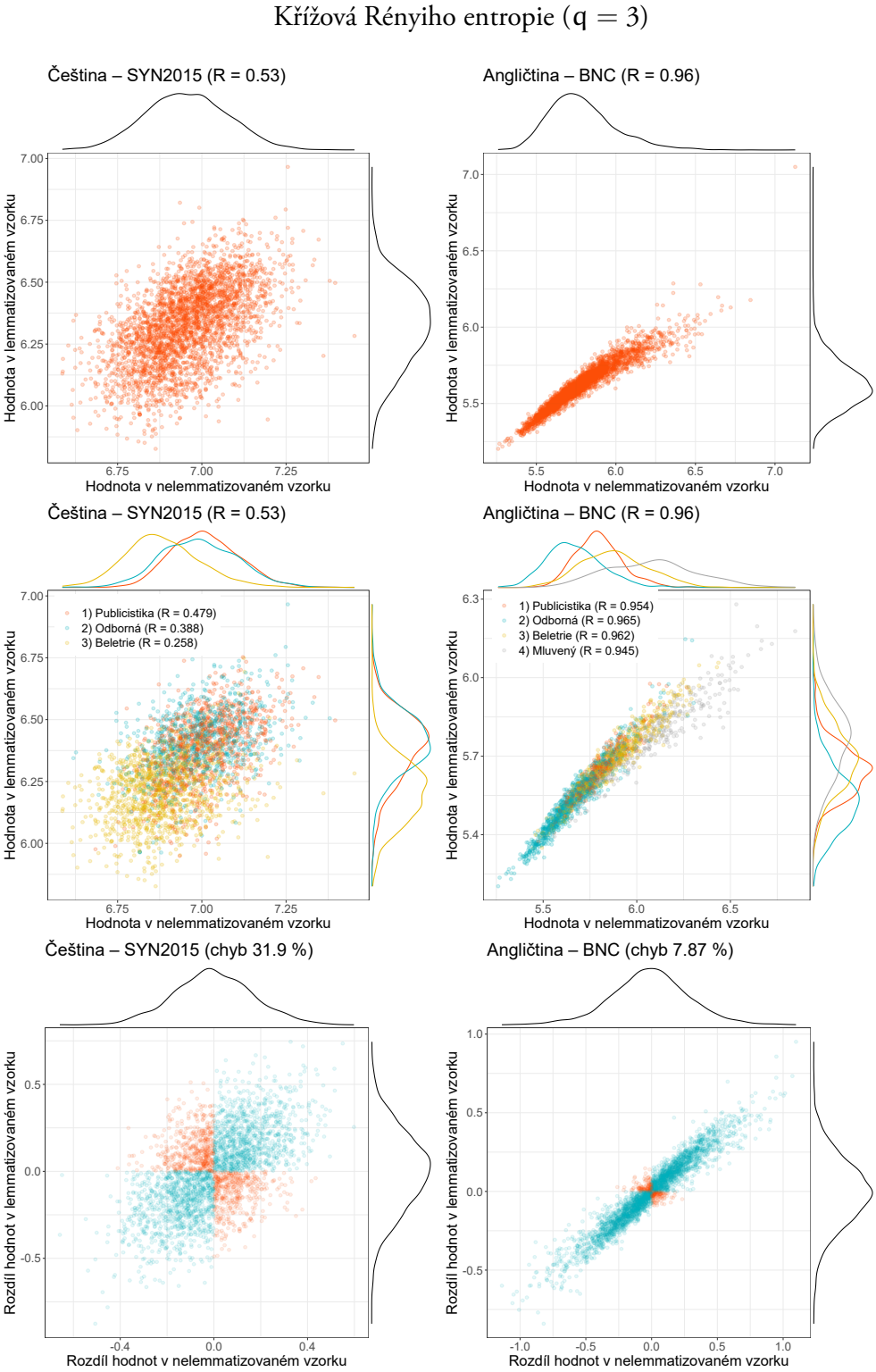
Obrázek 4.8: Korelace logaritmu křížového počtu typů nelemmatizovaného a lemmatizovaného textu.



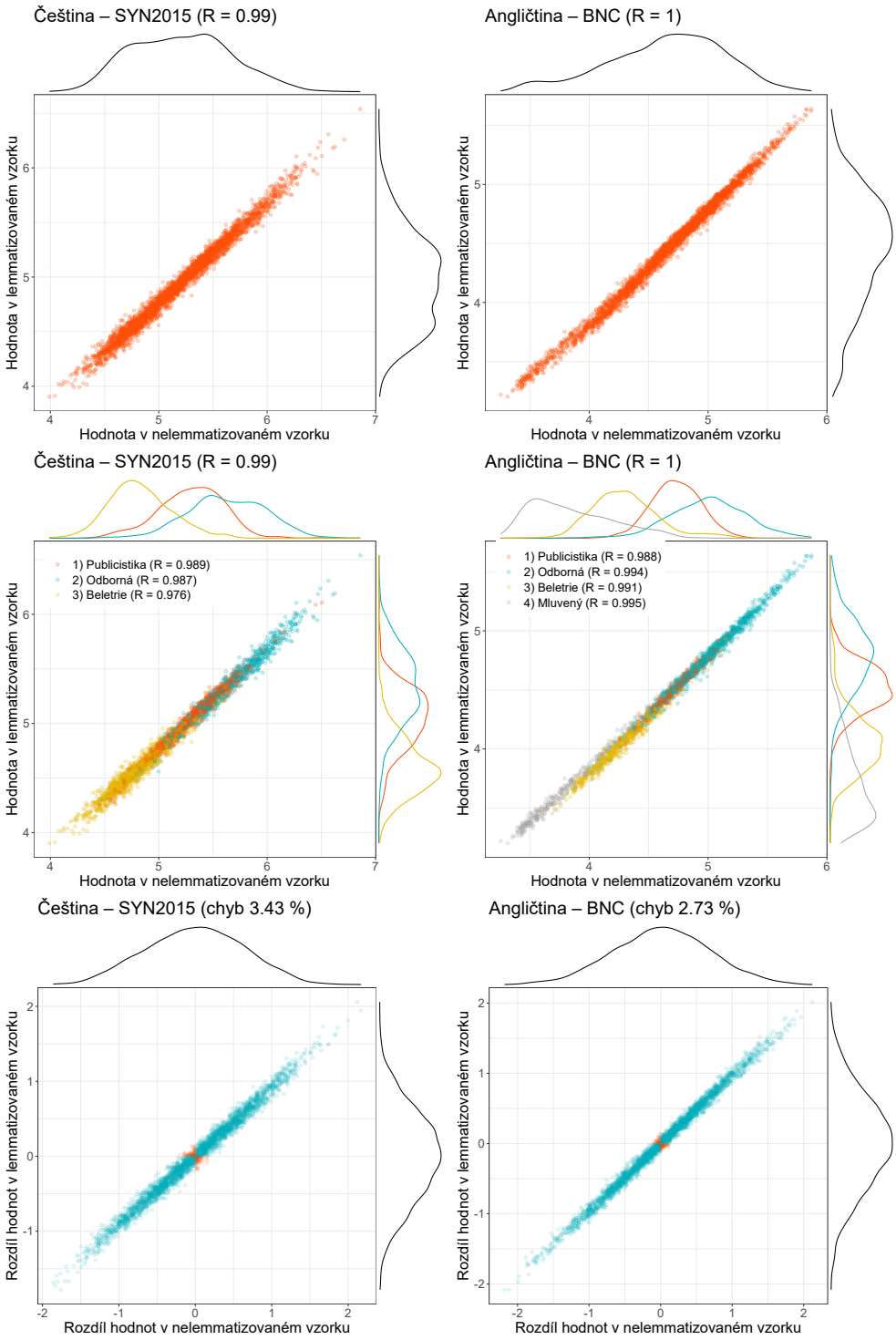
Obrázek 4.9: Korelace křížové Shannonovy entropie nelemmatizovaného a lemmatizovaného textu.



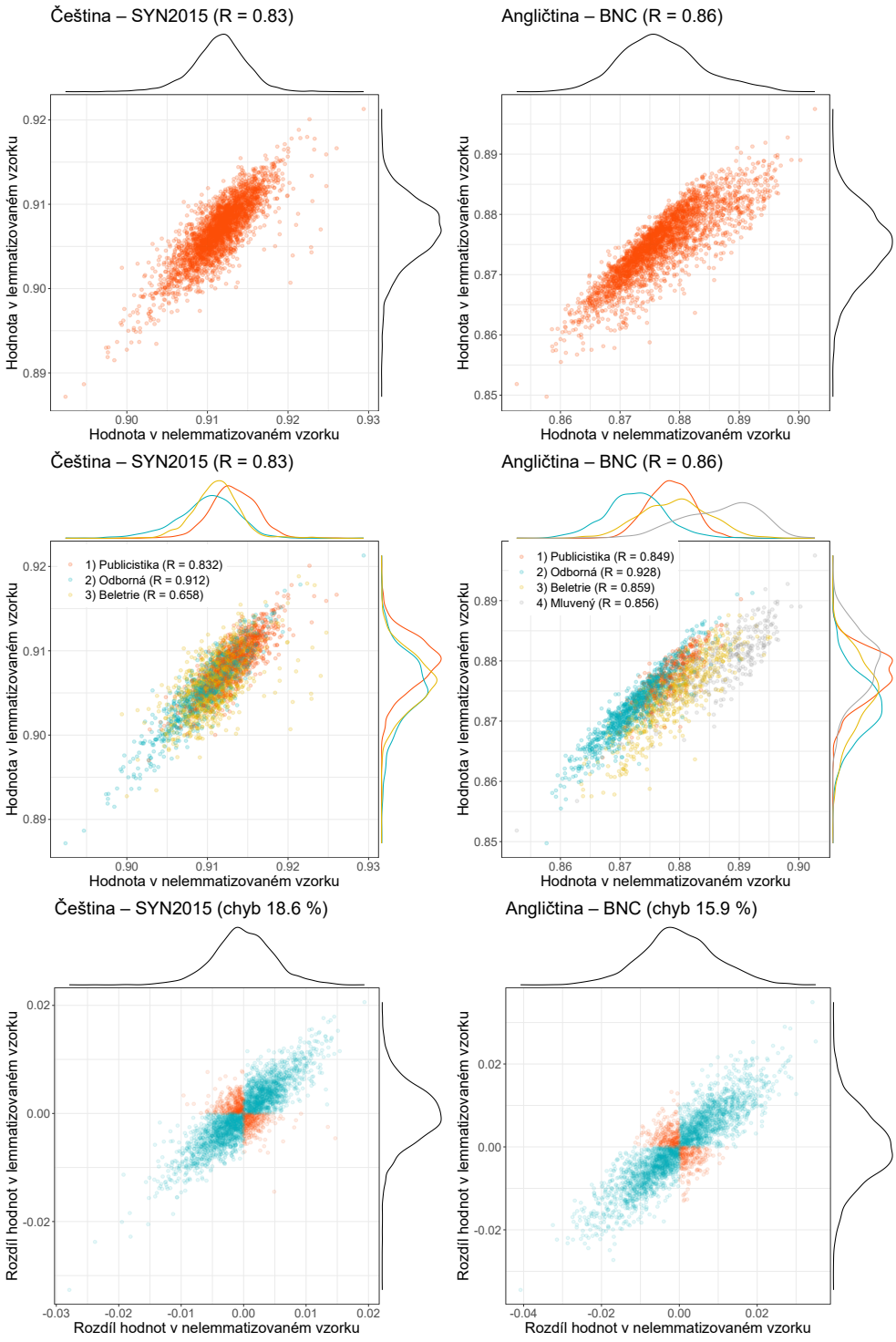
Obrázek 4.10: Korelace logaritmu křížové převrácené pravděpodobnosti opakování nelemmatizovaného a lemmatizovaného textu.



Obrázek 4.11: Korelace křížové Rényiho entropie (q = 3) nelemmatizovaného a lemmatizovaného textu.



Obrázek 4.12: Korelace délky tokenů nelemmatizovaného a lemmatizovaného textu.



Obrázek 4.13: Korelace rozdílnosti nelemmatizovaného a lemmatizovaného textu.

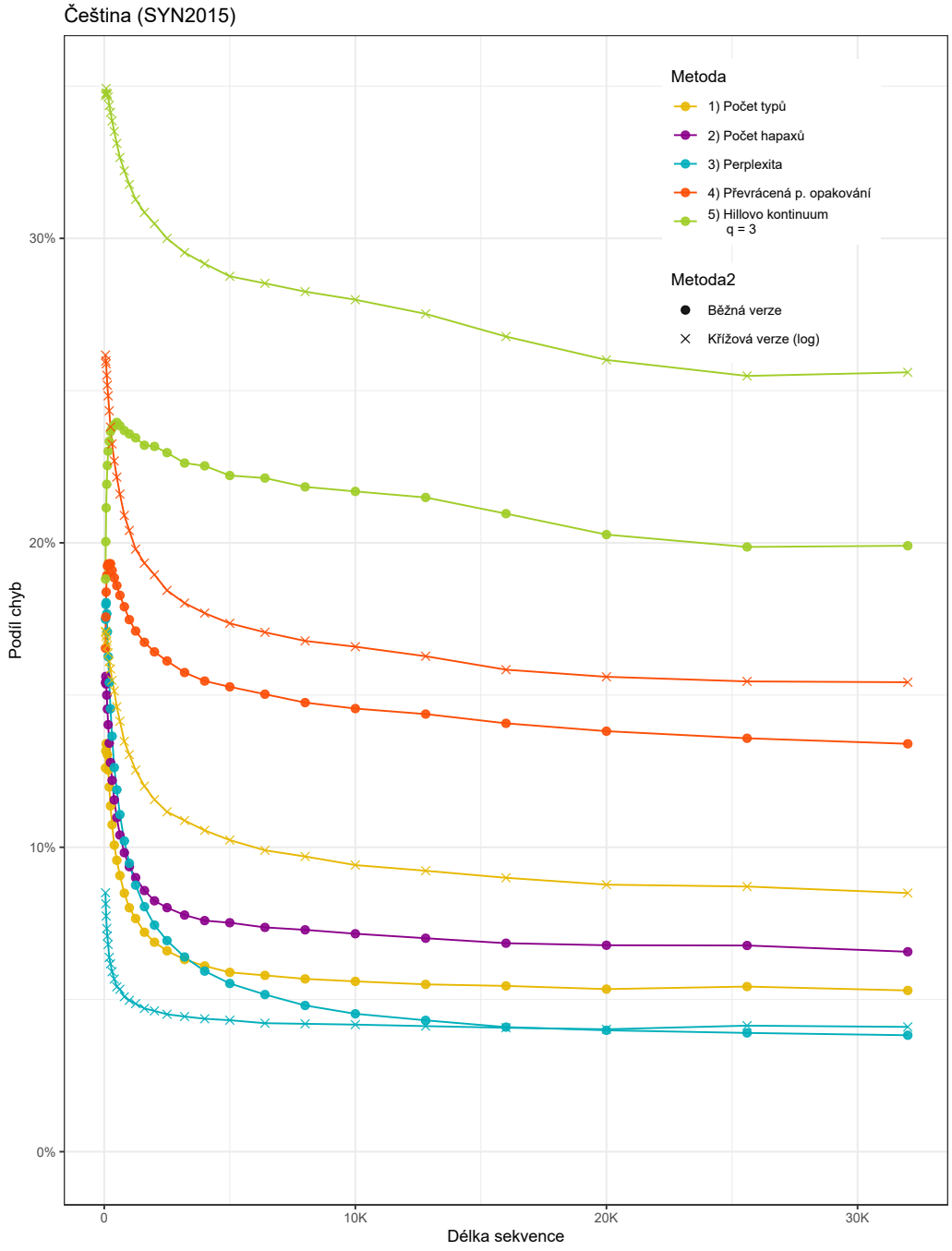
4.4 Různě dlouhé vzorky

Můžeme shrnout, že čím vyšší je hodnota parametru q a čím flektivnější je jazyk, tím míň je jedno, jestli lexikální diverzitu měříme na lemmatizovaném nebo nelemmatizovaném textu. Zní to docela rozumně a vysvětlitelně, ovšem k těmto závěrům jsme došli pouze na základě sekvencí o délce tisíc tokenů ve dvou jazycích. Co když jiné délky sekvencí přinesou jiné výsledky? Délka tisíc tokenů byla vybrána jednak kvůli tomu, že je to zhruba množství textu, které je možné napsat na jeden zátah a které se používá jako limit při zadávání esejů a slohových cvičení (cca tři normostrany), takže docela dobře může posloužit jako velikost okna při zkoumání akvizice jazyka, ale zejména kvůli numerologicko estetickým kritériím — kdybych vybral číslo řekněme 654, musel bych vysvětlit proč, zatímco takto je čtenáři jasné, že jde o arbitrární hodnotu.

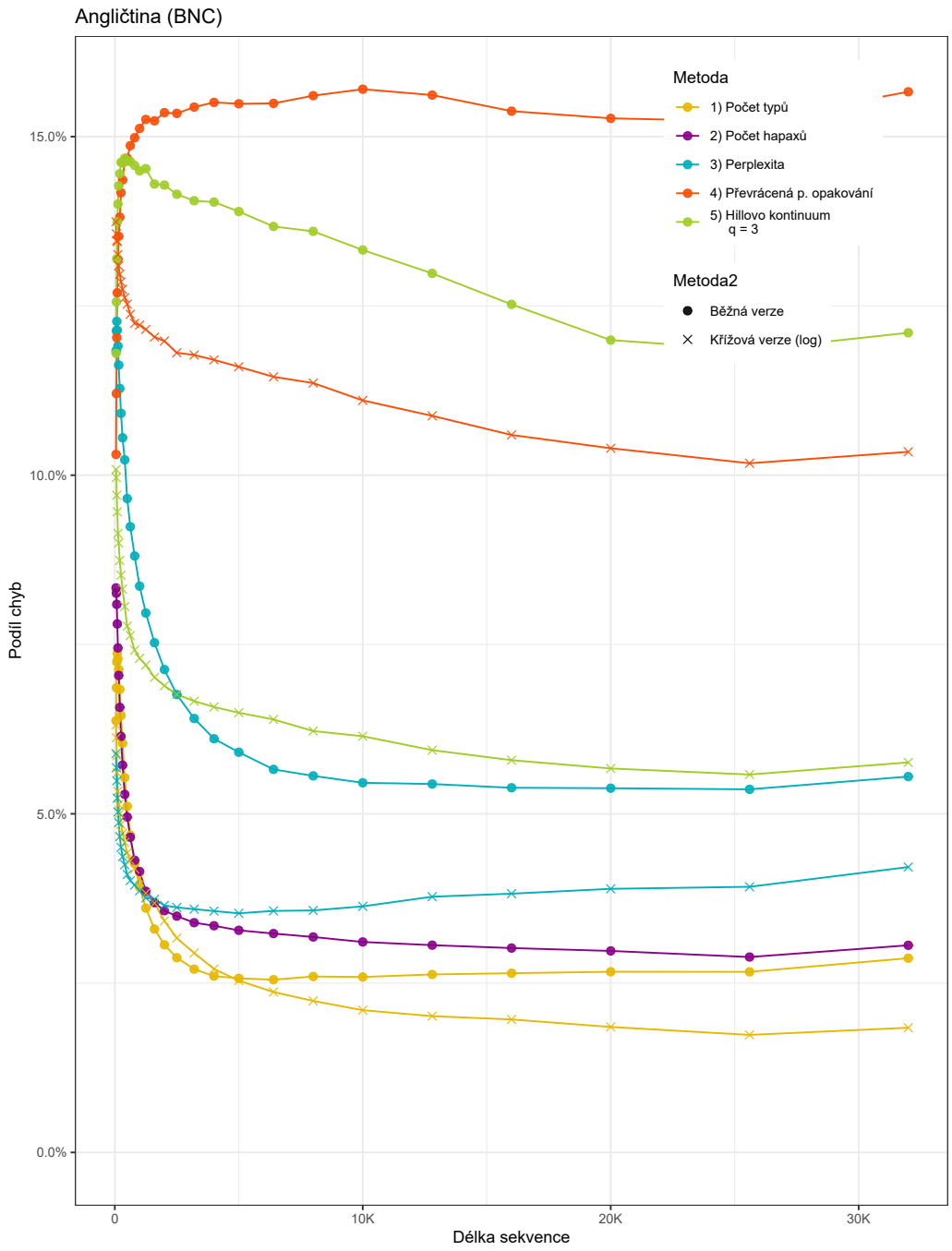
Jak vidíme na obrázcích 4.14 a 4.15, obavy jsou namístě, neboť délka textu dokáže s chybovostí docela zamíchat: právě okolo tisícího tokenu pořadí metrik není ustálené a teprve okolo desetitisícího můžeme hovořit o nějakém stabilním pořadí metrik co do úspěšnosti. Jak je vidět, naše předčasná zobecnění nejsou tak obecná, jak by se nám líbilo, chybovost sice evidentně roste s parametrem q i pro jiné délky sekvencí, ovšem u delších sekvencí vzatých z anglických textů nám toto pravidlo narušuje křížová Rényiho entropie pro vyšší q.

Vzhledem k vyšší variabilitě lexikální diverzity u krátkých sekvencí obecně platí, že čím je sekvence delší, tím menší můžeme očekávat chybovost, ovšem jak je vidět z grafů, ani toto pozorování nemůžeme plně zobecnit. Zejména na začátku, v sekvencích kratších než tři tisíce tokenů, je průběh křivky netriviální a pro RRR v angličtině k poklesu na pozorovaném intervalu vůbec nedojde, pouze se zastaví růst.

Předpokládám, že je to tím, že zde působí několik protichůdných sil, jednak zmiňovaná nižší variabilita pro krátké sekvence, která tlačí chybovost dolů, jednak se zvyšující se délkou sekvence se zvyšuje i počet lemmat, která se v textu vyskytují ve více tvarech, což naopak znamená větší důležitost lemmatizace. V každém případě modelování tohoto fenoménu bude mnohem složitější úkol, než se na první pohled zdá.



Obrázek 4.14: Srovnání, jak jednotlivé metriky ovlivňuje lemmatizace (české texty).



Obrázek 4.15: Srovnání, jak jednotlivé metriky ovlivňuje lemmatizace (anglické texty).

Kapitola 5

Vliv velikosti klouzavého okna

Pokud jste četli pozorně kapitolu 2.2.1, asi jste si všimli, že jsem nepokrytý fanoušek normování metrik lexikální diverzity pomocí klouzavého okna a že bych byl rád, kdyby se metoda nyní používaná u metriky zvané MATTR rozšířila i na ostatní metriky.

Hlavním důvodem je, že tato metoda problém normování jednoznačně a jednoduše řeší, aniž bychom museli obětovat interpretovatelnost, možnost představit si hodnoty metriky jako něco konkrétního.

Ale pak je tu ještě jeden důvod: tato metoda nám dává příležitost zvolit si *velikost klouzavého okna*. Ona velikost může být zcela arbitrární, od několika tokenů, kde nemůžeme čekat přílišnou variabilitu výsledků, až po velikost textu. Respektive, pokud srovnáváme více textů, tak velikost nejkratšího srovnávaného textu.

Mnozí tuto arbitrárnost vidí jako nevýhodu, pamatuji si na rozhovor s Reinhardem Köhlerem, který navrhoval, abychom zkusili najít nějakou ideální délku klouzavého okna, která by se pak univerzálně používala (ať už obecně, nebo pro ten který jazyk). Tedy že bychom si stanovili nějaká kritéria, například variabilitu nebo schopnost rozdělovat autory podle autorského stylu, a pak se snažili najít velikost okna, při které se daná kritéria maximalizují.

Osobně tuto arbitrárnost vidím naopak jako klíčovou výhodu této metodiky. Je to jedinečná příležitost změřit lexikální diverzitu na různých úrovních.

Vezměte si nejnižší úroveň, kterou pokrývají krátká okna, dlouhá řádově řekněme desítky tokenů. Někteří autoři s gustem opakují stejná slova v jednom odstavci nebo i větě. A naopak jsou autoři, kteří by si raději ukousli ruku, než aby na jedné stránce napsali slovo se stejným kořenem jako už jednou použité. Jsou literární tradice, ve kterých učitelé slohu nabádají k nahrazování stejných slov synonymy, a jiné, kde se použití slova se stejným základem považuje za estetickou figuru. Jsou textové typy a žánry, kde nejdůležitější je terminologická přesnost a autoři nemají příliš na výběr, jaké slovo použít, v jiných zase má autor nekonečno možností, jak danou myšlenku opsat metaforou.

Pojďme o úroveň výš, do hájemství oken zvíci stovek tokenů. Teprve tady se projeví celková stylistická úroveň. Vyšňořenost na úrovni vět a odstavců neznamená, že autor nevyčerpá svou zásobu synonym a metafor už na druhé stránce a že se nezačne opakovat.

Ještě vyšší úrovně. Okna dlouhá tisíce slov by měla odhalit, jak tematicky rozkročený text je na úrovni jedné kapitoly, okna délky desetitisíců slov pak shrnují lexikální diverzitu a tematickou promiskuitu na úrovni celých knih.

Už Covington (2010) navrhuje, abychom srovnávali MATTR pro různě velká okna, čímž odhalíme, jestli je nízká hodnota lexikální diverzity způsobená opakováním těchtýž slov na krátkém úseku, nebo jestli je dána tematickou jednotností daného díla, nebo jednoduše malou slovní zásobou autora.

Tato tvrzení bychom měli rigorózně potvrdit rozsáhlou studií na velkém množství textů. Ideálně za přispění literárních vědců a vůbec lidí, s citem pro literaturu a literární styl. Takové teď zrovna nemám po ruce, takže se omezím na drobnou sondu, která 1) ukáže, jestli je mezi měřeními pomocí různě velkých oken vůbec nějaký rozdíl; 2) prozkoumá, jak se v prostoru vyhrazeném dvěma velikostmi oken umisťují texty různých typů a modalit.

5.1 Metodika

Texty označíme podle jejich typů a modalit stejně jako ve 4. kapitole. A podobným způsobem z těchto textů vybereme vzorek sekvencí o délce tisíce tokenů a změříme na něm naše oblíbené metriky lexikální diverzity. Následně na stejných sekvencích změříme tytéž metriky, jenže za použití klouzavého okna o délce sta slov. Výsledek vidíte na následujících grafech v levém sloupci, každý bod představuje jednu sekvenci.

Totéž provedeme pro vzorek náhodně vybraných sekvencí o délce deseti tisíc tokenů, tentokrát ovšem druhotné měření provedeme pomocí klouzavého okna o velikosti tisíce tokenů. Výsledky tohoto měření obsadily levý sloupec.

5.2 Rozdíl mezi krátkými a dlouhými okny

U žádné z metrik patřících do Hillova kontinua není korelace zrovna velká, metriky pracující s různě velkým klouzavým oknem tedy popisují různé, byť podobné fenomény a přinášejí různé informace (obrázky 5.1–5.5).

Asi nejzajímavější zjištění v tomto ohledu je, že korelace mezi stotokenovým oknem a tisícitokenovým oknem je systematicky větší než korelace mezi tisícitokenovým oknem a desetittisícitokenovým oknem. Tedy efektu délky s přibývající délkou ubývá, což bychom, při pohledu na křivku vztahu typů a tokenů (type-token relation, připomínám obrázky 1.12 a 1.13), mohli docela i čekat.

Korelace je systematicky vyšší v arabštině než v češtině a angličtině, zejména pro krátké sekvence, což by nás nemělo překvapit, neboť v arabské stylistické tradici není tak akcentována nutnost nahrazovat použitá slova synonymy, odpadá tak jeden stupeň volnosti.

Podobně jako v případě lemmatizace lze s rostoucím parametrem q vysledovat určitý pokles korelace, ale rozhodně neklesá tak markantně jako v případě lemmatizace (viz kapitolu 4.1) — ostatně pro tento jev mě nenapadá ani žádné rozumné teoretické vysvětlení.

Otázka je, jestli je lineární korelace vůbec dobrý model pro takové srovnání. Právě u arabštiny i pouhým okem vidíme, že závislosti metrik změřených na různě velkých oknech lineární právě nejsou, zejména u metrik s vyšším parametrem q, tato tendence je ovšem méně zřetelná až neznatelná u ostatních dvou jazyků. Pokud bychom tedy chtěli zjistit, jestli je to obecné pravidlo nebo nějaká specialita daná složením CLAUDie, museli bychom prozkoumat množství korpusů. Osobně pro to nenacházím žádné dobré teoretické vysvětlení, takže spíše se přikláním k variantě, že se jedná o náhodu.

Metriky spadající pod Rényiho křížové entropie mají korelace mnohem vyšší než jejich nekřížové varianty, v některých případech téměř lízající jedničku, takže různě velká okna nepřinášejí mnoho různých informací o textu (obrázky 5.6–5.9). To by nás nemělo překvapit, neboť závisí z velké míry na frekvencích typů v referenčním korpusu, které se s velikostí okna nemění. Pokud průměrnou délku slova, za podmínek ideálního kódování, přirovnáváme ke křížové entropii s ultimátně velkým referenčním korpusem (podkapitola 1.7), pak nás invariance křížové entropie tváří v tvář změně velikosti klouzavého okna musí vyloženě potěšit, neboť délka slova je právě takto invariantní.

5.3 Klastrování pomocí lexikální diverzity dlouhých a krátkých oken

Jak jsem již zmínil, rozdíl mezi tisícitokenovými a desetitisícitokenovými sekvencemi je menší než v případě kratších sekvencí. Ovšem i tyto dlouhé sekvence dokáží hezky ve vzniklém prostoru rozdělit sekvence na jednotlivé textové typy a modality.

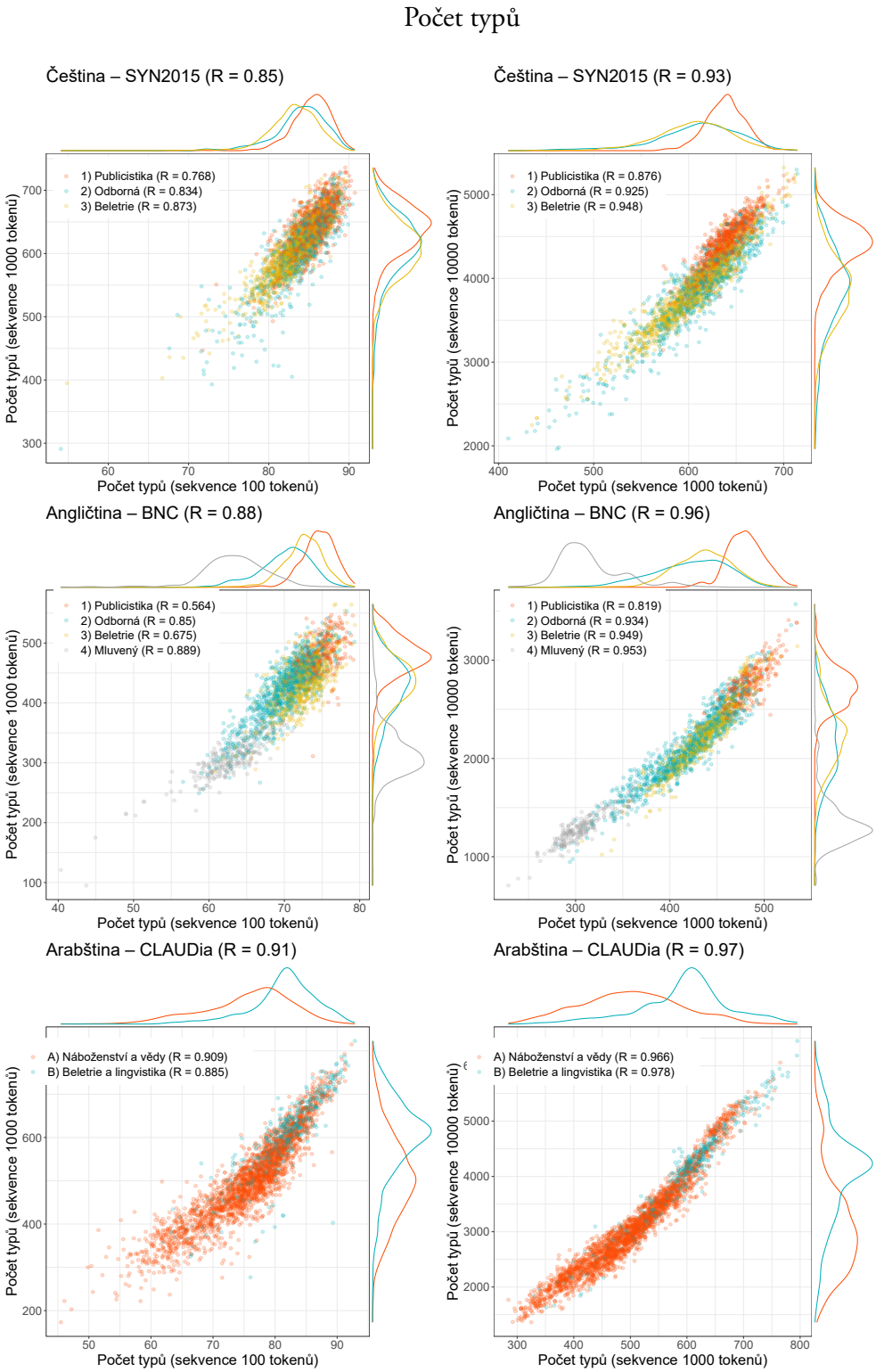
Podivuhodné je, jak nám sekvence rozklastrovala perplexita v angličtině: jednotlivé textové typy a modality zaujímají ve vzniklém prostoru prakticky stejné pozice jako v případě lemmatizovaného a nelemmatizovaného textu.[1] Nemyslím si, že by podobnost vycházela z toho, že vztah mezi perplexitou lemmatizovaného a nelemmatizovaného anglického textu je stejný jako vztah mezi perplexitou měřenou pomocí

[1] Srovnejte obrázek 5.3 s obrázkem 4.4. Pokud nemáte dobrou vizuální paměť, můžete si graf představit jako obrázek ryby s červenou hlavu, žlutým břichem, modrými zády a šedivou ocasní ploutví. Podobnost je tak zjevná, že jsem pro jistotu zkontroloval, jestli jsem neudělal chybu a na osu y nedal hodnoty pro lemmatizovaný text. Nedal.

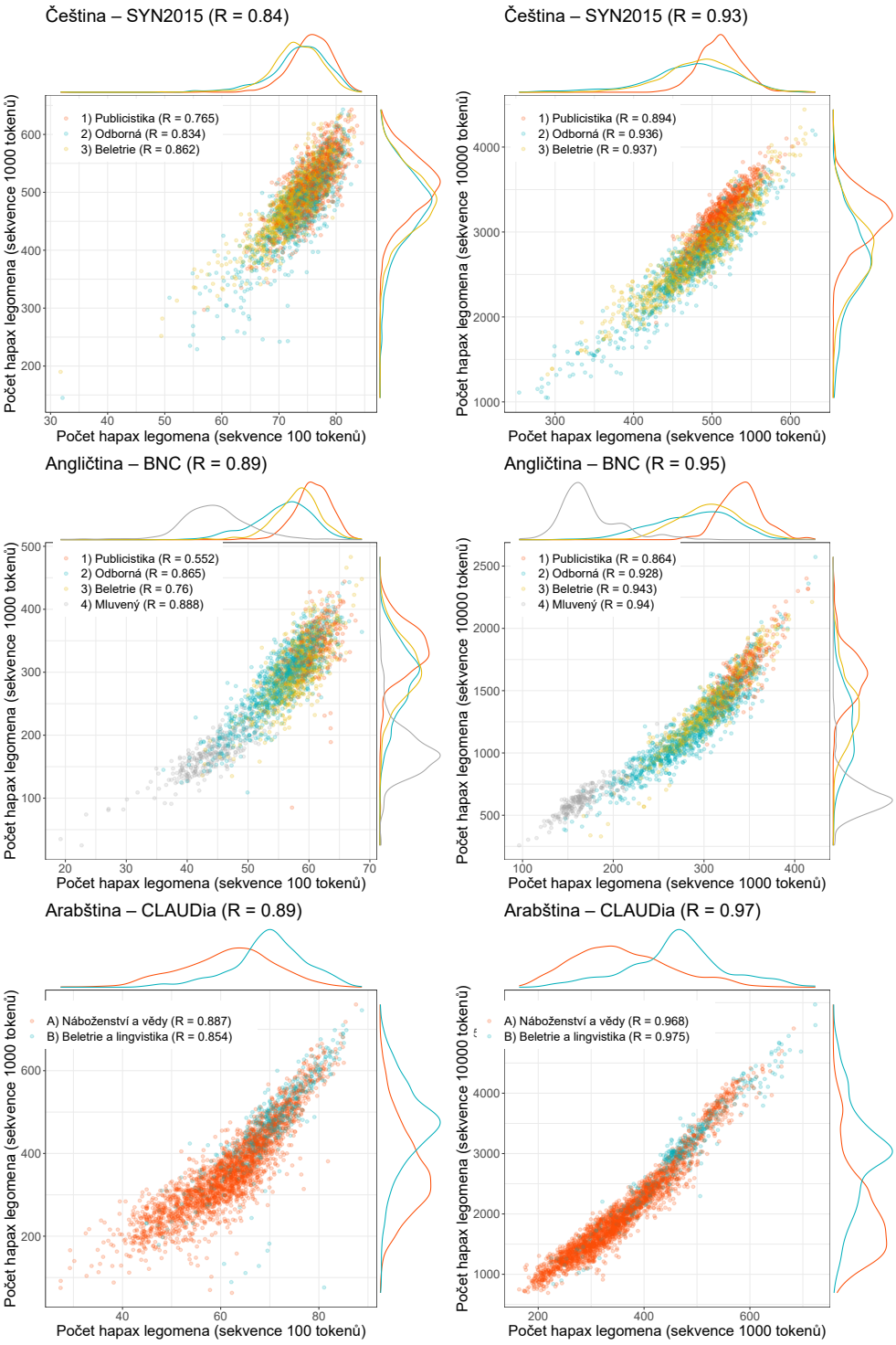
krátkého a dlouhého okna, spíše se jedná o náhodu, ostatně pro další metriky se sekvence chovají jinak.

To, že se sekvence vybrané z beletrie umístily pod sekvencemi z odborných textů, dává docela dobrý smysl: autor beletristického textu má mnohem větší motivaci vyhýbat se stejným slovům v rámci věty, odstavce či kapitoly než autor odborného textu. Přestože tedy může mít odborný text obrovskou lexikální diverzitu, danou velkým množstvím termínů a celkovou polytematičností, na krátkých úsecích bude naopak terminologicky konzistentní a stylisticky chudý.

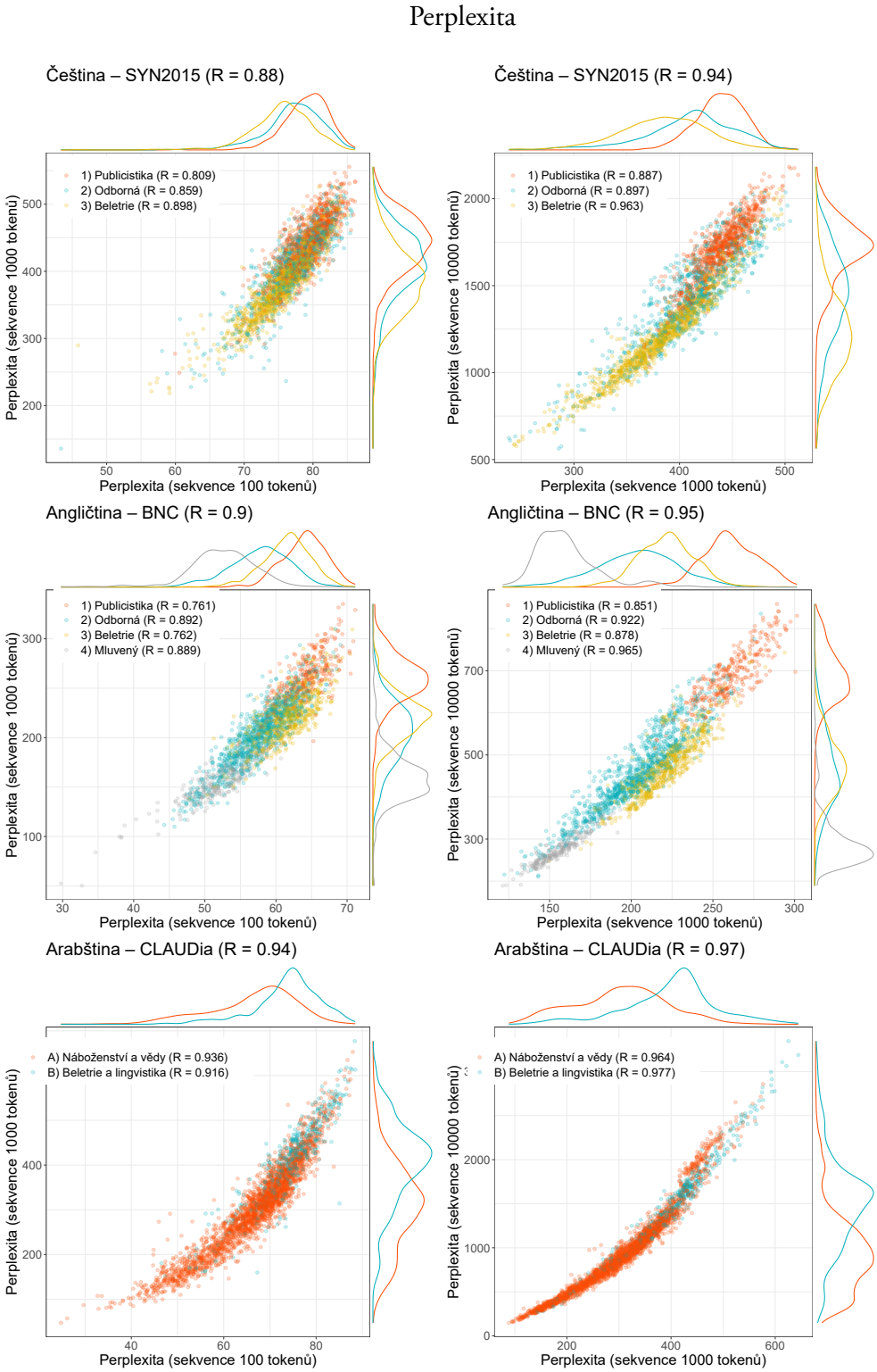
Zejména výsledky pro češtinu příjemně překvapily, neboť po tom, co jsme viděli v kapitole o vlivu lemmatizace (4.3), jsem trochu pochyboval o tom, jestli textové typy v SYNu 2015 nejsou třeba nějak špatně definovány, nebo jestli se vůbec v lexikální diverzitě nějak v češtině liší. Zde se ukazuje, že mezi textovými typy aspoň nějaké rozdíly jsou, i když nejsou tak výrazně viditelné na první pohled jako v angličtině a i když pořád splývají odborné texty s publicistikou do jakéhosi společného nonfiction klastru.



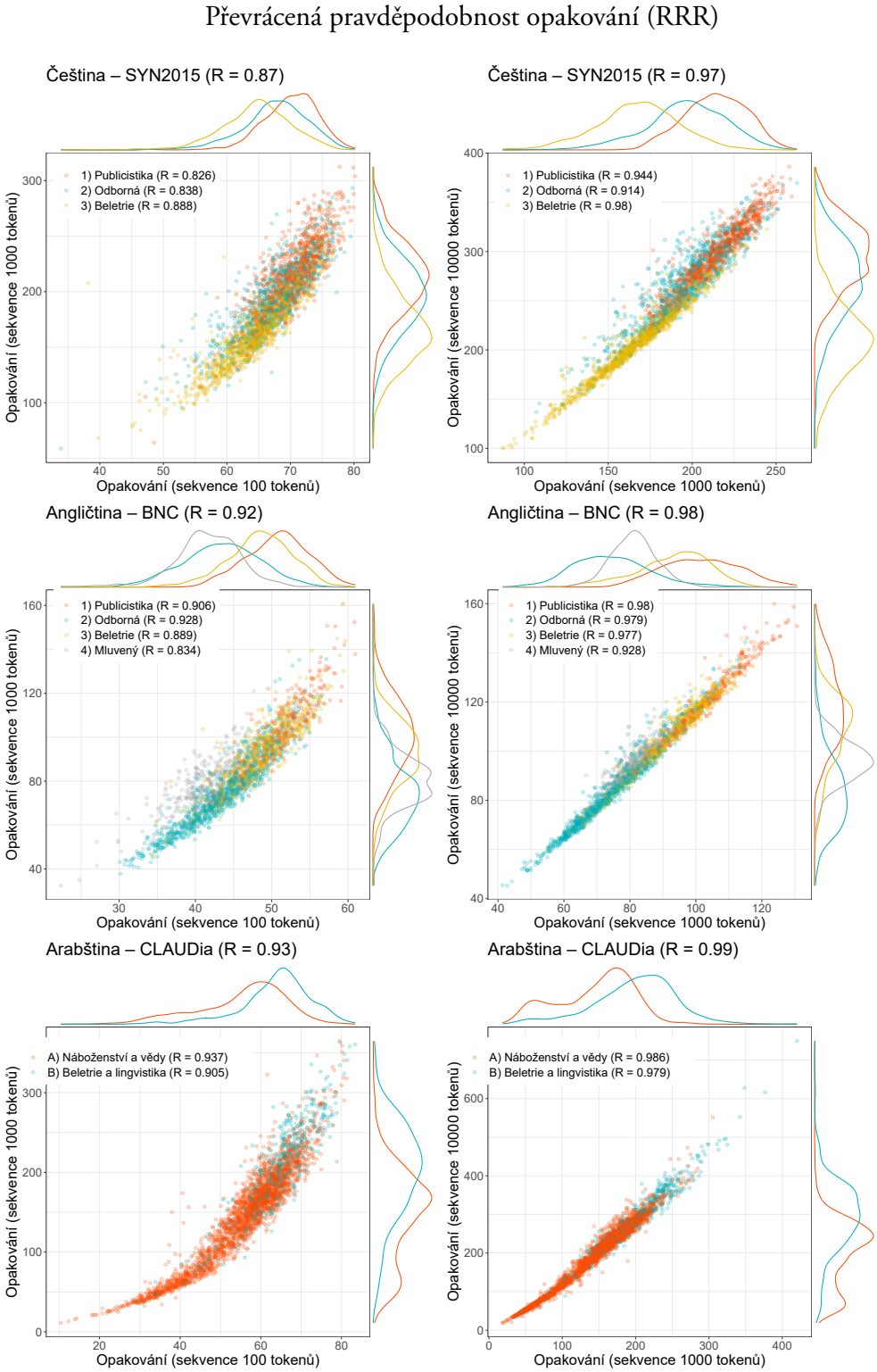
Obrázek 5.1: Korelace počtu typů v delším a kratším okně.



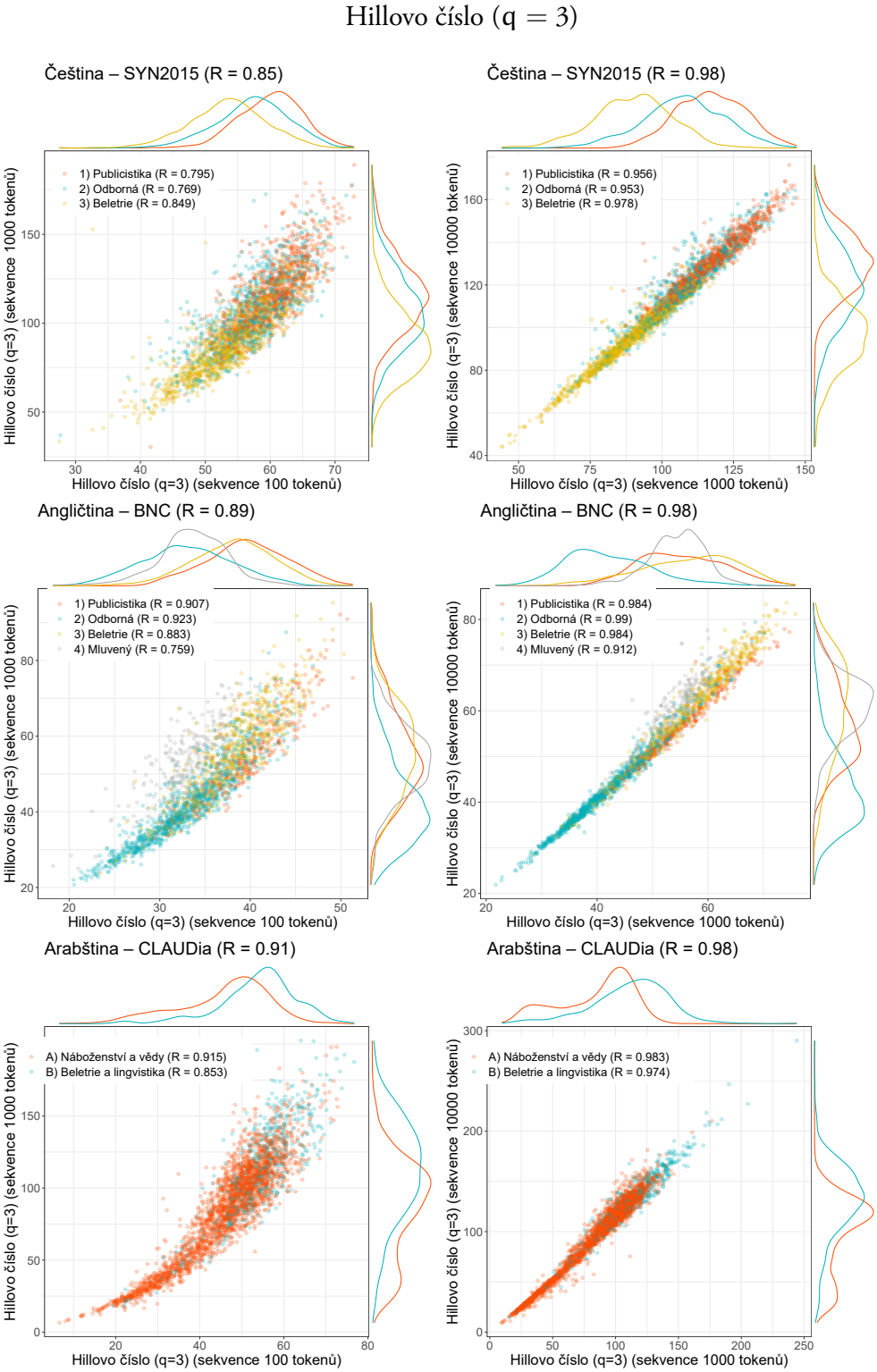
Obrázek 5.2: Korelace počtu hapax legomena v delším a kratším okně.



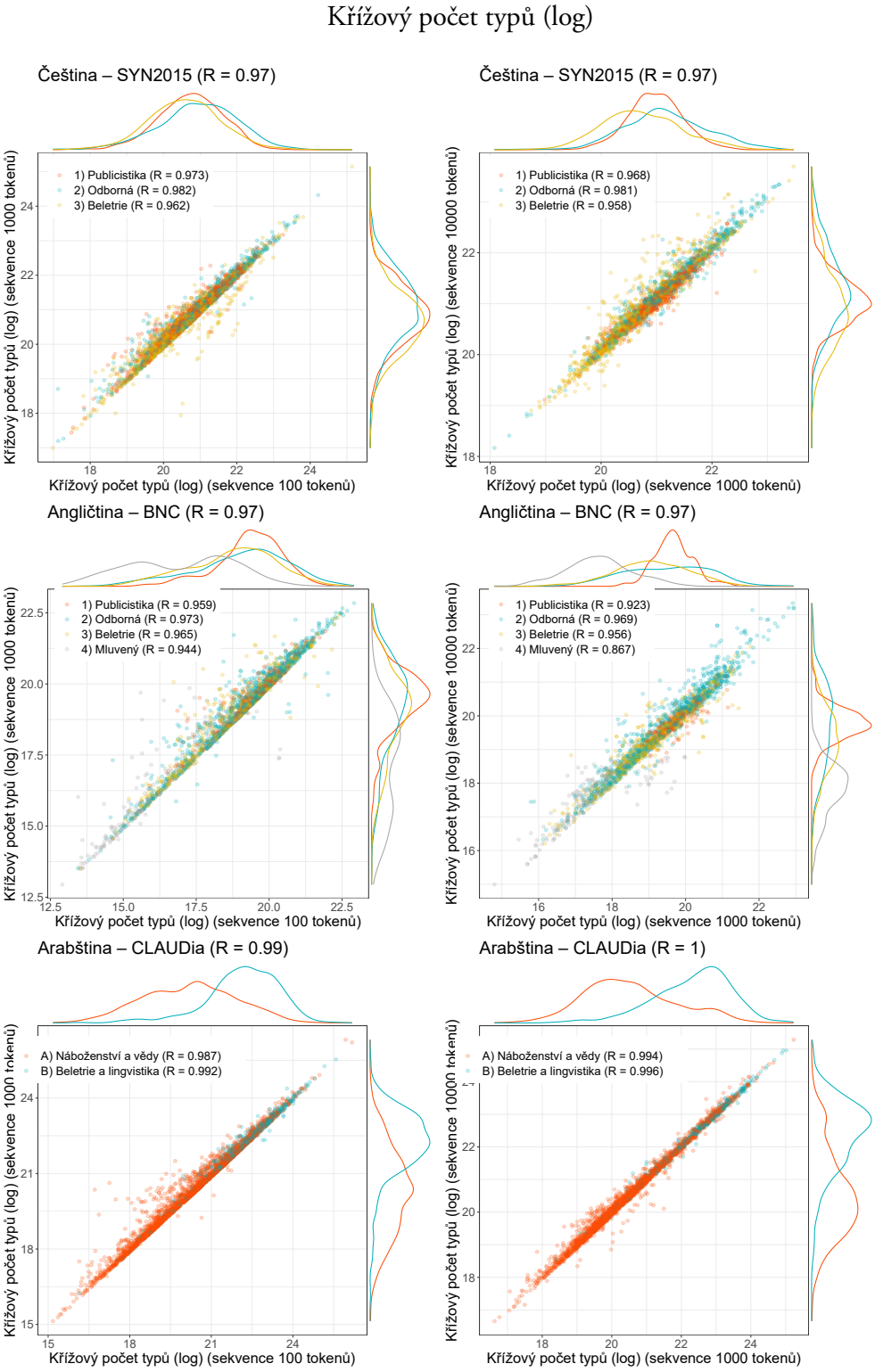
Obrázek 5.3: Korelace perplexity v delším a kratším okně.



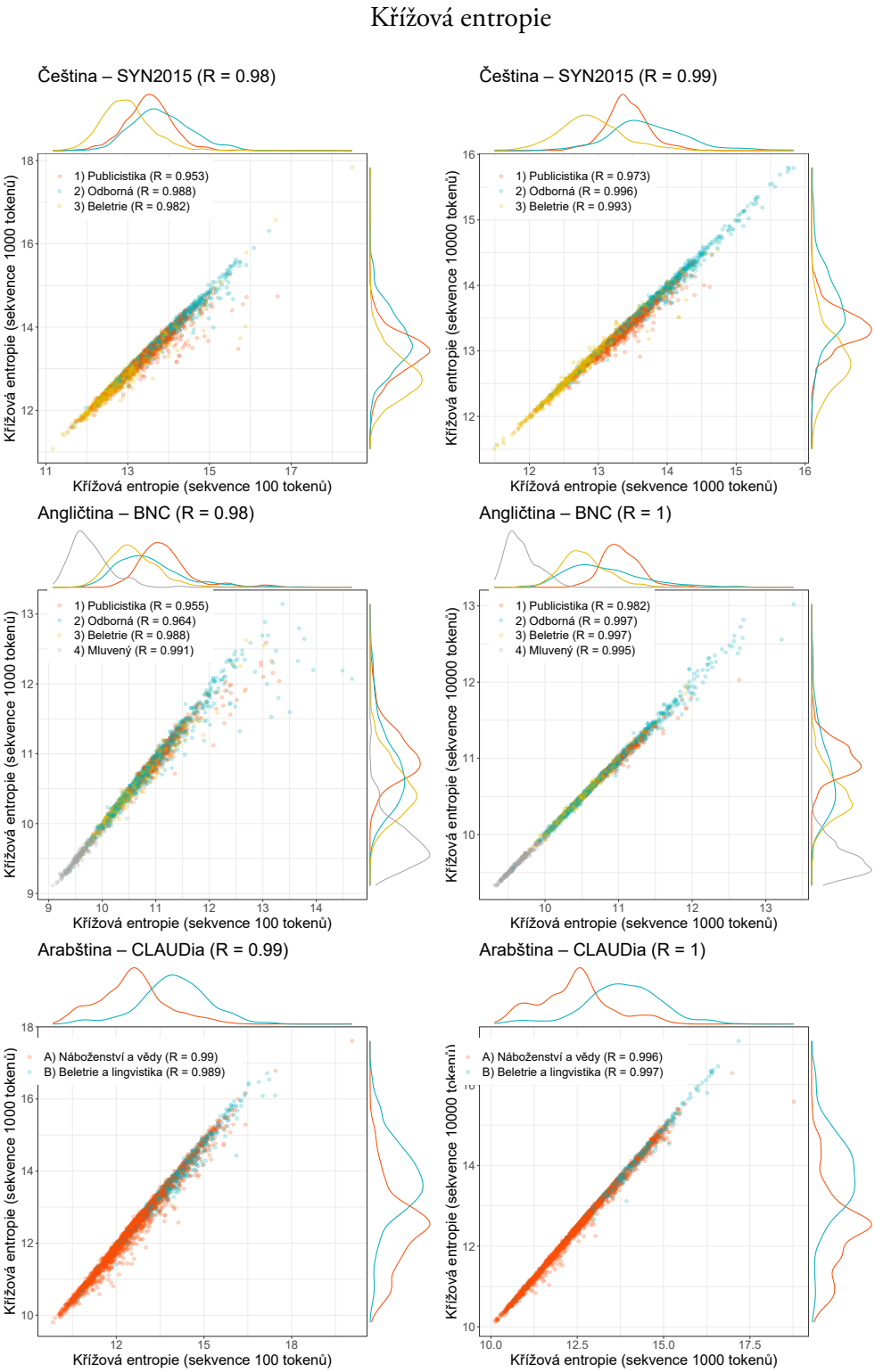
Obrázek 5.4: Korelace převrácené pravděpodobnosti opakování v delším a kratším okně.



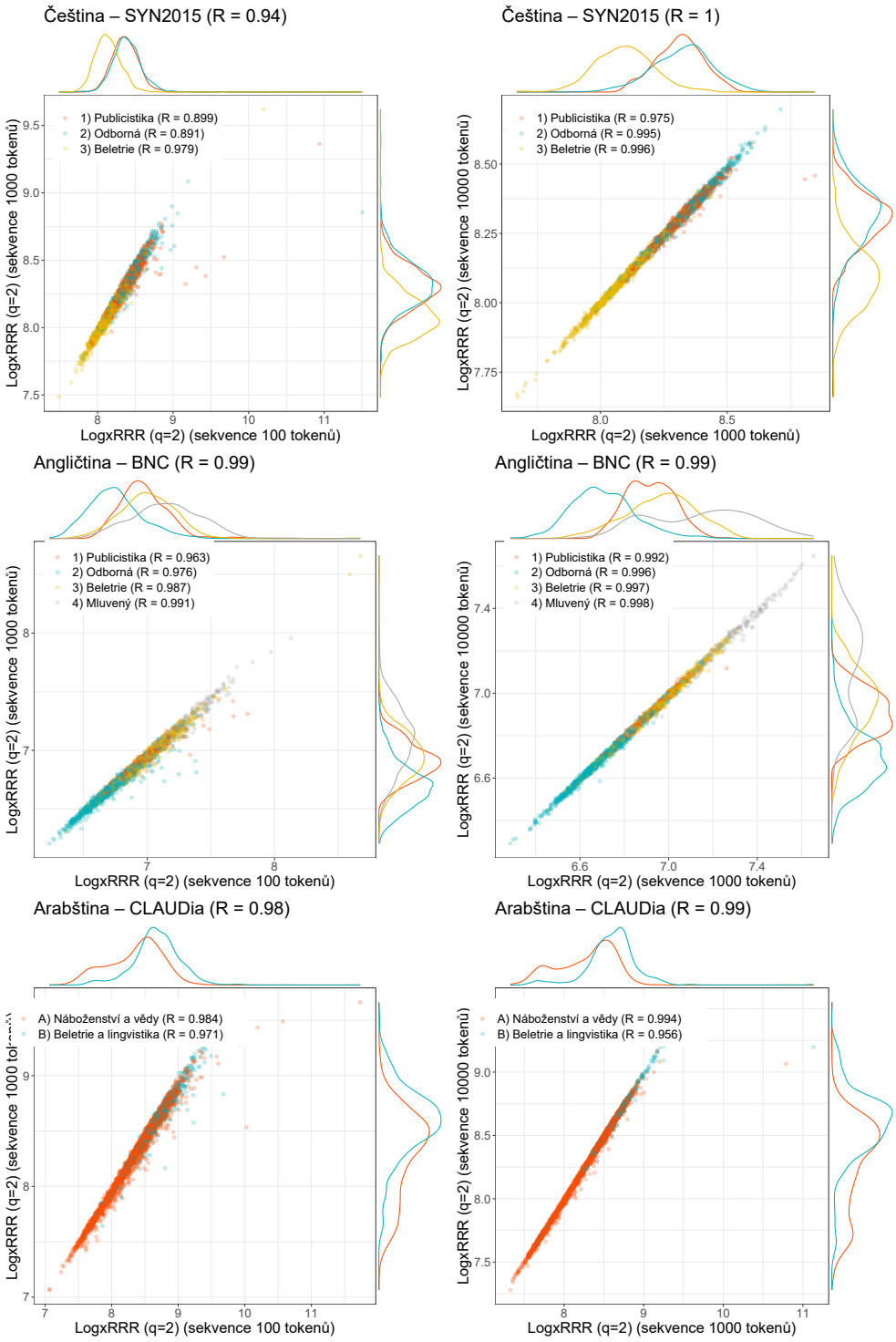
Obrázek 5.5: Korelace Hillova čísla (q = 3) v delším a kratším okně.



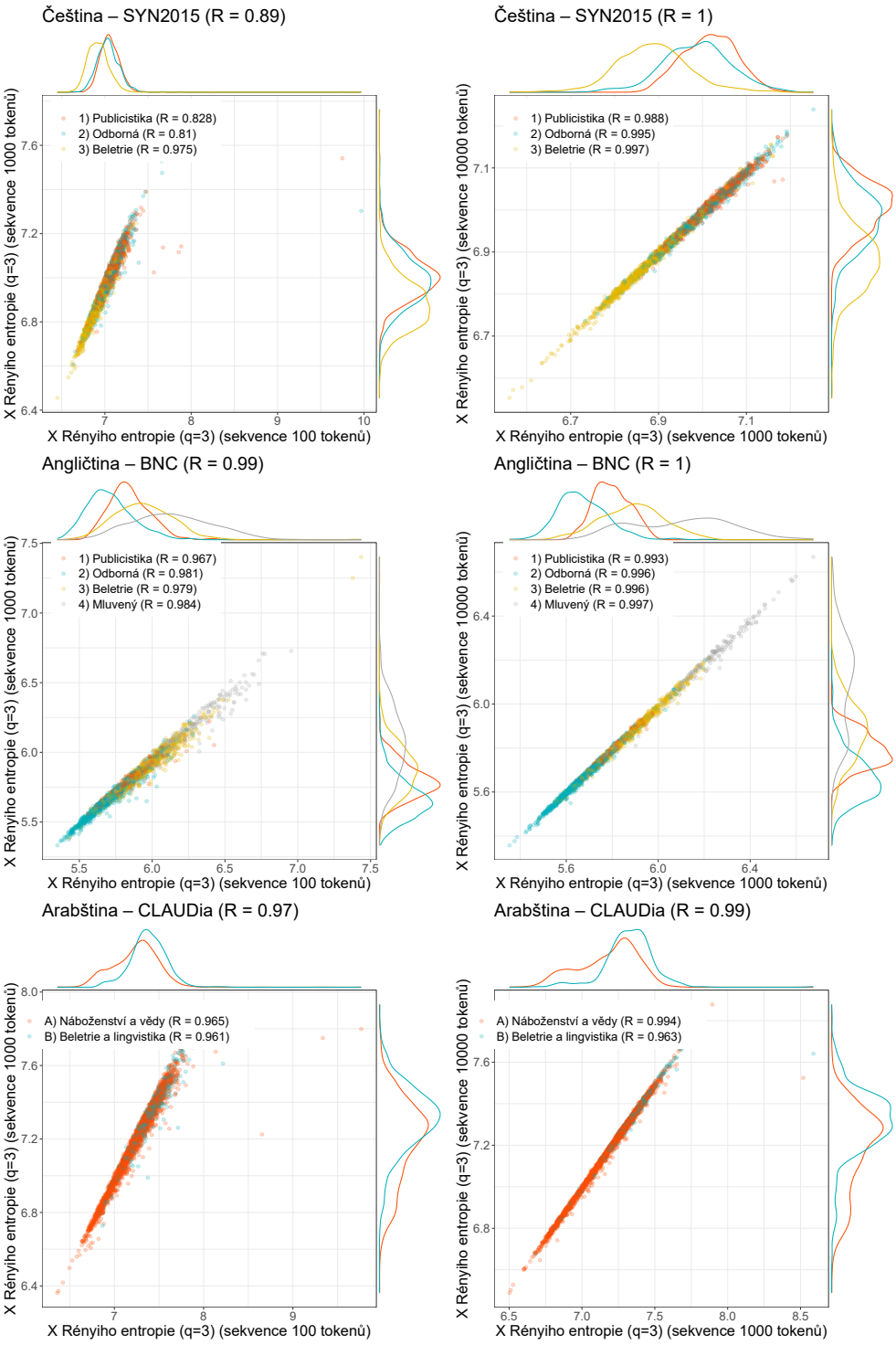
Obrázek 5.6: Korelace logaritmu křížového počtu typů v delším a kratším okně.



Obrázek 5.7: Korelace křížové Shannonovy entropie v delším a kratším okně.



Obrázek 5.8: Korelace logaritmu křížové převrácené pravděpodobnosti opakování v delším a kratším okně.



Obrázek 5.9: Korelace křížové Rényiho entropie (q = 3) v delším a kratším okně.

Kapitola 6

Syntéza

6.1 Systematizace metrik a indexů lexikální diverzity

V této knize jsem udělal všechno pro to, abychom metriky nechápali samostatně a izolovaně, aby z historicky daného chaosu povstal určitý řád, abychom byli schopní je rozdělit podle jejich vlastností. Přesto nebude na škodu si trochu utřídit myšlenky.

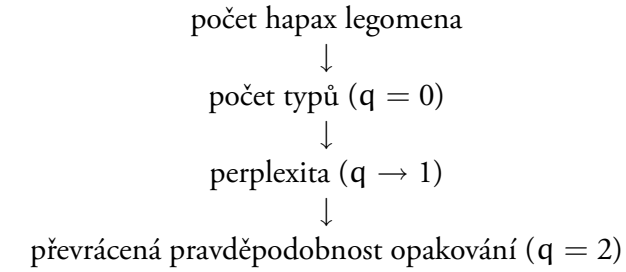
V této podkapitole identifikujeme pět proměnných, které společně dokáží charakterizovat metriky představené v této práci, odlišit jednu od všech ostatních. Jsou vybrány tak, aby podle nich bylo možno definovat opravdu co nejširší spektrum metrik, tedy i ty, jež nepovažuji za příliš šťastné, byť se historicky používaly. Tyto proměnné už byly podrobněji představeny v předchozích kapitolách, nyní jde tedy pouze o určité shrnutí.

Některé proměnné jsou kategorické a mají třeba i jen dvě hodnoty, ovšem některé mohou nabývat libovolného množství hodnot. Představme si pozici každé metriky v tabulce o pěti dimenzích — každá následující sekce je jednou dimenzí lexikální diverzity.

6.1.1 Metrika — vliv málo frekventovaných slov

Distribuce frekvencí slov v textu je přirozeně velmi nevyvážená a pokud sledujeme v naší metrice jednoduše pouhý počet typů, ztrácíme obrovské množství informace o tom, jak diverzifikované jsou nejčetnější typy. Přitom právě tato diverzita pro nás může být mnohem zajímavější než počet málo frekventovaných typů. Jindy nás naopak zajímají právě málo časté typy, neboť reflektují dynamiku slovní zásoby. Je tedy nejen možno, ale i záhodno metriku vyladit pro naše potřeby i v tomto smyslu. Tuto proměnnou můžeme chápat kategoricky, dejme tomu podle schématu v obrázku 6.1, kde čím níž jdeme, tím menší vliv mají méně frekventovaná slova.

Ovšem taky je možné sofistikovaněji použít Hillovo kontinuum, kde čím větší parametr q, tím větší mají na metriku vliv slova s vyšší frekvencí (kapitola 1.5).



Obrázek 6.1: Systematizace normování metrik podle délky.

Alternativně bychom se mohli jednoduše omezit na nějakou konkrétní frekvenční vrstvu — například zkoumat pouze hapax či dis legomena, slova, která se v textu vyskytují právě jednou nebo dvakrát. A také je samozřejmě možné tento princip s Hillovým kontinuem zkombinovat, není důvod, proč bychom například nemohli měřit perplexitu (tedy $q \to 1$) slov s frekvencí vyšší než tři.

6.1.2 Metrika — porovnávání s referenčním korpusem

Tedy v souladu s terminologií, kterou používám, jde o odlišení „křížových“ metrik, jejichž hodnota je odvislá nejen od měřeného textu, ale i od referenčního korpusu. Klasicky se používají pouze křížová entropie a křížová perplexita (1.6.1), nicméně není důvod, proč bychom se prizmatem referenčního korpusu nemohli podívat na jakoukoli metriku — tedy na křížovou variantu celého Hillova kontinua a Rényiho entropie, jako to činím v této knize. V celé práci chápu „křížovost“ jako binární kategorickou proměnnou, ovšem určitě si dokážeme představit nějaké kontinuum, kdy k referenčnímu korpusu přihlížíme „jen tak trochu“. Nebylo však dosud definováno.

6.1.3 Škálování

Škálování je také jako kategorická proměnná. Naše metrika může být škálována lineárně s počtem různých slovních tvarů, nebo logaritmicky. Způsobů škálování si sice umím představit mnoho, ovšem právě tyto dva dávají smysl podle toho, k čemu se chceme vztahovat: lineárně pojatá metrika je dobře pochopitelná i pro laiky a naplňuje dobře obecnou představu o diverzitě, logaritmovaná verze té samé metriky naopak škáluje s komplexitou, jak ji známe od Kolmogorova a s logaritmickou povahou vnímání našich smyslů. Jediným zástupcem logaritmované metriky, který se tradičně používá, je entropie, respektive její shannonovský odhad (kapitola 1.4). Ovšem není důvod, proč bychom za tímto účelem nemohli zlogaritmovat i všechny ostatní metriky z Hillova kontinua, čímž bychom dostali Rényiho entropie.

Podivuhodné je, že křížová entropie a vůbec ostatní metriky založené na Rényiho křížových entropiích (1.6.1) škálují s počtem typů lineárně, obdobně i délka slov (1.7).

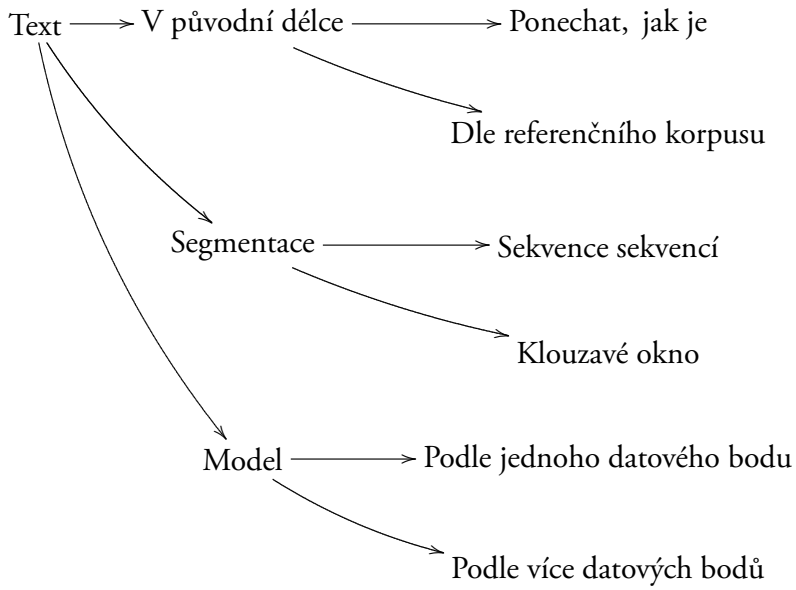
V této knize, s výjimkou kapitoly 3, škálování nevariujeme, naopak, ona kapitola slouží k tomu, abychom zjistili, jak metriky převést na stejné škálování, aby se s nimi dalo lépe počítat dál.

Na základu, při kterém logaritmujeme, nesejde. Pokud ale za základ logaritmu použijeme číslo dvě, budou výsledky interpretovatelné z pohledu teorie informace a výslednou jednotkou budou pro Shannonovu entropii bity a efektivní bity pro všechny ostatní Rényiho entropie. Společnou jednotkou metrik s lineárním škálováním je efektivní počet typů.

Pro jistotu ještě připomínám, že rozdílné škálování (pokud se neprovede nějak chybně) nezmění pořadí textů, ovšem pokud z tohoto čísla děláme průměr (například při normování pomocí metody klouzavého průměru), pak se pořadí těchto průměrů změnit může (aritmetický průměr zlogaritmované metriky je prakticky geometrický průměr metriky nezlogaritmované).

6.1.4 Metoda normalizace délky

V této práci představuji šest způsobů, jak se vyrovnat s problémem délky textu. Na obrázku 6.2 vidíme strom, který těchto šest způsobů pomáhá trochu utřídit.



Obrázek 6.2: Systematizace normování metrik podle délky.

Nejjednodušší způsob je nedělat nic — tedy text ponechat v původní délce tak, jak je — což ovšem bude dávat smysl jen s metrikami na délce textu inherentně nezávislými, tedy prakticky nikdy. Nebo také lze text ponechaný v původní délce porovnat s referenčním korpusem jako v kapitole 2.2.2.

Další možností je rozdělit text na větší množství sekvencí stejné délky, metriku změřit na nich a následně výsledky zprůměrovat či s nimi pracovat jinak. I tady se ovšem nabízejí dvě možnosti: dané sekvence mohou následovat jedna po druhé, nebo se mohou překrývat (2.2.1, metoda klouzavého okna, tuto možnost vřele doporučuji). Samotná délka sekvencí (velikost okna) je tady další důležitou proměnnou, se kterou můžeme dále pracovat, neboť metriky měřené na oknech různé velikosti se mohou zajímavě lišit.

Poslední možností (respektive poslední možností představenou v téhle publikaci, jinak možnosti jsou samozřejmě neomezené) je vybrat si nějaký model toho, jak daná metrika závisí na délce textu, onen model nafitovat na daný text a následně pracovat s parametry tohoto modelu (kapitola 2.2.3). Osobně mi tyto dvě metody nepřijdou příliš vhodné z důvodů popsaných v kapitole 1.10.1, nicméně v literatuře se hojně používají.

6.1.5 Metoda typizace

Tato kategorická proměnná určuje metodu, podle které se pozná, že dva tokeny patří ke stejnému typu. Jak podrobně rozebírám už v kapitole 1.1, může jít o řadu subtilních rozhodnutí, nicméně zhruba je můžeme rozdělit podle toho, zda používáme texty lemmatizované či nelemmatizované (podrobněji viz celou kapitolu 4), nebo jestli použijeme nějakou sofistikovanější metodu, například nějakou nebinární metriku podobnosti (respektive rozdílnosti) typů, ať už na základě vnější podobnosti (například nějak normovanou Levenshteinovu vzdálenost, kterou popisuji v kapitole 1.8 a se kterou v průběhu studie dále pracuji), nebo podobnosti vnitřní, například sémantické.

V případě mluvených textů je pak otázka, jestli se budeme řídit pravidly jazyka psaného, nebo jestli půjdeme blíž k fonetické či dokonce akustické podobě textu.

6.2 Vzájemné korelace metrik a klastrová analýza

Pojďme se nyní podívat, jak se metriky rozdělené podle těchto pěti dimenzí chovají. Jak moc se liší výsledné hodnoty, jak spolu metriky korelují v závislosti na tom, podle jaké dimenze se liší. Podle jakých principů se metriky sdružují.

Ideální by bylo prohlédnout si přímo bodové grafy, na kterých je korelace znázorněna mnohem přehledněji a intuitivněji, než když použijeme nějaké korelační metriky. Docela obstojné množství bodových grafů znázorňujících, jak spolu korelují různé metody měření lexikální diverzity, jste si prohlédli v předchozích kapitolách.

Ovšem mohl jsem jich ukázat jen omezené množství, už takhle je v knížce skoro víc obrázků než textu. Pokud byste toužili vidět jich ještě víc, připravil jsem pro vás ultimátní korelogram o několika stovkách grafů, který si můžete stáhnout na adrese `milicka.cz/habilitace.zip`.[1] Prohlížení tohoto korelogramu je asi nejlepší způsob, jak si vytrénovat intuici v tom, jak se k sobě jednotlivé metriky zhruba mají. Po vytištění s dobrým rozlišením na metrový papír se jedná i o docela estetickou záležitost.

Nicméně takový korelogram je docela nepřehledný, takže si práci s korelační maticí trochu ulehčíme pomocí klastrové analýzy, jež nám pomůže určit, která metrika souvisí s kterou, a najít v dimenzích lexikální diverzity, jak jsme si je představili v předchozí podkapitole, nějaký systém.

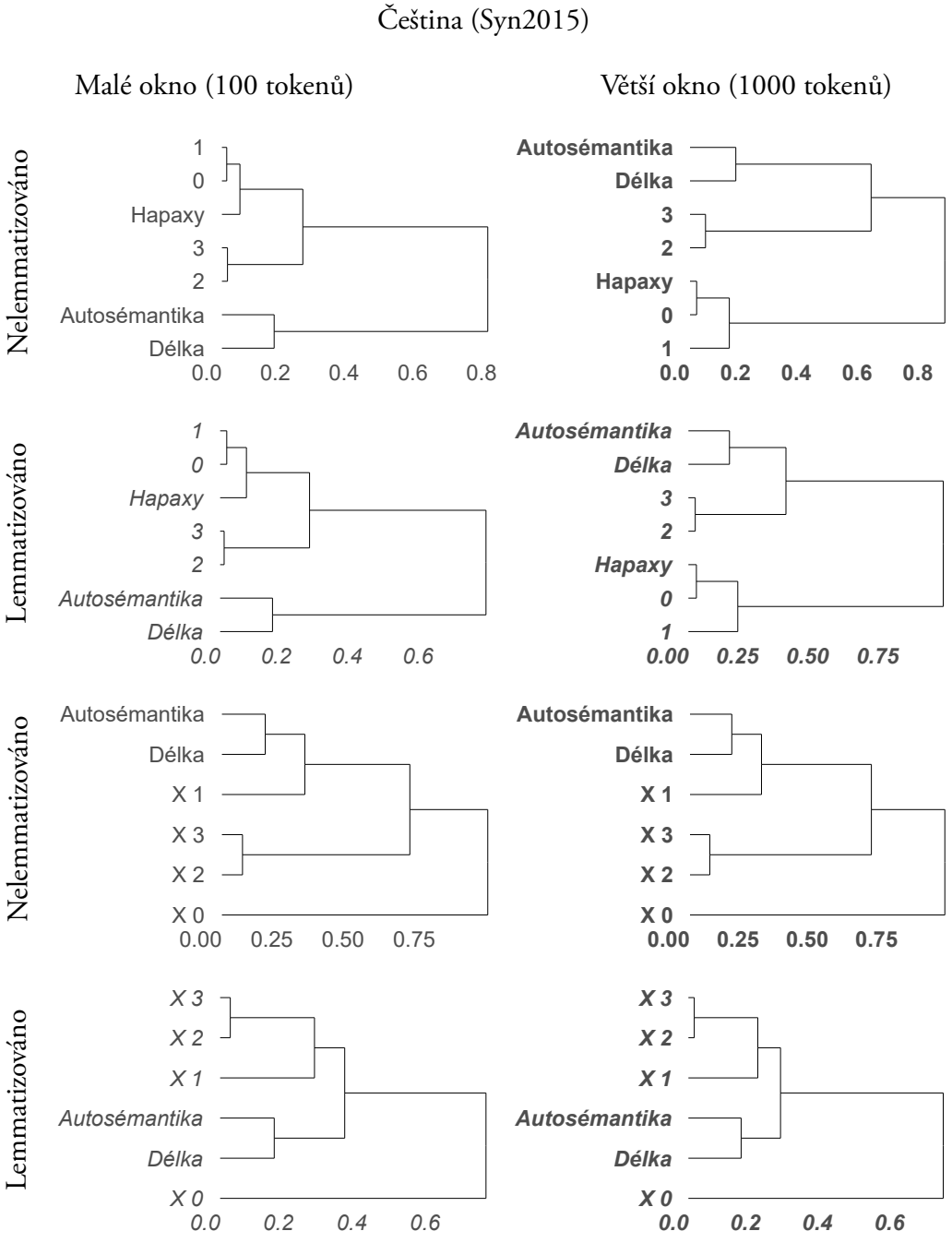
Na obrázcích 6.3–6.5 si můžete prohlédnout dendrogramy, které postupně sdružují metriky, které spolu lineárně korelují (jako metriku vzdálenosti používám Pearsonův korelační koeficient odečtený od jedničky). Je jich pro každý jazyk osm, neboť jsem je rozdělil podle tří dimenzí: velikosti okna, lemmatizace a křížovosti. Výsledky jsou naměřeny na deseti tisíci vzorcích o stovce tokenů (malé okno) a tisíci tokenů (velké okno), tedy velikostech, které se docela hodí pro reálný výzkum na povídkách či slohovkách a mezi kterými je dostatečně velký rozdíl na to, aby vzájemně kontrastovaly. Metrika změřená na velkém okně je v dendrogramu znázorněna tučným řezem písma. Kurzívou jsou znázorněny metriky změřené na lemmatizovaném textu.[2] Abychom spolu křížové a nekřížové metriky mohli srovnat, místo křížových Hillových čísel používáme jejich log. transformaci, tedy Rényiho křížové entropie. Křížové verze označuji písmenem X. Jednotlivé metriky jsou pro přehlednost označeny pouze hodnotou parametru q.[3]

Je velmi potěšující, že nezávisle na jazyce, lemmatizaci a délce okna vidíme velmi podobné výsledky. Metriky se hezky sdružují podle hodnoty parametru q, přičemž dělicí linie leží mezi perplexitou ($q = 1$) a převrácenou pravděpodobností opakování ($q = 2$). U křížových entropií se trochu vymyká čeština, kde se zlogaritmovaný křížový počet typů odmítá kamarádit s ostatními, nicméně tato metrika se nikdy nepoužívala a vlastně mě překvapuje, že se jinak chová docela ukázněně.

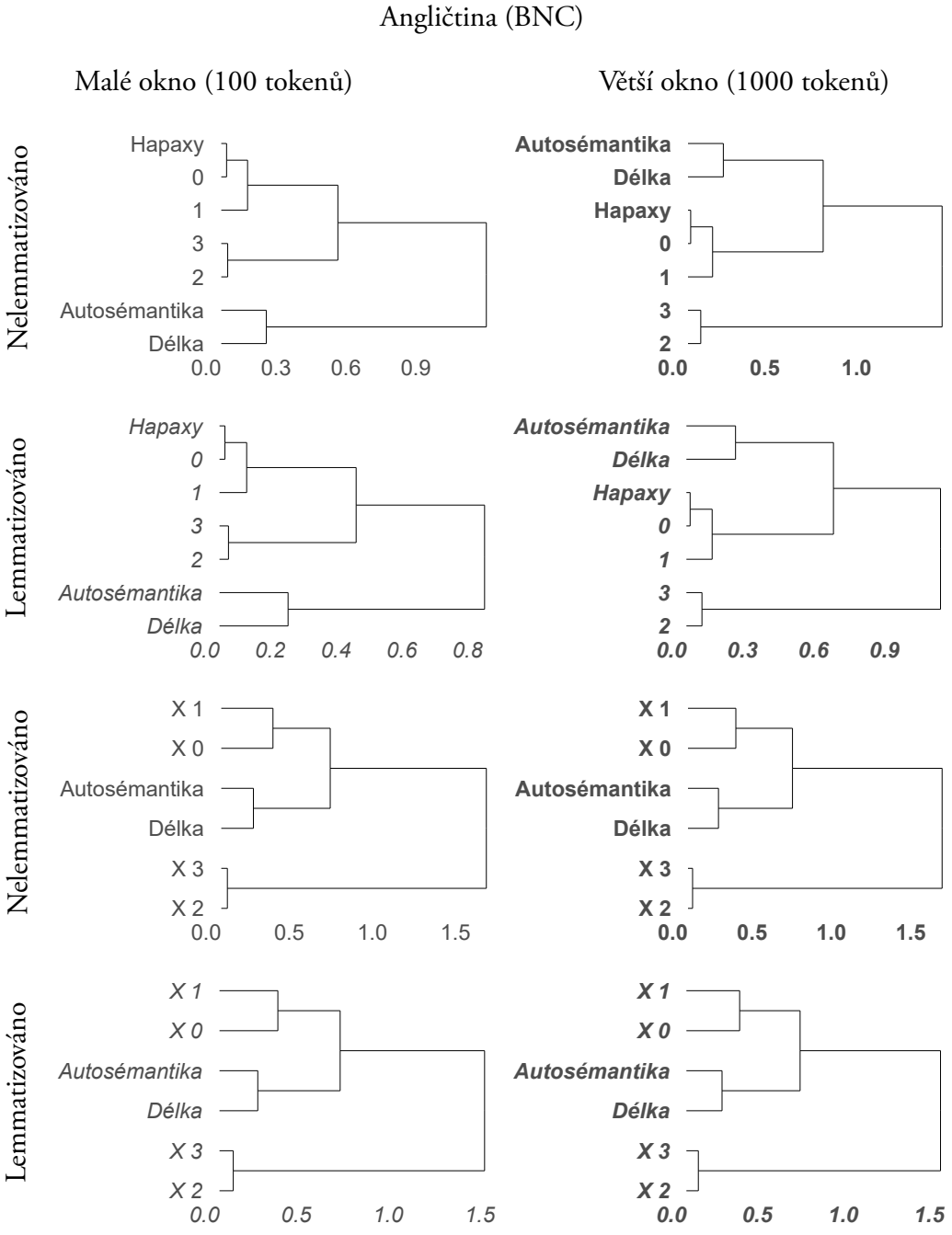
[1] Ostatně jako i veškeré programové vybavení a data, která byla při této studii použita, s výjimkou těch, u kterým mi copyright sdílení neumožňuje.

[2] Musím se přiznat, že toto typograficky nehezké řešení vzniklo kvůli tomu, že se mi několik hodin nedařilo přesvědčit R, že má lemmatizované texty obarvit červeně a nelemmatizované modře. Nicméně vedlejším produktem je, že grafy na obrázcích 6.6–6.8 jsou vhodné i pro daltoniky. Tímto bych se jim ovšem rád omluvil, v této knize jsou to totiž první a poslední grafy, na kterých něco rozumného uvidí. Pokud se mezi čtenáři a čtenářkami daltonik či daltonička najde, může si upravit barevnou škálu a grafy si vygenerovat znova pomocí přiložených R skriptů. Popřípadě mě můžete najít a přimět, abych je pro vás vygeneroval.

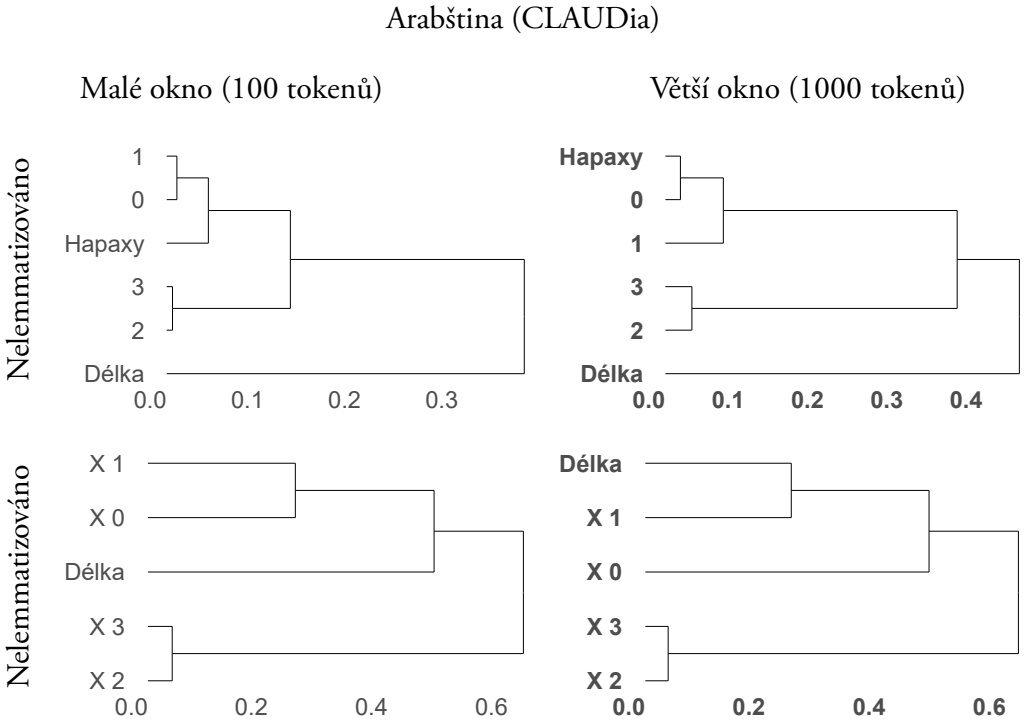
[3] Pokud jste přečetli celou knihu až sem, předpokládám, že si klasická pojmenování jednotlivých metrik podle parametru q již pamatujete, nicméně pro jistotu připomínám, že 0 odpovídá počtu typů, 1 perplexitě, 2 převrácené hodnotě pravděpodobnosti opakování a 3 nemá pojmenování.



Obrázek 6.3: Dendrogram popisující jak souvisejí metriky za různých okolností v češtině.

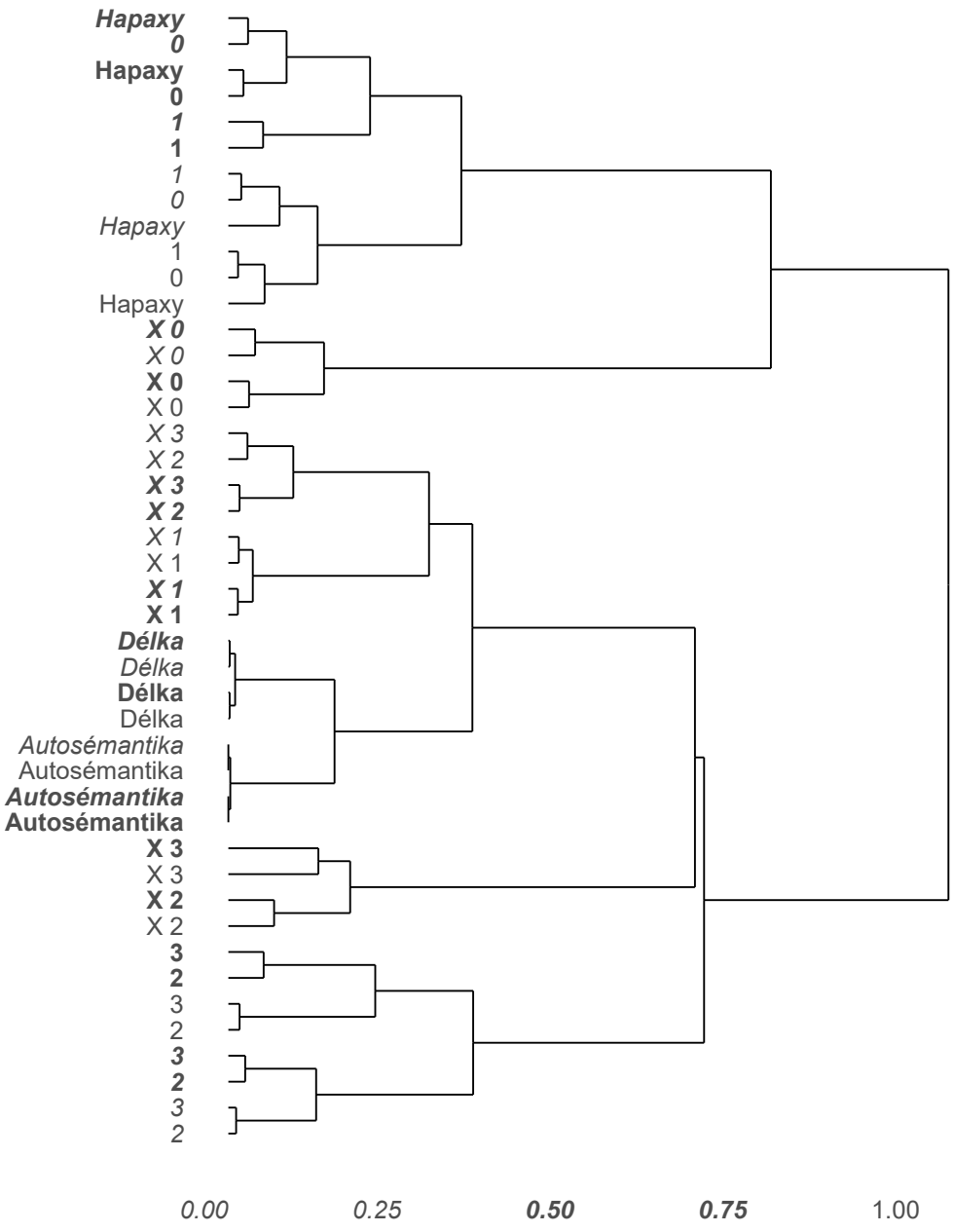


Obrázek 6.4: Dendrogram popisující jak souvisejí metriky za různých okolností v angličtině.



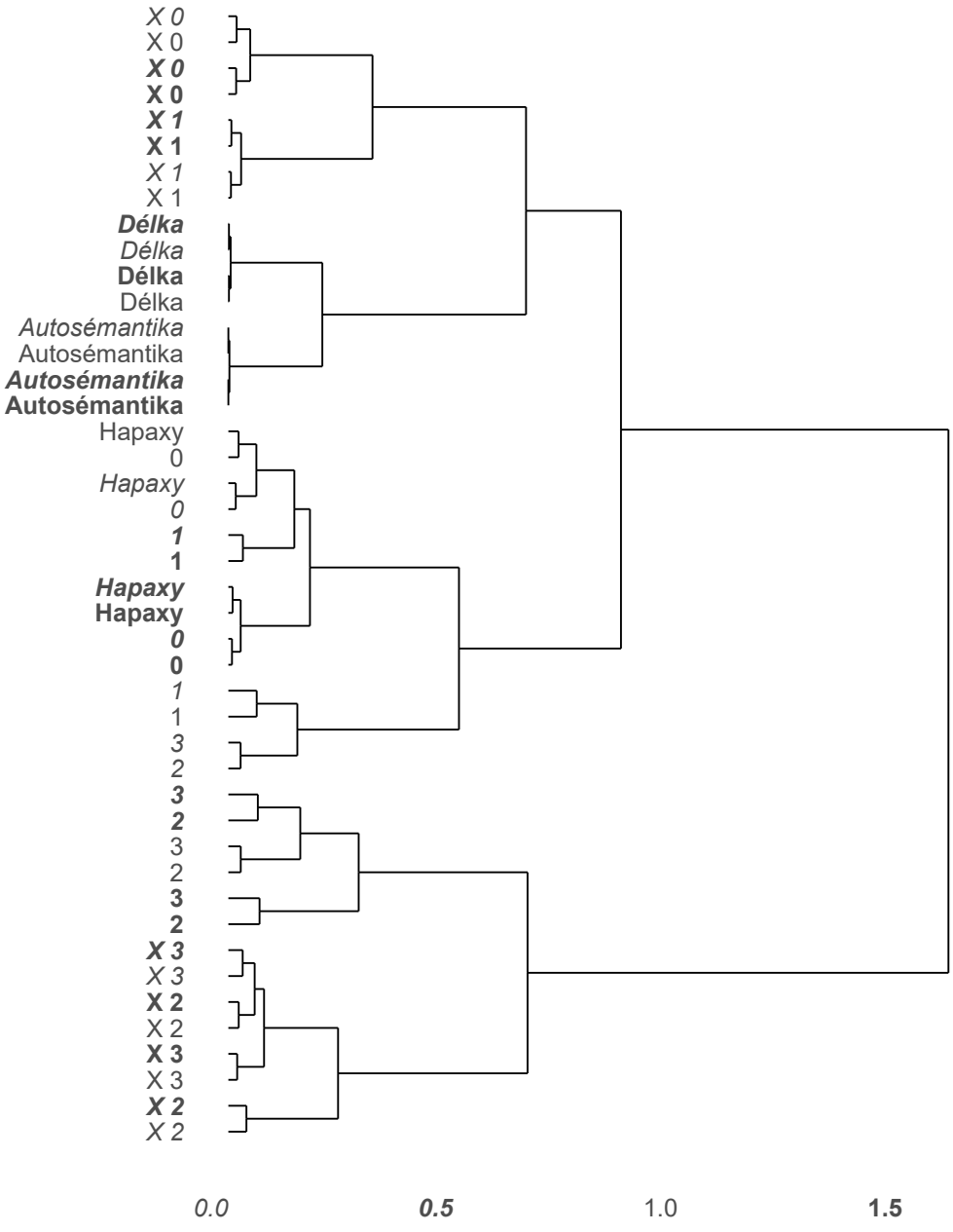
Obrázek 6.5: Dendrogram popisující jak souvisejí metriky za různých okolností v arabštině.

Čeština (Syn2015)



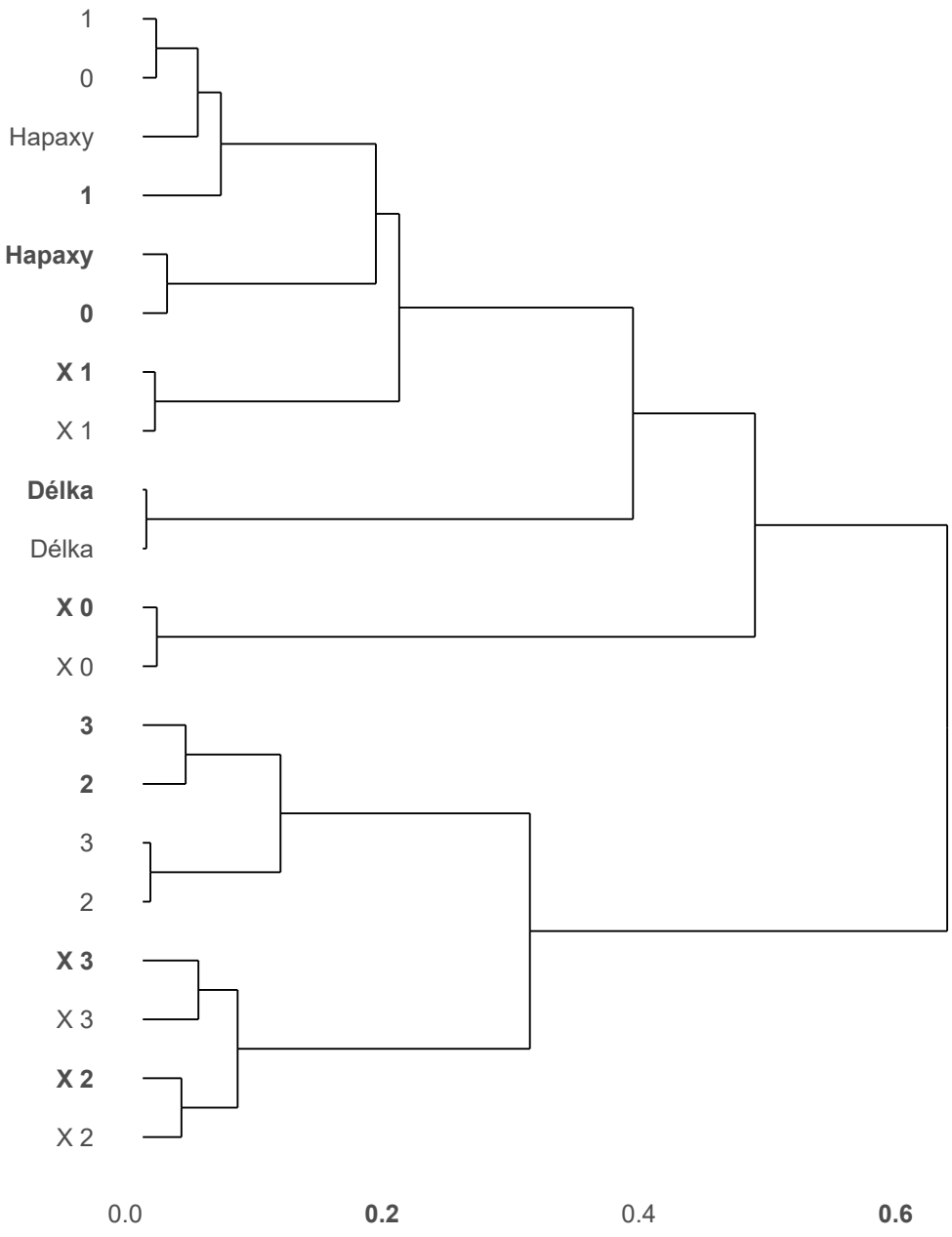
Obrázek 6.6: Dendrogram popisující jak souvisejí metriky v češtině.

Angličtina (BNC)



Obrázek 6.7: Dendrogram popisující jak souvisejí metriky v angličtině.

Arabština (CLAUDia)



Obrázek 6.8: Dendrogram popisující jak souvisejí metriky v arabštině.

Kromě metrik na Hillově kontinuu jsem do dendrogramu zahrnul ještě průměrnou délku slova a podíl autosémantik, které se vydělují a tvoří samostatný klastr. Je příjemné, že tento klastr nejblíže koreluje buď přímo s křížovou Shannonovou entropií, nebo alespoň klastrem, který ji obsahuje, což je vztah, který jsme si odvodili v kapitole 1.7 z teorie efektivního kódování. K Hillovým číslům se průměrná délka slova nejčastěji připojuje až naposled, ovšem situace není tak jednoznačná jako u křížových entropií. Počet hapaxů, přesně podle očekávání, klastruje s počtem typů ($q = 0$).

Na obrázcích 6.6–6.8 jsou pak dendrogramy, které sdružují podle lineární korelace kompletně všechny použité metriky a metody měření.

Na angličtině se metriky chovají hezky systematicky: Nejblíže k sobě mají metody, které se liší pouze lemmatizací, popřípadě velikostí okna (u metrik s nižší hodnotou parametru q tento závod obvykle vyhrává lemmatizace, u těch s vyšší naopak délka okna). Podobný efekt můžeme vidět i na češtině, ovšem vzhledem k bohatší morfologii je rozdíl v lemmatizaci patrnější a hodnoty měřené na lemmatizovaných textech se od svých nelemmatizovaných kamarádů občas docela zatoulají, zejména pro vyšší parametry q, což je v souladu s kapitolou 4. Arabský korpus bohužel nemáme lemmatizovaný a morfologicky značkovaný, tedy odpadá nám jedna dimenze (a také nejsme schopní určit podíl autosémantik). Tím více vynikne, že se sdružují metody, které se liší pouze velikostí okna.

V českém dendrogramu se přímo ukázkově vydělily křížové metriky, které jsou v jednom klastru (s výjimkou log. křížového počtu typů ($q = 0$), o jehož podivnostech již byla řeč, ovšem společně s podílem autosémantik a průměrnou délkou slova). Tento výsledek vcelku souhlasí s tím, čeho jsme byli svědky v předchozích kapitolách, totiž že křížové metriky se chovají odlišně od svých nekřížových variant a že vlastně měří něco tak trochu jiného. Ovšem v angličtině není situace tak jednoznačná, neboť klastr s křížovými metrikami pro nižší q se spřáhl klastrem obsahujícím jejich nekřížové varianty a stejně tak učinil i klastr s křížovými metrikami pro vyšší q. Podobné výsledky vidíme i na arabštině.

V další podkapitole se tedy zaměříme mimo jiné i na tuto otázku: dává smysl oddělit Hillovo kontinuum od Rényiho křížových entropií?

6.3 Empirické určení dimenzí a jejich redukce

V kapitole 6.1 jsem rozdělil metriky a metodiky měření lexikální diverzity do pěti dimenzí. Ty jsem ovšem určil racionálně, bez ohledu na empirii. Pojďme nyní zkusit přesně opačný postup: do jakých dimenzí se nám naše metriky roztřídí, když se budeme dívat *pouze* na data, která produkují. Ideálním nástrojem pro tuto práci je analýza hlavních komponent — PCA.

Tato podkapitola nám také pomůže odpovědět na otázku, kterou ohledně metrik

lexikální diverzity vznáší Altmann s Wimmerem, řka: „If they do not say the same, how many indices do we need in order to obtain a complete picture of vocabulary richness?“ (Wimmer – Altmann, 1999)

Pojďme se napřed podívat na dvě hlavní komponenty Hillova kontinua na obrázku 6.9, které vznikly analýzou stejných dat, jaká byla použita v předchozí podkapitole. V levém sloupci vždy vidíme, kde našly podle těchto dimenzí své místo jednotlivé vzorky vytažené z korpusu, vpravo pak nalezneme pozici eigenvektorů — pro přehlednost naznačuji pouze střed, ke kterému se vztahují (velkým černým křížem), a koncové body, přičemž stejně jako v předchozí kapitole jsou jednotlivé metriky označeny číslicí podle svého parametru q. Větší kroužek značí větší vzorek, menší kolečko menší, červená barva metriky změřené na slovních tvarech, zatímco modrá metriky změřené na lemmatizovaném textu.

Zejména nás zaujme první komponenta, na které se podílejí všechny metriky téměř stejnou měrou, přičemž všechny ukazují stejným směrem. Jedná se o jakousi destilovanou lexikální diverzitu. Tuto komponentu bychom mohli zavést jako shrnující metriku lexikální diverzity, pokud bychom byli ochotní obětovat interpretabilitu a další informace, které lze vydedukovat z dalších komponent.

Druhou komponentu můžeme snadno interpretovat jako vliv málo frekventovaných slov, neboť proti sobě jsou metriky s nízkým a vysokým parametrem q, vlastně se nám metriky podle tohoto parametru seřadily.

Přestože PCA nemá jako primární cíl klastrovat vzorky podle jednotlivých textových typů a modalit, je příjemné vidět, že se podle těchto dvou nejdůležitějších dimenzí docela hezky rozdělily — publicistika v češtině i angličtině má prostě vysokou surovou lexikální diverzitu a odborný text lze odlišit od beletrie na základě druhé komponenty.

Další dvě hlavní komponenty Hillova kontinua máme na obrázku 6.10, třetí komponenta rozděluje prostor na červené a modré, tedy podle lemmatizace zdrojových textů. To tedy platí pro češtinu a angličtinu, kde lemmatizace je k dispozici, v arabštině rozděluje třetí komponenta podle velikosti vzorku. Velikost vzorku je pak doménou čtvrté komponenty pro angličtinu a češtinu, čtvrtá komponenta pro arabštinu už není právě interpretovatelná a má i velmi nízkou eigenhodnotu.

Jak je vidět, empirická data se pro Hillova čísla chovají přímo ukázkově: prostor se nám rozdělil na čtyři velmi dobře interpretovatelné dimenze, přičemž ta nejdůležitější je prostě lexikální diverzitou, druhá nejdůležitější je závislá na parametru q, třetí je dána metodou typizace a čtvrtá velikostí okna.

Křížové Rényiho entropie (obrázky 6.11 a 6.12, konce eigenvektorů označuji křížkem) se chovají o poznání chaotičtěji. Klastrování podle textových typů je sice pozorovatelné, nicméně situaci znepřehledňuje množství outlierů.

Samotná „destilovaná“ křížová entropie se v češtině a arabštině opět ukazuje jako nejdůležitější dimenze a parametr q jako druhá nejdůležitější, ovšem rozdíl mezi jejich eigenhodnotami není tak velký jako v případě Hillových čísel a v angličtině jsou první

dvě komponenty dokonce prohozené. Třetí dimenzi nedokážu interpretovat v žádném z jazyků a teprve ve čtvrté se dá vypozorovat poněkud nesoustavná tendence: v češtině a angličtině čtvrtá dimenze rozděluje metody měření podle lemmatizace, v arabštině, která tuto distinkci nemá, pak podle velikosti okna. Ve všech případech je ovšem eigenhodnota této čtvrté dimenze docela malá.

Jistě jste napjatí, jaké komponenty vykrystalizují, když smícháme všechny naše oblíbené metriky dohromady, tedy Hillova čísla, Rényiho entropie, průměrnou délku slova (značeno jako trojúhelník s písmenem L), počet autosémantik (trojúhelník s písmenem A), počet hapaxů (kroužek s písmenem H). S výsledkem se seznámíme na obrázcích 6.13 a 6.14.

Podivuhodné je, že se PCA stále daří najít nějaké vlastnosti společné všem metrikám a že obsazují nejdůležitější komponentu, která se tak stává velmi obecnou metrikou lexikální diverzity. Tedy s výjimkou křížových entropií s vyšším parametrem q v angličtině, které z neznámých důvodů působí proti této obecné metrice.

Druhou komponentu v češtině zaujímá distinkce křížových a nekřížových metrik, třetí pak řazení podle parametru q, v ostatních dvou jazycích jsou tyto komponenty prohozené. Čekali bychom, že čtvrtou komponentu zaujme buď velikost okna, nebo lemmatizace, ale není tomu tak, jedná se o podivnou směs, kterou nedokážu rozumně interpretovat.

Je příjemné vidět, že v prostoru prvních dvou komponent zaujímá průměrná délka slova polohu blízkou poloze křížové Shannonovy entropie ($q = 1$).

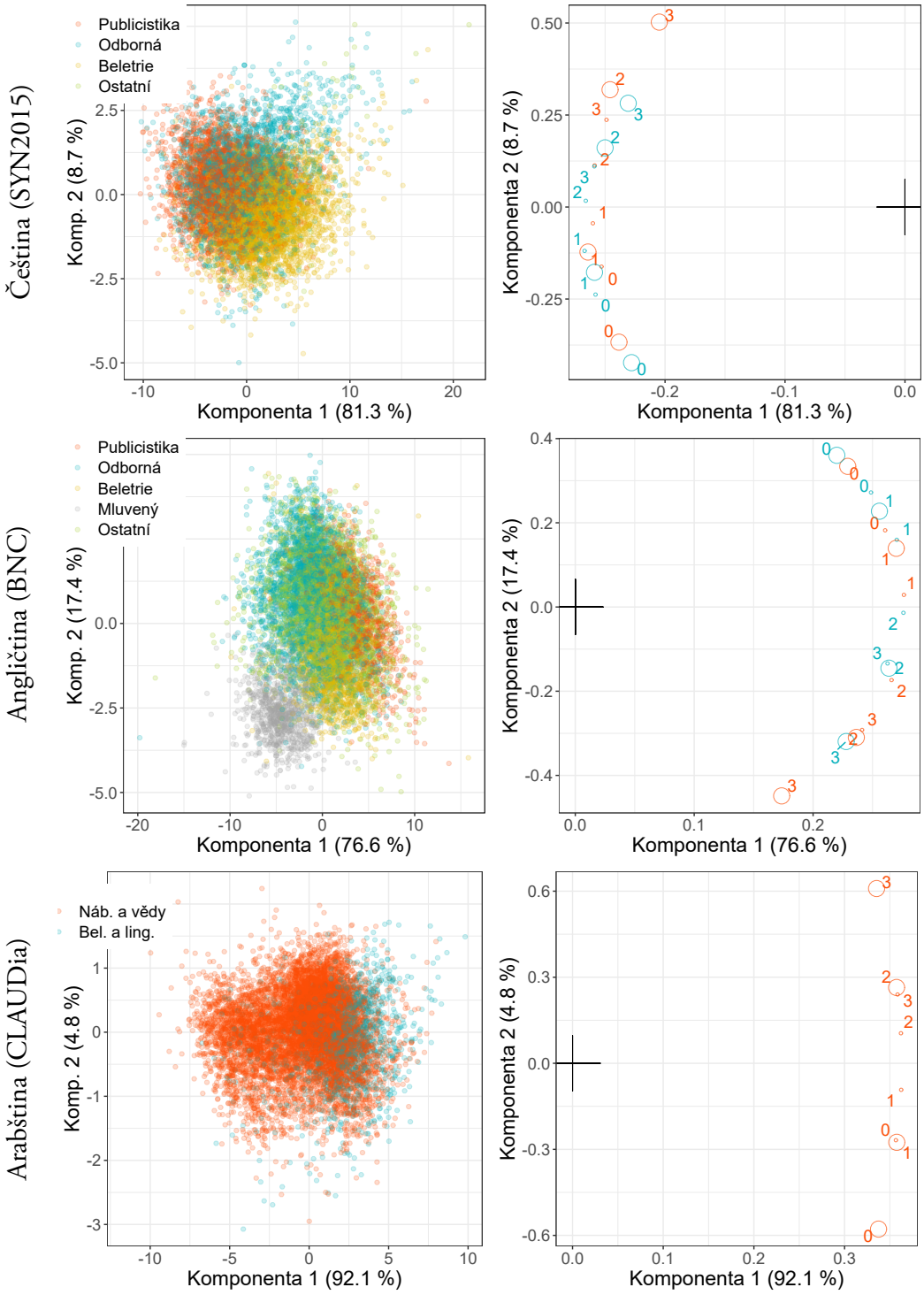
* * *

Na základě posledních dvou podkapitol bych shrnul, že všechny metriky Hillova kontinua i Rényiho entropie, průměrná délka slov a počet autosémantik ukazují obdobným směrem, byť každá z těchto metrik měří něco trochu jiného. Je to právě výběr metriky, co má největší vliv na výsledek, lemmatizace textu a délka okna jsou podružné, byť též důležité.

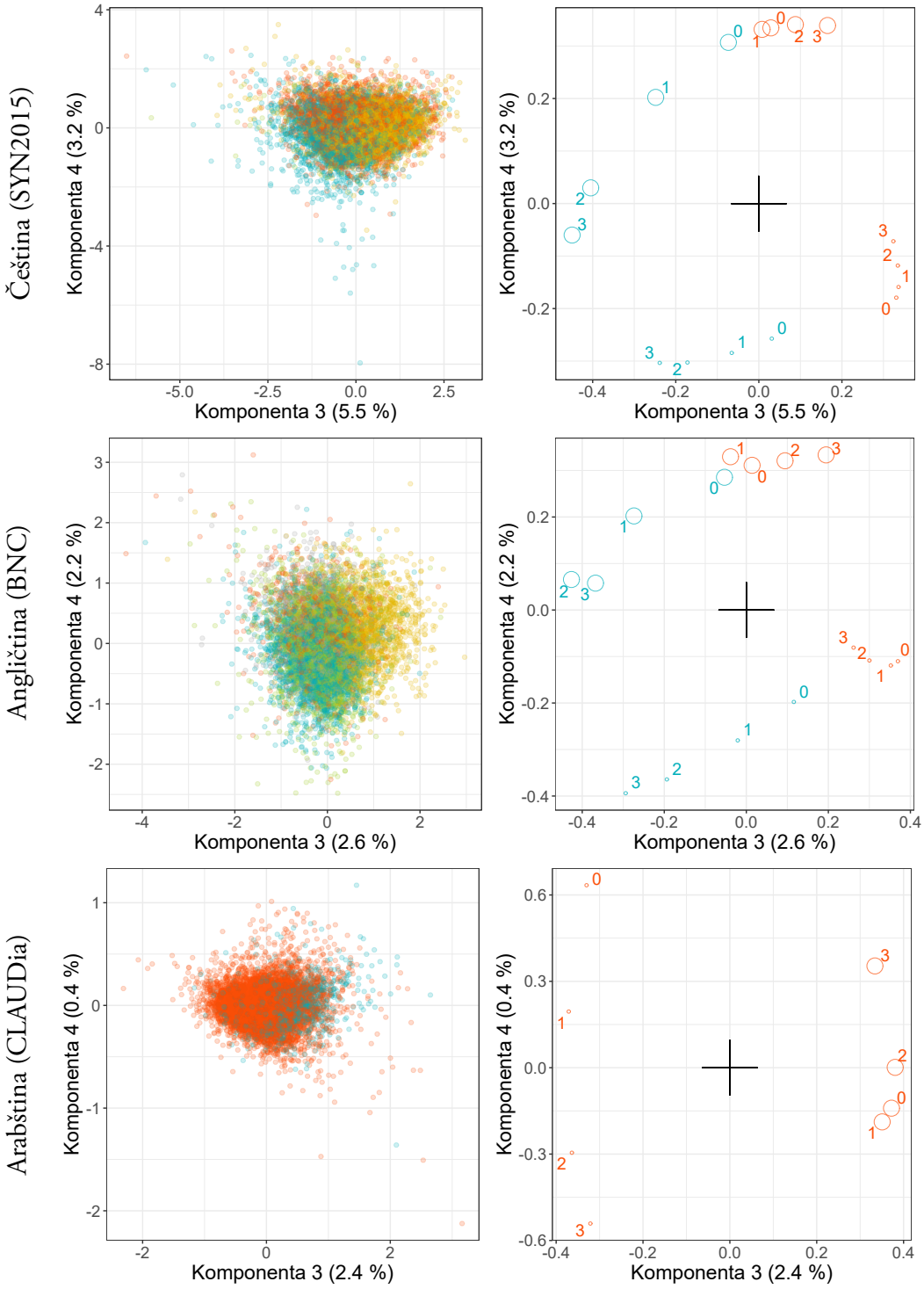
Metriky Hillova kontinua, analyzované zvlášť od ostatních metrik, chovají se příjemně systematicky a předvídatelně, mají stejnou škálu a vzájemně korelují. Přimíchají-li se k nim Rényiho entropie a další metriky, vznikne situace o poznání chaotičtější, navíc Rényiho entropie a další zmíněné metriky měří vlastně něco jiného,[4] což je nejlépe vidět na českých datech, ovšem podobný, byť ne tak silný rozdíl můžeme vidět i v ostatních jazycích.

Lexikální diverzitu, vyjádřenou Hillovým kontinuem a měřenou v efektivních typech, a lexikální křížovou entropii (vyjádřenou pomocí Rényiho entropií a měřenou

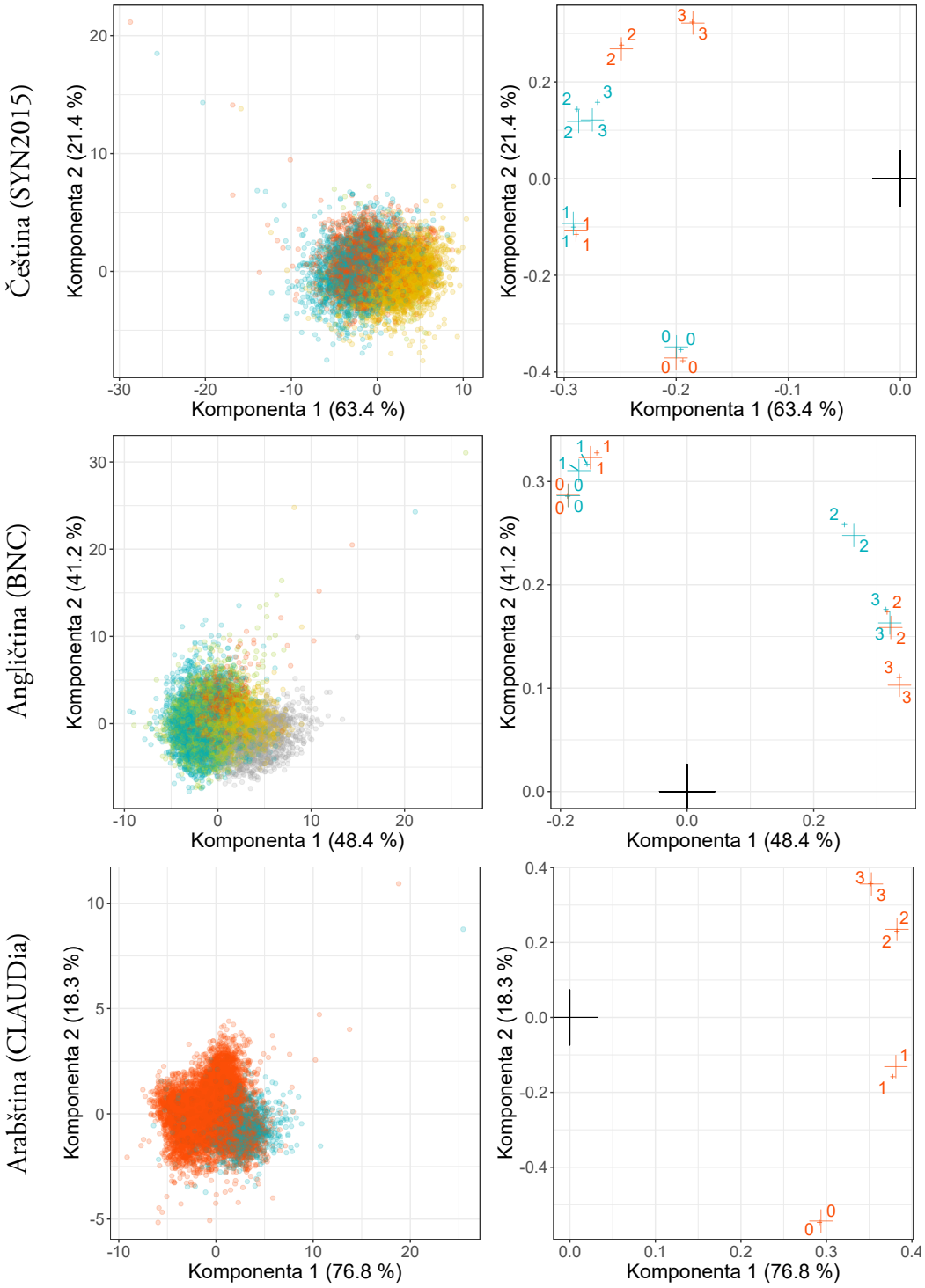
[4] Předpokládám, že má spíš blíž k lexikální sofistikovanosti (lexical sophistication) než k lexikální diverzitě, ovšem to je téma, které nechávám pro příští publikace.



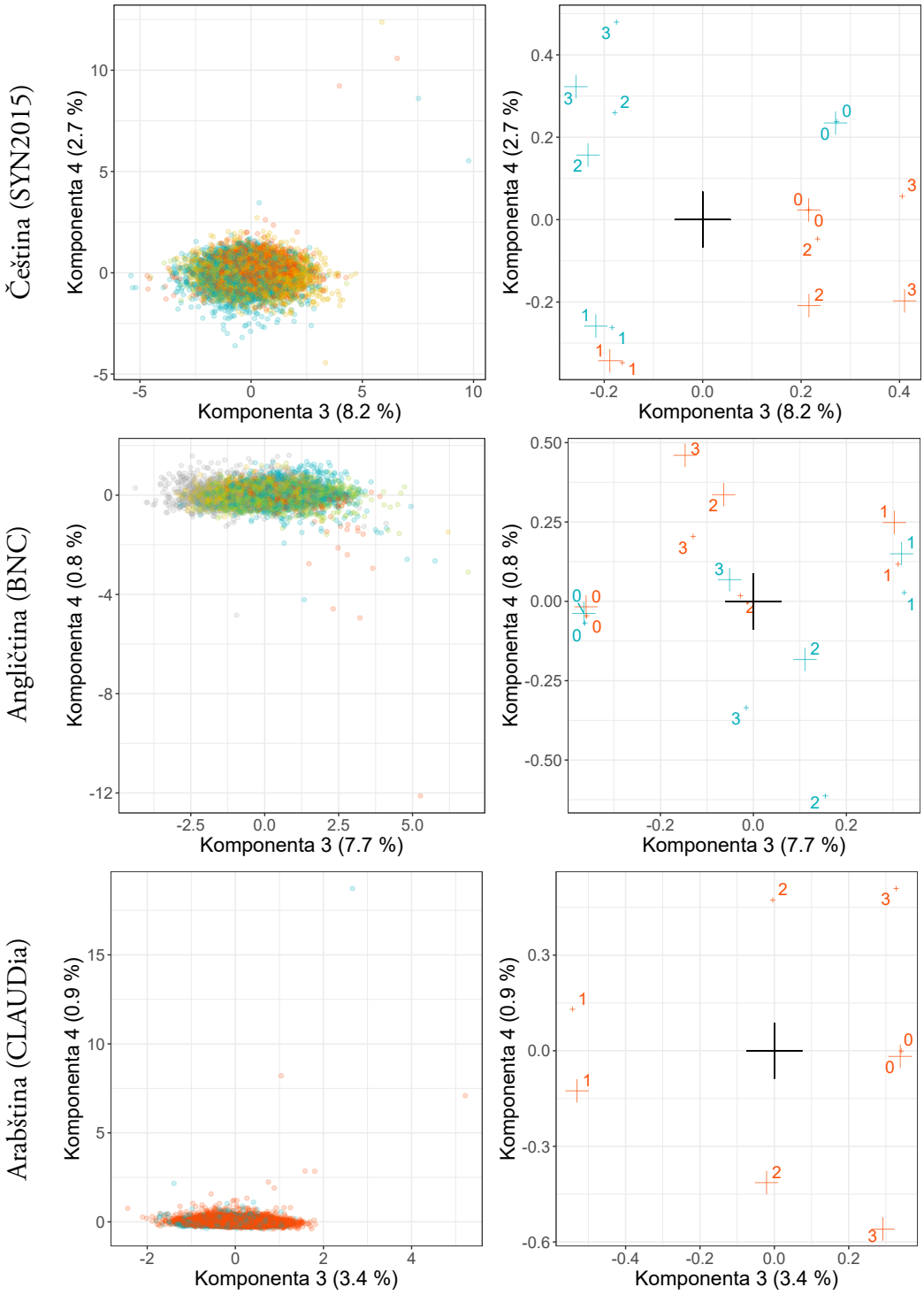
Obrázek 6.9: První dvě komponenty lexikální diverzity.



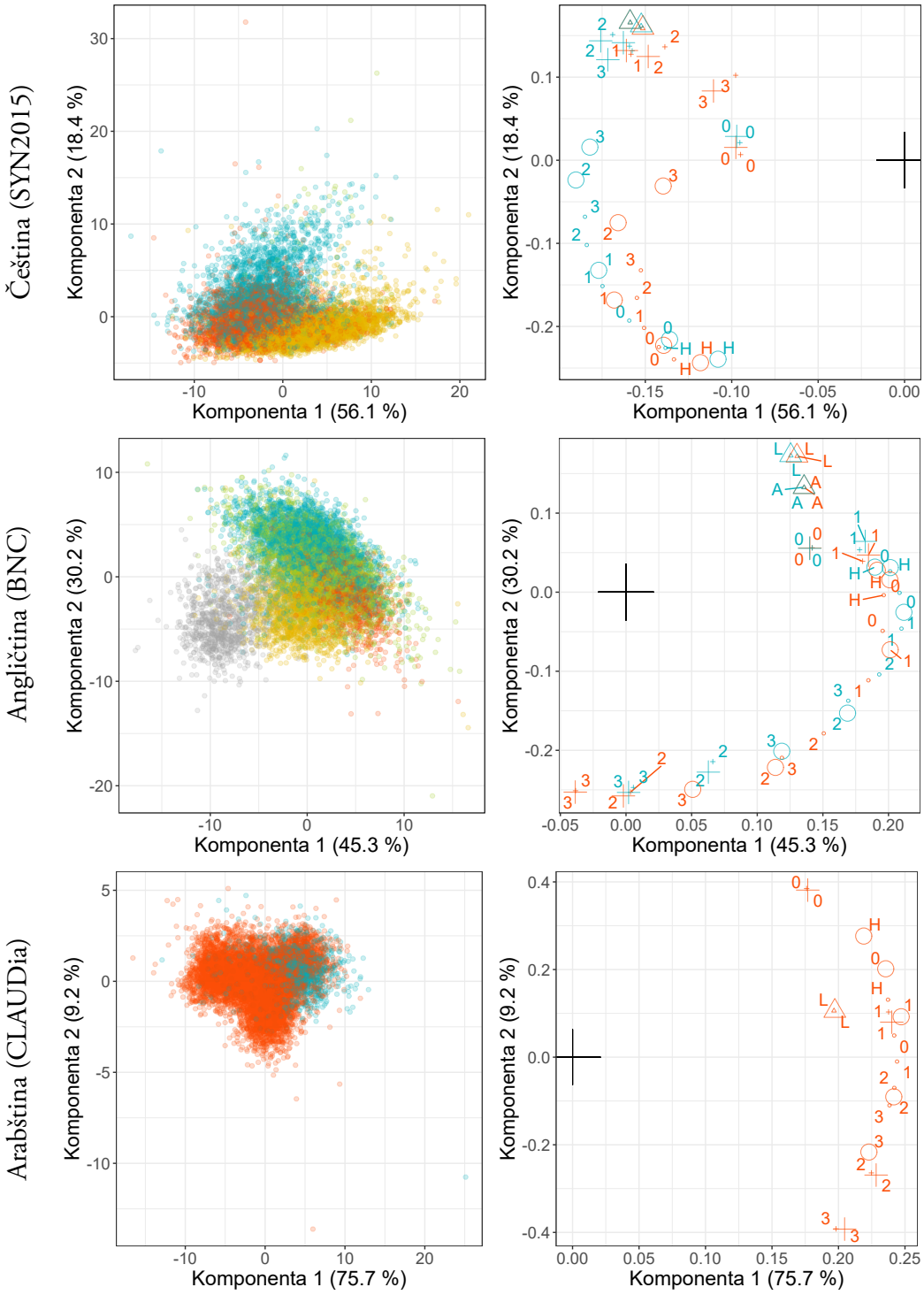
Obrázek 6.10: Další dvě komponenty lexikální diverzity.



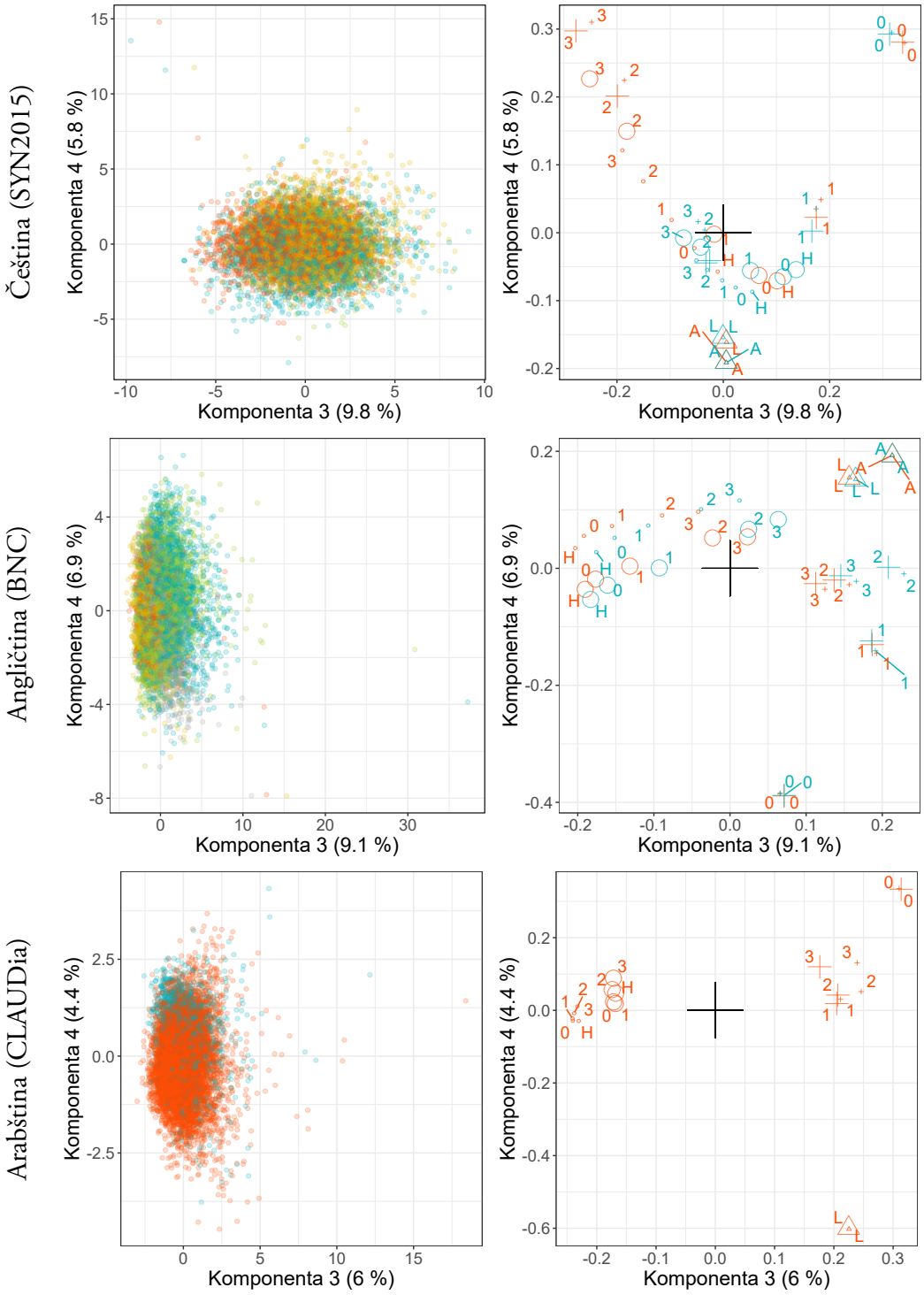
Obrázek 6.11: První dvě komponenty křížové lexikální diverzity (křížových entropií).



Obrázek 6.12: Další dvě komponenty křížové lexikální diverzity (křížových entropií).



Obrázek 6.13: První dvě komponenty všech představených metrik.



Obrázek 6.14: Další dvě komponenty všech představených metrik.

v efektivních bitech) bych tedy rozdělil do dvou lingvistických veličin, byť vzájemně souvisejících. V souladu s titulem této knihy se nyní budu věnovat pouze té první.

6.4 Lexikální diverzita jakožto lingvistická veličina

Až doteď jsem psal o metrikách lexikální diverzity jako o *metrikách lexikální diverzity*, aniž bych vysvětlil, proč jim neříkám třeba *míry*, *veličiny* nebo *indexy*. Věřím, že drtivá většina z vás prostě přijala mou terminologii a vlastně ani nepřemýšlela nad tím, že by mohla být nevhodná, a ten zbytek mou volbu blahosklonně akceptoval.

Nyní ovšem nazrál čas se nad touto volbou pozastavit. Termíny *metrika* (*metric*) a *míra* (*measure*) mají v matematice velmi konkrétní definici, nicméně v literatuře o lexikální diverzitě jsou to pojmy volně zaměňované a vlastně se používají jako synonyma, zvyšujíce tak lexikální diverzitu odborných textů o lexikální diverzitě pojednávajících. Zároveň jsou to pojmy v tomto kontextu nejpoužívanější. Termín *metrika* jsem si vybral v podstatě z estetických důvodů, neboť *míra* se používá v přirozené češtině v ještě větším množství roztodivných významů.[5] Pojem *veličina* (*quantity*) či *vlastnost* (*property*) se v anglofonní literatuře v tomto kontextu prakticky neužívá, jednak také nejspíš kvůli vysoké ambiguitě, ale hlavně proto, že *veličina* předpokládá určitou míru shody společenství na tom, co a jak se vlastně měří, v jakých jednotkách a jak se měření typicky operacionalizuje. Vlastně bych neměl problém uznat, že počet typů nebo převrácená pravděpodobnost opakování docela přesně specifikují metodu měření i užité jednotky a jsou dostatečně intersubjektivní a interpretovatelné, nicméně jen těžko můžeme označit lexikální diverzitu jako takovou za lingvistickou veličinu měřenou pomocí některé z těchto metrik, zejména kvůli tomu, že bychom se jen těžko shodli na tom, která z nich je ta správná.

Přesto podle mě stojí za to snažit se lexikální diverzitu jako lingvistickou veličinu etablovat.

Možným přístupem je chápat ji jako veličinu určovanou subjektivním hodnocením určitého počtu posuzovatelů.[6] O to se pokouší asi deset let Scott Jarvis, který ve svých experimentech nechává hodnotit různorodé texty lidmi s různou expertizou a následně hodnocení srovnává s klasickými metrikami lexikální diverzity, které nazývá „objektivními“ (dal bych spíše přednost označovat je za intersubjektivní).

Nechtěl bych Jarvisovi křivdit, základní myšlenka, se kterou přišel ve svém prvním článku na toto téma (Jarvis, 2013), byla bohulibá: celé dvacáté století jsme se snažili najít ideální metriku lexikální diverzity podle toho, že je nezávislá na délce textu, aniž

[5] Příkladem budiž Prótagorás a jeho „mírou všech věcí je člověk – jsoucích, že jsou, a nejsoucích, že nejsou“ nebo klasická školní definice fyzikální jednotky jako „míra veličiny“. V těchto smyslech metriky lexikální diverzity *mírou* rozhodně nejsou.

[6] V češtině se pro takovou veličinu občas používá termín *kvalimetrická veličina*, ovšem jde o označení omezené na G. G. Azgal'dova a jeho následovníky, mimo postsovětský prostor prakticky nepoužívané.

bychom se starali o to, co ona metrika vlastně měří; pojďme se podívat, jak klasické metriky korelují s tím, jak si lexikální diverzitu představují lidé. Jenomže když během svých experimentů zjistil, že tyto metriky se subjektivně určenými výsledky zrovna nekorelují ani jemu samotnému (Jarvis, 2017), ani jiným týmům (Bonvin – Lambelet, 2017), a to ani tehdy, když se pokusí různě měnit operacionalizaci měření (Jarvis – Hazhangmoto, 2021), a že nepomůže ani, když nechá hodnotitele, aby se „zkalibrovali podle objektivních metrik" (Kyle et al., 2021), tak z výsledků dedukuje, že holt jsou špatně ony metriky.[7]

Přitom subjektivní hodnocení různých fyzikálních veličin také zrovna lineárně nekoreluje s jejich etablovanými intersubjektivními metrikami (například v úvodu první kapitoly zmiňovaná teplota), a to jde o konkrétní a dobře vnímatelné vlastnosti fyzických objektů, oč horší to musí být s něčím tak abstraktním, jako je lexikální diverzita. Pochybuji, že někdy v dějinách někdo dejme tomu *sílu* měřenou siloměrem „validoval" podle nějakého subjektivního hodnocení síly. Osobně si tedy nemyslím, že snaha o subjektivizaci pojmu lexikální diverzita míří správným směrem, zejména kvůli tomu, že takto určované veličiny nejsou praktické pro normální vědecké zkoumání.

Pokud bychom za pravou a správně konceptualizovanou lexikální diverzitu označili ono subjektivní hodnocení a intersubjektivní metriky za pouhé indexy diverzitě se blížící, jen těžko budeme diverzitu usouvztažňovat do hypotéz, lingvistických zákonitostí a teorií, což považuji za základ našeho vědeckého snažení.

Představme si, že takto subjektivně definujeme třeba hmotnost. K určení hmotnosti balíku pak necháme skupinu lidí ohodnotit, jak je těžký na škále jedna až sedm, a za hmotnost určíme medián jejich hodnocení. Takováto subjektivní hmotnost pak může mít docela slušné využití — pokud například chceme varovat skladníka před těžkým balíkem, napíšeme na něj číslo označující jeho subjektivní hmotnost. Takové číslo může být pro skladníka i užitečnější než klasická hmotnost, jak si ji definuje fyzika, neboť některé balíky, byť fyzikálně lehké, se prostě špatně drží, jsou neforemné, kloužou, kteréžto všechny vlastnosti subjektivní hmotnost nejspíš vezme v potaz. Pokud ale na takto definované hmotnosti budete chtít postavit nějaký model reality schopný ji predikovat, tedy fyzikální zákony, pak to bude extrémně obtížné. Nikoli kvůli subjektivitě jako takové, ale protože subjektivní hmotnost pravděpodobně nebude aditivní a nebude mít nějaký rozumný vztah se subjektivní silou a subjektivním zrychlením. Také přeji hodně štěstí při určování subjektivní energie.

Proto bych uvítal, kdybychom diverzitu nechápali subjektivně, ale jako běžnou veličinu měřenou pomocí Hillova kontinua, tedy složenou z několika komponent. Všechny metriky Hillova kontinua mají stejnou jednotku (efektivní typy) a škálu,

[7] První z řetězu studií na toto téma zakončil doslova: „In this paper, I have argued that existing measures of LD, which are fundamentally measures of lexical repetition, are of limited usefulness because of their lack of construct validity and because of the fact that they do not reflect the psychological, linguistic, and social forces that give rise to the curious orderliness of language use to which Zipf (1935) referred" (Jarvis, 2017, str. 551). Nicméně v pozdějších publikacích se zdá být poněkud smířlivější.

když se tedy celočíselné metriky vydělené z Hillova kontinua normují nějak rozumně na délku textu či vzorku, můžeme s čistým svědomím mluvit o *intenzivní vektorové veličině*,[8] kde každá metrika Hillova kontinua je reprezentována jednou dimenzí vektoru.

Pokud jste dočetli až sem, tak vás asi nepřekvapí, že jako *rozumné normování* na velikost textu navrhnu metodu klouzavého okna. A že nebudu preferovat nějakou jednu konkrétní velikost okna, ale že naopak doporučím používat celé spektrum délek, aby tak vynikla lexikální diverzita na různých úrovních.

Tím se dostáváme k definici lexikální diverzity textu T jakožto *tenzoru druhého řádu* $\mathsf{D}_{\mathsf{ql}}(\mathsf{T})$, kde první sada dimenzí q určuje pozici metriky na Hillově kontinuu a druhá sada dimenzí l pak určuje délku klouzavého okna. Samozřejmě je možné využívat pouze část tenzoru, v praxi je to vlastně nutné, neboť q není shora omezeno.

Intenzivní tenzorová veličina zní docela sofistikovaně. Pokud vás to děsí, prostě si představte, že lexikální diverzitu neurčuje jedno číslo, ale několik čísel v dvourozměrné tabulce, například tab. 6.1.[9]

l \ q	0	1	2	3
100	73.2	59.7	43.1	31.43
300	180.7	118.0	59.0	35.48
1000	465.0	216.4	68.5	36.43
3000	1045.7	331.8	72.3	36.52
10000	2347.8	465.4	74.3	36.60

Tabulka 6.1: Výseč tenzoru charakterizujícího lexikální diverzitu románu *The Last of the Mohicans* od J. F. Coopera.

Takto definovaná lexikální diverzita jednak pokrývá široké spektrum toho, jak je diverzita chápána (užívá klasické metriky slovního bohatství, pravděpodobnost opakování a shannonovskou entropii), a zároveň bere v potaz, že text není vertikálně homogenní, tedy že diverzita krátkých úseků textu se může radikálně lišit od diverzity úseků větších, přičemž pro určení celkové lexikální diverzity jsou zásadní obě hodnoty.

Zde je několik důvodů, proč si myslím, že by se takto definovaná lexikální diverzita mohla a měla ujmout.

[8] Podle Helmova klasického rozdělení (intensive property). Bylo by samozřejmě užitečné mít i extensivní variantu, ale to je cíl nedosažitelný, tedy alespoň v našem konceptuálním prostoru, neboť je nedokážeme definovat jako aditivní. Asi se ptáte, jak to, že entropie, jakožto metrika lexikální diverzity, není aditivní, když fyzikální entropie aditivní je. Tady se právě projevuje, že entropie textu je jen básnická zkratka pro entropii systému, který vyprodukoval daný text.

[9] Z této tabulky můžeme například vyčíst, že v románu *The last of the Mohican* naměříme v sekvencích o tisíci tokenech průměrně 465 typů, v sekvencích délky 300 tokenů je perplexita průměrně 118 efektivních typů atd.

Tradice Počet typů je nejpoužívanější metrikou lexikální diverzity a slovní bohatství je pro mnoho autorů prostě synonymem lexikální diverzity. Shannonova entropie je univerzální metrika diverzity, na které je založený celý moderní svět, bez její existence byste si nemohli přečíst tento text. Pravděpodobnost opakování navrhlo jako metriku diverzity tolik lidí nezávisle na sobě a v tolika různých oborech, že se pro ni dodnes používá asi deset různých pojmenování. Hillovo kontinuum jenom dává těmto klasickým metrikám společný rámec, společnou škálu a společnou jednotku. Jak rozebírám v kapitole 3, právě škálování Hillova kontinua je příjemně intuitivní. Všechny tyto metriky se chovají obdobně jako prostý počet typů a také podle toho jednotku měření nazýváme jako efektivní typy.

Interpretovatelnost Zmíněné metriky jsou snadno interpretovatelné každá zvlášť. Hillovo číslo jako celek je možné přiblížit tak, že s rostoucím parametrem q roste vliv častějších slov na výslednou hodnotu. S rostoucí velikostí okna se mění měřítko, ve kterém operujeme.

Intersubjektivita Měření takto definované lexikální diverzity je deterministické a je možno ho libovolně opakovat. Pokud se upřesní detaily operacionalizace, není důvod, aby různé subjekty došly k různým výsledkům měření pro stejný text.

* * *

Pokud by nám pojetí lexikální diverzity jako tenzoru nevyhovovalo a skutečně bychom toužili po jedné jediné charakteristice, bylo by možné z celého tenzoru vydestilovat jedno číslo, například použitím první komponenty PCA, kterou jsme si osahali v kapitole 6.3. Takto definovaná lexikální diverzita by ovšem ztratila interpretovatelnost a jednotku, takže celý komplikovaný proces by byl spíše cestou zpět.

Přimlouval bych se tedy spíše za použití nějaké metriky vybrané z Hillova kontinua tak, aby její hodnota byla tak nějak uprostřed, dobrým kandidátem je Shannonova perplexita, už pro její prominentní pozici v rámci teorie komunikace, fyziky i matematiky.

Definovat lexikální diverzitu tak, aby splňovala podmínku interpretovatelnosti, intersubjektivity a určité intuitivnosti, je ovšem pouze první krok. Aby se definice ustálila a zaužívala, je třeba pomocí ní vyjadřovat hypotézy, musí být součástí lingvistických teorií. Pokud se nějakou takovou teorii chystáte vystavět, pevně doufám, že takto zúžená lexikální diverzita vám bude dobrým souputníkem. Pokud z nějakého důvodu selže, dejte mi prosím vědět.

Kapitola 7

Závěr

Kniha je o lexikální diverzitě *v textu*. Jenže jen málokdy nás zajímá text jako takový, obvykle ho bereme jen jako vzorek, jako okno do nějakého fenoménu, který nemůžeme pozorovat přímo. Do psychiky člověka, který onen text vyprodukoval. Do psychiky člověka, který onen text rád konzumuje. Do způsobu komunikace mezi nimi, do způsobu kódování a zpracování onoho kódu. Do historie onoho kódování, do procesu, pomocí kterého vzniklo a vzniká. Do procesů, pomocí kterých se ono kódování přizpůsobuje aktuálním okolnostem. Do témat, o kterých ona komunikace probíhá, do diskurzu, kterého je součástí, do stavu světa, který ona témata a onen diskurz ovlivňuje či přímo determinuje…

Podobně jako zvýšený obsah hemoglobinu ve vzorku krve může znamenat, že její zdroj bere doping, že je nemocný, že se tak narodil, nebo že strávil posledních pár let někde v horách, i při interpretaci lexikální diverzity záleží na kontextu.

Zde ovšem má práce končí. Je na vás, do jakého kontextu hodnoty, které jste naměřili, zasadíte a jak je interpretujete. Může pomoci, že oněch hodnot může být víc, že nemusí jít jen o jeden index zahrnující v sobě vše, ale o mnoho rozměrů, které mají potenciál podat plastický obraz předmětu zkoumání, ať už je jakýkoli.

Takto jsem také definoval lexikální diverzitu v poslední kapitole, ne jako jediné číslo, ale jako lingvistickou veličinu charakterizovanou několika hodnotami. Tolika hodnotami, kolik je potřeba. Celá monografie směřovala k tomu, abych na konci mohl sebevědomě prohlásit, že ony hodnoty nejsou náhodné uskupení nesouvisejících indexů, ale že mají stejnou škálu, stejnou jednotku, jasnou interpretaci a dobře určené vztahy mezi sebou. A také k tomu, abyste se vy, drahé čtenářky a čtenáři, mohli sebevědomě a kvalifikovaně rozhodnout, jak změřit lexikální diverzitu, aby měření reflektovalo potřeby vašeho výzkumu, jakou metodu zvolit, jak měření technicky provést a jak výsledku porozumět.

S úsměvem se teď dívám na první osnovu této studie, koncepci celé knížky jsem původně plánoval úplně jinak. Po krátkém úklidu a nalezení smysluplných metrik mělo následovat hledání lingvistických zákonů, které s lexikální diverzitou pracují,

formulace hypotéz, jejich empirické testování a usouvztažňování do teorií. Právě proto ostatně veličiny definujeme, abychom s nimi tohle mohli provádět, je to jejich smysl.

Nakonec jsem stihl vlastně jenom ten úklid a na testování hypotéz došlo jen, když jsem na ně narazil pod vrstvou prachu a nebylo možné je jenom tak odsunout na později.

Ta zábavnější část nás tedy teprve čeká.

Přílohy

Příloha A

Software LxDiversity

Pokud jste dočetli až sem, téma vás pravděpodobně zajímá. Pro ty z vás, kteří by rádi pracovali s popsanými metrikami, ale nemají čas nebo chuť programovat svou vlastní implementaci, jsem připravil uživatelské rozhraní pro implementace, které používám v této studii. Program je dostupný ze stránky `http://milicka.cz/LxDiversity`.

LxDiversity (zkráceně LxD) je jedním z mnoha dostupných programů pro měření lexikální diverzity, které vznikly jako vedlejší produkt nějaké teoretické práce. Některé z nich se rozrostly do obludných rozměrů a čítají několik desítek metrik (například *Text Complexity* od Thomase Proisla, 2021), jiné se skromně omezují na několik málo indexů relevantních pro tu kterou práci, například *TAALED* (Kyle et al., 2018), který obsahuje deset metrik spojených s články Kyla a Jarvise (Kyle et al., 2021). Osobně si myslím, že je lépe nezahltit uživatele velkým množstvím náhodných metrik, proto se také omezuji na ty, které jsou popsány v této práci a které dle mého mínění dávají smysl.

Výjimečnost LxD spočívá v tom, že umožňuje vypořádat se s problémem délky textu pomocí klouzavého okna u *všech* metrik, zatímco ostatní tuto možnost nabízejí pouze u slovního bohatství pod názvem MATTR (viz kapitolu 2.2.1). Dále umožňuje jednoduše spočítat metriky používající referenční korpus — křížová perplexita a entropie, Kullback-Leiblerova divergence a křížová pravděpodobnost distinkce, o kterých pojednává kapitola 1.6; také je to první software, díky kterému si lingvisté a textologové mohou osahat Hillovo kontinuum (kapitola 1.5) a rozdílnost (dissimilarity, kapitola 1.8).

Dále jsem se snažil o jazykovou univerzálnost, tedy je možno měřit i předtokenizované texty, nebo dokonce texty už zpracované jako korpus v dnes nejčastěji používaných formátech — různé vertikály, CoNLL-U a XML ve stylu BNC.

Podstatné pro mnohé také může být, že nástroj respektuje hranice textů a že umožňuje zpracovat více textů najednou. Software je plně offline a nevyžaduje instalaci

žádných dalších součástí, na nichž by byl závislý, tedy stačí program stáhnout, rozbalit a je připraven k používání, podmínkou je ovšem použití operačního systému Windows nebo nějakého jeho emulátoru.

A.1 Popis funkcionality a uživatelského rozhraní

Klasický rozpor mezi verzatilitou a jednoduchostí, tedy otázku, jak dát uživateli dostatek možností, aby si mohl program dobře přizpůsobit, ale zároveň aby se ve všech možnostech neztratil, řeším tak, že na začátku má uživatel jen minimální možnost volby, a teprve až podle jeho přání se postupně odkrývají další a další relevantní nastavení. Zejména jsem si dal pozor, aby alespoň jedna možnost (ta, která je nastavena jako default) byla pokud možno intuitivní, pochopitelná a snadná na interpretaci.

Po otevření programu na uživatele čekají dva panely: na levém určí, jaké texty se mají zpracovat a jak vypadají, na tom pravém pak má možnost specifikovat, jaké metriky si přeje změřit a jaké mají mít parametry.

A.1.1 Vstupy (levý panel)

Úplným začátečníkům je určena volba *Just copy and paste the text*, jak název napovídá, po jejím zvolení se nedá dělat nic jiného než prostě vzít libovolný text a zkopírovat ho do okna. Díky tomu se není třeba bát, jestli má soubor správné kódování a formátování. Program se už pak sám pokusí text natokenizovat, přičemž se předpokládá, že se jedná o jazyk, který při zápisu nějak přirozeně rozděluje text do slov (tedy ne například čínština). Při této volbě je ještě možno nastavit, zda si přejete, aby program chápal velikost písmen jako distinktivní, tedy jestli slovo *Lednice* a *lednice* patří ke stejnému typu, či nikoli. Osobně doporučuji ponechat defaultní nastavení, tedy *case insensitive*, neboť v textech psaných latinkou mají slova s velkým písmenem na počátku věty převahu nad slovy, které takto začínají ze sémantických důvodů. Dále je možné odfiltrovat číslice a interpunkci, což rovněž doporučuji, neboť čísla zapsaná číslicemi mají tendenci být v nevyčištěném textu spíše na obtíž — čísla stránek, poznámek pod čarou, tabulky se sportovními výsledky a podobně, naroste tak počet hapaxů a celkově typů. Naopak interpunkce je uzavřená množina prvků, takže zase neúměrně ovlivní metriky, které dávají větší váhu typům s velkou frekvencí (typicky pravděpodobnost distinkce), přestože má spíše formální charakter. Nemůžu ovšem vyloučit, že váš výzkum naopak vyžaduje zahrnutí číslic a interpunkce, proto tuto volbu nevylučuji.

Druhá možnost, tedy *Analyze texts in txt files*, dává uživateli mnohem větší kontrolu nad tím, v jakém formátu je vstupní soubor, zároveň ale předpokládá, že uživatel

bude tento formát dodržovat. Asi nejdůležitější je, aby textové soubory byly kódovány v UTF-8.[1] K dispozici jsou následující formáty.

Plain text

Program se k vašim textovým souborům bude chovat, jako kdybyste je zkopírovali do okénka (jako v předchozím případě), tedy pokusí se je nějak rozumně roztokenizovat. Za správnost tokenizace v různých jazycích ovšem neručím, pro náročnější jazyky a pro úkoly vyžadující preciznost je určena následující volba.

Words delimited by new lines

Program předpokládá, že každý token je na novém řádku, tedy jakási primitivní vertikála. Tento formát je ze všech nejobecnější — pokud zapnete volbu *Always case sensitive* a vypnete *Omit digits* a *Omit punctuations*, pak bude program zcela agnostický k obsahu jednotlivých řádků a vůbec se nebude plést do vašich rozhodnutí ohledně tokenizace či čehokoli jiného. Na řádcích může být cokoli: slova, fragmenty slov, morfémy, sousloví, přepisy ptačího zpěvu, části genetického kódu, identifikační čísla zákazníků či jména zvířat seřazená podle toho, jak přišla ke krmelci...

Words delimited by white spaces

Prakticky totožné s předchozím případem, pouze místo nových řádků jsou očekávány mezery.

Standard vertical

Poněkud ČNKcentricky očekávám, že standardní vertikála je klasický formát používaný k ukládání velkých korpusů řady SYN (tradice byla ovšem bohužel přerušena pro SYN2020, který má jiné pořadí sloupců). Tedy jedná se o jakousi tabulku, kde každý řádek reprezentuje jeden token, přičemž v prvním sloupci je nelemmatizované slovo, ve druhém sloupci, který od prvního odděluje tabelátor, je lemma, ve třetím morfologické značkování atd. Díky přítomnosti druhého sloupce dává smysl zaškrtnout políčko *Lemma* — tak je možné změřit metriky jak na nelemmatizovaném, tak na lemmatizovaném textu zároveň. Třetí sloupec s morfologickými tagy zase umožní změřit poměr autosémantických slov (ratio of autosemantics). Tyto možnosti se otvírají i při použití následujících korpusových formátů.

[1] Mimochodem, důrazně doporučuji používat kódování UTF-8 pro všechna textová data a tiše doufám, že pokud tuto knihu čtete po roce 2030, že teď nevěřícně kroutíte hlavou přemítajíce nad tím, jestli skutečně v roce 2022 někdo ještě používal jednobytové kódování textu. Bohužel používal.

COCA vertical

Korpus COCA[2] je jedna z nejjednodušších možností, jak se dostat k plnému neporušenému většímu korpusu angličtiny (byť ne zdarma), je tedy výhodný pro různé testování nebo k tvorbě referenčního slovníku. Vertikála má prakticky stejný formát jako klasická vertikála ČNK, ovšem první tři sloupce zabírají identifikační čísla souboru, textů a slov a také se liší meta tagy pro hranice textů. Starší vertikály (cca před rokem 2019) měly jiný formát a jednobytové kódování, pokud tedy máte texty staršího data, doporučuji zkontrolovat kompatibilitu. Tyto starší varianty (nebo novější) můžete zpracovat pomocí vertikály s vlastním nastavením (custom vertical).

Custom vertical

Švýcarský nožík, který vám dovolí pracovat v podstatě s libovolnou vertikálou. Nejprve specifikujete, jestli vertikála obsahuje lemmata a morfologické značkování, a následně určíte, ve kterém sloupci se nacházejí. Sloupce jsou číslovány od jedničky. Dále je potřeba určit, jak začíná řádek, který označuje hranici textů (například ve standardní vertikále ČNK je to dnes řetězec `<doc`, dříve to však byl `<opus`). Nakonec, pokud chcete měřit poměr autosémantických slov, je třeba určit, kterými písmeny začínají morfologické tagy značící autosémantická slova. Pokud potřebujete využít k určení autosémantických slov i jiné informace (například mezi ně chcete zařadit i některá slovesa, která se značkováním nijak neliší od ostatních), tak je nutné učinit změny přímo ve vertikále pomocí jiného vhodného nástroje (například regulárních výrazů v textovém editoru).

CoNLL-U

Tento formát se rozšířil díky popularitě projektu Universal Dependencies (Nivre et al., 2016, 2017; Marneffe et al., 2021),[3] jehož hlavní ambicí je sjednotit různé formalismy dependenční syntaxe, nicméně jako vedlejší produkt je díky němu k dispozici k volnému užití obrovské množství textového materiálu různé kvality ve stejném formátu.[4] CoNLL-U vychází z jednoduché vertikály v UTF-8, přičemž další sloupce obsahují bohatou syntaktickou anotaci, jenž je pro UD klíčová (Buchholz – Marsi, 2006), kterou však k měření indexů lexikální diverzity nijak nevyužíváme. Díky popularitě formátu CoNLL-U naleznete spoustu nástrojů, které umožňují automatickou konverzi z různých formátů právě do něj, nehledě na to, že pro mnoho jazyků skvěle funguje UD-pipe, nástroj pro automatickou anotaci prostých textů do formalismu UD (Straka, 2018).[5]

[2] https://www.english-corpora.org/coca/
[3] https://universaldependencies.org/format.html
[4] https://universaldependencies.org/
[5] https://ufal.mff.cuni.cz/udpipe

BNC xml

Obecně nejsem příliš velkým zastáncem užívání XML pro uchovávání korpusů, oproti vertikále jsou náročnější na parsování, a to jak z pohledu člověka (vertikálu můžete číst po otevření v textovém editoru holým okem a pochopit její strukturu je možné prakticky bez dokumentace, oproti tomu xml může, díky své verzatilitě, skrývat lecjaká překvapení), tak z pohledu stroje — skutečné programatické parsování XML zabírá řádově víc času než parsování vertikály, samotný formát je pak velmi redundantní a zabírá zbytečně moc místa. Přitom korpusová data jsou již z povahy věci „plochá“, tedy variabilní stromová struktura XML je zbytečná. Pokud se ovšem nesnažíme pomocí stromové struktury XML reprezentovat stromové struktury syntaktické, což je praktika, se kterou se doufejme budeme setkávat čím dál míň, mimo jiné i díky popsanému formátu CoNLL-U.

Nicméně v XML je uložena volně přístupná verze Britského národního korpusu, konkrétně The British National Corpus, version 3 (BNC XML Edition) z roku 2007 (dostupná po přihlášení z `https://cqpweb.lancs.ac.uk`), což je kvalitní zdroj dat, který stojí za to používat. Data jsou ovšem původní, tedy z roku 1993.

A.1.2 Výstupy (pravý panel) — frekvenční seznam

Primárním úkolem LxD je měření lexikální diverzity, ovšem některé metriky vyžadují frekvenční seznam nějakého referenčního korpusu. Konkrétně křížová a relativní entropie a perplexita a křížová pravděpodobnost distinkce a opakování vztahují váš měřený text k nějakým jiným referenčním textům. Nepotřebujete je ovšem pokaždé celé, stačí vám vytáhnout si z nich frekvenční seznam.

Toho dosáhnete tak, že na pravém panelu zaškrtnete možnost *Get frequency list of types*, čímž se otevře záložka, kam můžete zadat jména souborů, kam budete chtít seznam uložit, a stisknete tlačítko *Process the texts*. Tímto způsobem je možno zpracovat seznamy slovních typů a lemmat zároveň — pokud ovšem vstupní soubory lemmatizovaný text obsahují.

Samozřejmě nic vám nebrání již hotový slovník odněkud stáhnout, vyžadovaný formát je nejjednodušší možný: v textovém souboru (opět ve standardním kódování UTF-8) je seznam typů, na každém řádku jeden typ, za ním následuje tabelátorem oddělená jeho absolutní frekvence.

Nástroj na tvorbu frekvenčních seznamů můžete využít i jindy než při měření křížové lexikální diverzity, třeba pro zjišťování zipfovské distribuce nějakého textu nebo prostě pro jednoduchý přehledový frekvenční slovník.

Pokud vám tvorba referenčního frekvenčního seznamu přijde příliš složitá, docela dobře se bez ní obejdete, neboť většina metrik ho nepotřebuje.

A.1.3 Výstupy (pravý panel) — měření lexikální diverzity

Po stisknutí volby *Measure lexical diversity* uvidíte nejdůležitější volbu v celém programu: velikost okna (*Window length*). Všechny metriky se totiž zbavují závislosti na délce textu metodou klouzavého okna (podrobně popsáno v kapitole 2.2.1). Defaultní hodnota je malá — okno o délce 200 slov. Je nastavena tak, aby se vypořádala i s krátkými texty. Pokud ovšem máte texty delší, doporučuji zvolit mnohem větší velikost okna, řádově třeba polovinu délky nejkratšího textu. Respektive ještě zajímavější je zvolit více variant, dejme tomu 200, 400, 800, 1 600 atd., a následně se podívat, jestli se vzájemně liší a proč, jde o dimenzi, kterou je škoda nevyužít (viz kapitolu 5).

Dále se objeví tři metriky, které považuji za nejdůležitější, z toho pouze dvě metriky jsou zaškrtnuty, neboť nevyžadují referenční frekvenční slovník — cílem je, aby i zcela nezkušenému uživateli, který se zdráhá měnit jakákoli defaultní nastavení, na konci vyšlo něco, co může dobře interpretovat. Další metriky jsou přístupné teprve po zaškrtnutí položky *Other metrics*, aby totiž na uživatele nevyskočilo hned na začátku dvacet metrik najednou, rozdělil jsem je do tří kategorií.

V první kategorii jsou následující metriky:

Počet typů (number of types)

Prostý počet různých slov v textu (slovní bohatství, varieta), kterému se blíže věnuji v kapitole 1.1, považuji za nejdůležitější metriku lexikální diverzity především kvůli její jednoznačné a snadné interpretovatelnosti. „Počet různých slov“ je koncept, který je možno vysvětlit i člověku, který o tématu nic neví. Proto je na prvním místě.

Perplexita (perplexity)

Je o poznání hůře interpretovatelná, její význam by se dal laicky popsat přibližně jako „pokud by v textu měly všechny typy stejnou frekvenci, kolik typů by v textu muselo být, aby byl stejně překvapivý jako měřený text?“ (podrobněji v kapitole 1.4). Navzdory tomu si podle mě zasluhuje druhé místo, neboť zahrnuje frekvence typů, tedy postihuje mnohem více informací z textu, a navíc se přímo dotýká teorie informace a lexikální komplexity.

Pravděpodobnost distinkce (1.3) je sice také postavená na celé frekvenční distribuci typů, ovšem oproti perplexitě je bezrozměrným číslem, pravděpodobností, zatímco jednotkou perplexity je počet (rovnoměrně distribuovaných) typů, čili má srovnatelné měřítko jako předchozí metrika.

Křížová perplexita (cross perplexity)

Křížovou perplexitu podrobně popisuji v kapitole 1.6.1. Lidově řečeno vyjadřuje, jak moc by byl překvapen naším měřeným textem čtenář referenčního korpusu (jednotkami překvapení jsou opět počet rovnoměrně distribuovaných typů jako v případě

perplexity). Pokud má tedy text nízkou perplexitu, ale vysokou křížovou perplexitu, znamená to, že si autor slova vymýšlí, nebo že je text přímo napsán v nějakém jiném jazyce, než je jazyk referenčního korpusu.

* * *

Ve druhé kategorii jsou další metriky popsané v první kapitole, které může být zajímavé použít.

Počet hapax legomena

Viz kapitolu 1.2. Počet hapaxů zveličuje lexikální kreativitu, ať už záměrnou (neologismy, okazionalismy, jazykový humor), či nezáměrnou (překlepy, nesystematické chyby a „chyby“).

Převrácená pravděpodobnost opakování a převrácená křížová pravděpodobnost opakování (reverse repeat rate a reverse cross repeat rate)

Podobně jako perplexita bere v potaz celou distribuci frekvencí typů, ovšem slova s větší frekvencí zde hrají ještě větší úlohu (kapitola 1.3). Převrácená křížová pravděpodobnost opakování (1.6.1) opět počítá lexikální diverzitu měřeného textu optikou referenčního korpusu, dá se tedy interpretovat obdobně jako křížová perplexita.

Pravděpodobnost opakování a křížová pravděpodobnost opakování (repeat rate a cross repeat rate)

Totéž jako převrácená pravděpodobnost opakování (tedy opět kapitola 1.3), pouze převrácená hodnota. Ovšem tato varianta, známá pod jménem repeat rate, je v literatuře mnohem častěji zmiňována než distinction rate, proto ji zde zařazuji. Čím větší je lexikální diverzita textu, tím menší je pravděpodobnost opakování stejných slov — což sice dává intuitivně smysl, ovšem znamená to, že metrika je nepřímo úměrná ke všem ostatním zde zmiňovaným, což je poněkud nepříjemné.

Entropie a křížová entropie (entropy a cross entropy)

Entropie je prakticky stejná jako perplexita, jiné je pouze jen měřítko, které je u entropie logaritmické. Jelikož používáme binární logaritmus, jednotkou entropie a křížové entropie jsou bity. Škála je tedy stejná jako škála komplexity (pokud ji pojímáme kolmogorovovsky). Samotného výsledku metriky je možno dosáhnout prostým zlogaritmováním perplexity, nicméně pokud nás zajímá průměr pro celý text, dostaneme mírně jiná čísla (aritmetický průměr entropie odpovídá totiž geometrickému průměru perplexity).

Relativní perplexita a relativní entropie (relative perplexity a relative entropy)

Relativní entropie (Kullback-Leiblerova divergence, viz kapitolu 1.6.2) je alternativní způsob jak vztáhnout měřený text k referenčnímu korpusu, který zvýrazňuje rozdíly mezi nimi.

Rozdílnost (dissimilarity)

Při programování jsem kladl důraz na efektivitu, takže vše by mělo fungovat svižně. Například zpracovat Jiráskovo *Temno* do všech možných výstupů a metrik s výjimkou rozdílnosti vyjde zhruba na dvě vteřiny na současném mainstreamovém hardware, a to včetně načítání referenčního slovníku z beletrie korpusu SYN 2015. Pokud ovšem budeme chtít spočítat i rozdílnost, protáhne se doba výpočtu na dvacetinásobek. Rozdílnost, která je založená na porovnávání každého typu s každým a počítání nějaké metriky pro podobnost řetězců mezi všemi typy (v našem případě jde o starou dobrou Levenshteinovu editační vzdálenost — Levenshtein (1965); Levenshtein et al. (1966)) je totiž už z principu komputačně náročné, což rozebírám v příslušné kapitole (1.8). Za to se nám odvděčí schopností vnímat rozdíly mezi typy, tedy nebinární typizací.

Hillovo kontinuum metrik (Hill's continuum)

Pokud je mi známo, LxD je první program, který umožňuje lingvistům snadno využít celé Hillovo kontinuum metrik diverzity (kapitola 1.5). Je tak možné vyzkoušet, jak se budou lišit metriky diverzity, když umenšíme roli slov s vysokou frekvencí (nízký parametr Q), nebo naopak když ji zdůrazníme (vyšší parametr Q). Jelikož Hillovo kontinuum s nižšími parametry Q odpovídá jiným, již dobře vyzkoušeným metrikám (počet typů, perplexita, pravděpodobnost distinkce), zajímavé mohou být právě vyšší hodnoty, popřípadě gradient výsledků v rámci celého spektra.

Podíl autosémantik (ratio of autosemantics)

Program spočítá podíl autosémantických slov, což může být zajímavou pomocnou metrikou při zkoumání lexikální diverzity, neboť autosémantika jsou otevřená třída, zatímco synsémantika třída uzavřená (kapitola 1.9). Ovšem program můžeme použít na měření poměru *jakýchkoli* slov, která mají vlastnosti, které si sami určíte. Tedy není nutné program používat jen na distinkci autosémantických — synsémantických slov, ale prakticky na jakoukoli binární opozici, co do vertikály vměstnáte. Stačí využít možností, které nabízí *Custom vertical*, a do sloupce s tagy vložit informace, které si přejete zpracovat.

Délka slov (word length)

Jak je popsáno v kapitole 1.7, slovní délka není tak úplně metrikou slovní diverzity, nicméně souvisí s ní tak úzce, že má smysl ji měřit dohromady. Defaultně program měří délku slova v písmenech. To je ovšem v mnoha případech zcela nevhodné. Pokud tedy chcete měřit délku slov ve fonémech, slabikách či jakýchkoli jiných jednotkách, můžete využít toho, že referenční korpus může mít i třetí sloupec — opět oddělený tabelátory. Ten naplňte délkou daného typu v jednotkách podle svého uvážení. Proto se také při zaškrtnutí této metriky objeví možnost použít referenční slovník (stejně jako u křížových metrik). Délka se nemusí týkat celého slova, ale třeba jen jeho kořene či kmene… možnosti jsou neomezené.

* * *

Třetí kategorie metrik zahrnuje několik tradičních metrik, které naopak nedoporučuji používat, o čemž jsem se rozepsal v kapitole 1.10.1, nicméně jsou natolik oblíbené, že je můžete potřebovat, abyste mohli porovnat své vlastní výsledky s výsledky předchozích výzkumů.

TTR

Type-token ratio, tedy počet typů podělený počtem tokenů, je pozůstatkem raných neúspěšných pokusů, jak se zbavit vlivu délky textu na počet typů. Bohužel (respektive z pohledu efektivního fungování přirozeného jazyka naštěstí) závislost počtu typů na počtu tokenů není lineární, takže normování prostým podělením nedává smysl. Přesto se tato metrika s oblibou dodnes používá. Pokud vás zajímá její hodnota pro celý text, použijte výstup typu *Values for the whole texts*. Pokud zvolíte jako výstup průměrnou hodnotu v klouzavých oknech (typ *Averages*), získáte hodnotu klasické metriky zkracované jako MATTR (Covington – McFall, 2010).

LogTTR a RootTTR

Další metriky, které vznikly jako neúspěšný pokus vypořádat se s problémem nelineární závislosti počtu typů na počtu tokenů. Vzhledem k tomu, že všechny metriky normujeme pomocí klouzavého okna, problém tohoto vztahu není třeba řešit a jejich použití tak nedává smysl. Pouze pokud potřebujete porovnat LogTTR nebo RootTTR celého textu s těmito metrikami v nějakém starším článku, použijte výstup typu *Values for the whole texts* (viz následující podkapitolu).

A.1.4 Výsledky měření a jak s nimi pracovat

Výsledky nejsou zobrazovány přímo v uživatelském rozhraní programu, ale jsou rovnou exportovány do souborů formátu CSV, abyste je mohli dále zpracovat — zobrazit pomocí tabulkového procesoru[6] či analyzovat pomocí dalších skriptů. Může zvolit následující čtyři druhy výstupů:

Průměry (averages)

Pro každý text je změřena průměrná hodnota dané metriky normované pomocí klouzavého okna, metody, kterou definovali Covington – McFall (2010) a kterou rozebírám v kapitole 2.2.1. Pokud chcete mít lexikální diverzitu každého textu charakterizovánu jedním číslem pro každou metriku, pak je to právě toto číslo. Výsledky jsou exportovány do tabulky, kde každý řádek reprezentuje jeden text a každý sloupec jednu zvolenou metriku. Výsledky jsou uloženy do zvolené složky do souboru začínajícího řetězcem `Averages`.

Distribuce (distributions)

Jedno číslo reprezentující průměr může být zavádějící nebo prostě nedostatečné, protože texty mohou být, co se týče lexikální diverzity, vnitřně značně heterogenní, a právě ona heterogenita může být zajímavá. Proto dávám uživateli možnost prohlédnout si celou distribuci hodnot pro zvolené indexy pro jednotlivá okna. První zmínku o této metodologii najdete v Kubát – Milička (2013), kde je také celý proces podrobně popsán. Na výstupu má každá metrika svůj vlastní soubor a v těchto souborech každý text zabírá svůj řádek. Soubory začínají řetězcem `Distribution`.

Celé série (series)

Maximalistický výstup, který vám dá možnost prozkoumat výsledky zvolených metrik pro všechna okna v celém textu. Jelikož takto dlouhé série nemohou zabírat řádky (respektive tabulkové procesory, které mají striktněji omezený počet sloupců než řádků,

[6] Asi nejoblíbenější tabulkový procesor Microsoft Excel je bohužel známý tím, že otevřít standardní formát CSV neumí, neboť z nepochopitelných důvodů v české mutaci místo oddělovacích čárek vyžaduje středníky. To je možné vyřešit tak, že soubor vložíte pomocí *Data → Text do sloupců*. Pokud budete soubory otevírat pomocí Libre Office Calc, RStudia či standardních knihoven Pythonu, neměli byste narazit na problém. Aby byly soubory bez problémů přenositelné a bylo možné je používat i při mezinárodní spolupráci, pro reprezentaci desetinných čísel je vždy použita desetinná tečka nezávisle na lokálním nastavení vašeho počítače. Standardní knihovny skriptovacích jazyků obvykle počítají právě s desetinnou tečkou, zde by tedy mělo být vše v pořádku. Při otevírání v Excelu či Calcu naopak může být potřeba nahradit všechny tečky čárkami. Abyste se tomu vyhnuli, doporučuji v lokálním nastavení vašeho operačního systému nastavit jako defaultní desetinnou tečku, čímž se obvykle zbavíte i různých jiných problémů s jinými programy.

by s nimi měly následně problém), má zde každý text svůj vlastní výstupní soubor. Soubory začínají řetězcem `Serie` a následné číslo značí identifikátor textu.

Hodnoty metrik nijak nenormované, spočítané pro celý text (values for the whole texts)

Dává smysl pouze pro metriky, které jsou přirozeně nezávislé na velikosti textu (podíl autosémantik, průměrná délka slova). Vzhledem k tomu, že všechny ostatní metriky na délce textu závislé jsou, silně nedoporučuji pro ně tento typ výstupu využívat. Výsledky je možno nalézt ve zvolené složce v souboru začínajícím řetězcem `WholeTexts`.

Příloha B

Návrh nomenklatury

Je zvykem pojmenovávat indexy lexikální diverzity po náhodných mužích, kteří si kdesi v hloubi dvacátého století mysleli (obvykle mylně), že jsou prvními, kdo danou metriku použil (Good, 1982), popřípadě pomocí kryptických zkratek nějakého generického a málo popisného sousloví. Není možné jednoduše dekódovat, co se skrývá pod jmény jako Wittbergova kappa, Ziebrzyńského nTTR nebo HDgD. Uvedené názvy nejsou ani nijak mnemotechnické a kdybych vám teď neřekl, že jsem si je vymyslel, docela dobře byste mohli věřit tomu, že jsou jedny z desítek nebo stovek názvů, které něco skutečně znamenají.

Tento stav podle mě není udržitelný, protože, jak naznačuji v předchozím textu, možnosti jsou nevyčerpatelné a i když je nějak omezíme, díky kombinatorice se dostáváme k obrovskému množství metrik. Podobně jako si chemie vybudovala na troskách tzv. triviálních názvů sloučenin systematické názvosloví, mohli bychom i my zkusit názvy metrik lexikální diverzity systematizovat ve smyslu dimenzí popsaných v kapitole 6.1. Ideálně tak, aby podobné metriky měly podobný název, a bylo tak snadné dekódovat, v čem konkrétně ona podobnost spočívá.

Navrhuji poziční systém, v němž každá vlastnost metriky má své pevně dané místo, přičemž jako vhodný princip pro řazení oněch vlastností se mi jeví, aby modifikátory byly vždy vlevo od modifikovaného, podobně jako shodné přívlastky v přirozené češtině či v angličtině, a zhruba v souladu se současnou praxí (například modifikované TTR označujeme jako rootTTR či zTTR). Tedy čím více vpravo, tím více se daná vlastnost dotýká textu. Podrobnější specifikace modifikátorů pak navrhuji kódovat ve spodních a horních indexech — byla by možná také závorková notace a není důvod, proč by se paralelně nemohla používat. Celý systém je vidět v tabulce B.1. Je samozřejmě otevřený pro další rozšíření.

Na první pozici je škála. Vzhledem k tomu, že tradiční metriky lexikální diverzity jsou lineárně škálovány vůči počtu typů v textu, lineární škálování navrhuji chápat jako defaultní a nijak ho neoznačovat. Logaritmické škálování pak bude mít příznak

Škála	Porovnání	Metrika	Normování	Typizace
			oNF	
		T	oRF_{korpus}	eT
∅	∅	P	oOS^{i}	eL
Log	Ref_{korpus}	R	oSS^{i}	$eF_{lev...}$
		$Q_i^j k$	oNT^{i}	
		$L_{syl\|phon...}$	oBW^{i}	
			$oMP_{lin\|log\|root...}$	

Tabulka B.1: Schéma nomenklatury

Log. Na tomto místě bych také kumuloval různé další transformace, například *Rec* pro převrácenou hodnotu (reciprocal).

Na další pozici je naznačeno, jestli metrika vyžaduje referenční korpus (příznak *Ref*), a pokud ano, tak jaký (označeno v dolním indexu).

Na další pozici je konečně samotná metrika, buď pod svým původním názvem, nebo specifikovaná podle parametru q v Hillově kontinuu. Tedy T nebo Q0 pro počet typů, P nebo Q1 pro perplexitu, Q0.5 pro něco mezi, Q2 nebo R pro převrácenou pravděpodobnost opakování atd. Frekvenční omezení navrhuji značit pomocí indexů: spodní index pro nejnižší započítávanou frekvenci, horní pro nejvyšší. Chybějící horní index znamená maximální frekvenci, chybějící dolní index pak nulu. Tedy T^1 značí počet hapaxů, P_5 pak perplexitu s vyloučením slov s frekvencí nižší než pět. Samozřejmě se nebráním, aby byly do nomenklatury zahrnuty i metriky mimo Hillovo kontinuum, například délky tokenů (navrhuji značit jednotku, v níž se délka měří, spodním indexem, tedy L_{syl} pro délku měřenou ve slabikách).

Další dvě pozice zaujímá způsob, jakým je text nasegmentován, zkrácen nebo namodelován, aby se normalizoval vliv délky. V souladu s kapitolou 6.1.4 jsou k dispozici zkratky pro několik nejdůležitějších možností (tabulka B.2).

NF	nijak nezměněný plný text (natural full text)
RF_{korpus}	plný text normovaný podle referenčního korpusu (referenced full text)
OS^{i}	segmentace na sekv. o délce *i*, které se překrývají (overlapping sequences)
SS^{i}	segmentace na sekvence, které se nepřekrývají (subsequent sequences)
NT^{i}	text zkrácený na *i* tokenů (natural truncated text)
MP_{model}	text interpolovaný modelem (model parameter)
BW^{i}	náhodný vzorek nesousedících tokenů o velikosti *i* (bag of words)

Tabulka B.2: Možnosti normování

A konečně poslední pozici zabírá způsob typizace. Momentálně zavádím pouze tři: prostý nelemmatizovaný text, lemmatizovaný text a typizace podle nějaké složitější funkce, například normované Levenshteinovy vzdálenosti.

Pro snadnou vyslovitelnost a lepší orientaci v pozicích navrhuji vložit před poslední a předposlední pozici vokály. Vzniknou tak procedurálně generovaná pseudoslova, která jsou i v anglické výslovnosti vcelku použitelná.

Tímto získáváme systematické názvy pro několik set základních metrik, se započítáním Hillova kontinua a různých podrobných specifikací pak pro neomezené množství metrik. Každá z nich se hodí na něco jiného a mnohé se nehodí vůbec na nic, což ovšem neznamená, že nebyly v literatuře používány, pročež i ony si svůj systematizovaný název zaslouží. Pokud na nějaké dimenzi metriky nezáleží nebo ji nechceme specifikovat, je možné pozici vynechat, popřípadě ji pro přehlednost nahradit hvězdičkou. Asi nejlepší bude uvést nějaké příklady:

ToOS^{200}eL Počet lemmat normovaný pomocí pohyblivého okna o délce 200 tokenů.

ToNFeT Prostý počet slovních typů v celém textu.

LogPoSS$*$ Shannonův odhad entropie normovaný pomocí nepřekrývajících se segmentů nespecifikované délky, způsob typizace zde také není specifikován.

LogRef$_{\mathrm{BNC}}$PoNFeL Křížová entropie celého lemmatizovaného textu používající korpus BNC jako referenční korpus.

ToMP$_{\mathrm{lin}}$eT Počet typů v celém nelemmatizovaném textu, normovaný pomocí lineárního modelu, jinými slovy type token ratio, TTR.

ToMP$_{\mathrm{root}}$eT Počet typů v celém nelemmatizovaném textu normovaný pomocí odmocninového modelu, tedy metrika známá jako rootTTR či Guiraudův index.

RecRoOSeF$_{\mathrm{lev}}$ Metrika, kterou označujeme jako rozdílnost (dissimilarity) normovaná metodou překrývajících se oken.

RoNT$^{1000}*$ Převrácená pravděpodobnost opakování spočtená na prvních tisíci tokenech textu, lemmatizace není specifikována. Jedná se vlastně o Hillovo číslo druhého řádu, takže synonymní označení je Q2oNT$^{1000}*$.

Jak už to tak u procedurálně generovaných pseudoslov bývá, některá se mohou podobat slovům již existujícím, obzvláště v kontextu mezinárodní vědy tomu nelze zabránit a nezbývá než doufat, že se nějaká méně smysluplná metrika nezačne používat jenom kvůli vtipnému názvu.

Jednou z významných předností navrhované nomenklatury je, že zrovnoprávňuje indexy, které z historických důvodů dostaly krátké názvy, jako například TTR, s těmi, které přišly později. Název ToMP$_{\mathrm{lin}}$eT explicitně popisuje rozhodnutí, která byla

učiněna (v tomto případě normalizace pomocí lineárního modelu pro type-token relation a typizaci podle slovních forem), která se podle triviálního názvosloví jeví jako defaultní, avšak nejsou a neměla by být. Systém samozřejmě není uzavřený a je možné do něj přidávat další prvky.

Příloha C

Technické protokoly

Cílem této přílohy je podat podrobné vysvětlení všech technických aspektů, jakými jsem dospěl ke všem výsledkům. Hlavní snaha je, aby bylo možné všechny procedury bez problémů zopakovat, a zrekonstruovat tak tvorbu každého jednotlivého datového bodu i jeho vizualizace. Zdrojové kódy programů, hotové grafy a další koncové výstupy použité v této knize naleznete na adrese `http://milicka.cz/habilitace.zip`. Veškeré výstupy, které jsem oprávněn redistribuovat (tedy kromě samotných korpusů a nástrojů třetích stran), si můžete stáhnout na adrese `http://milicka.cz/habilitace_komplet.zip`. Jedná se o poměrně velký objem dat, takže pokud mě znáte osobně, je jednodušší mě navštívit s externím diskem.

Každý soubor má své pevné místo v adresáři na disku, nepoužívám relativních adres, neboť jak projekt rostl, byla adresářová struktura několikrát změněna. Považuji to tedy za bezpečnější.

C.1 Programové vybavení

Celý projekt byl komputačně poměrně náročný, vzhledem k množství zpracovaného textu a také kvůli k tomu, že jsem používal resampling přímo zdrojových korpusů. Bylo tedy výhodné vyvinout programové vybavení v kompilovaném programovacím jazyce, který dovoluje nízkoúrovňovou manipulaci s daty a správu paměti. Vzhledem k dobrým zkušenostem jsem zvolil *Embarcadero Delphi 10.4*,[1] což sice není volný software ani open source, ale pro nekomerční využití je zdarma (stav v roce 2022).

Pro grafickou reprezentaci dat využívám *RStudio*.[2] Konkrétně verzi 2021.09.1 Build 372, které pracuje s R verze 4.1.2. Grafy jsou tvořeny a renderovány zejména za použití knihovny `ggplot` a přidružených balíčků `ggExtra` a `ggpubr` a jsou následně

[1] Dostupná na adrese `https://www.embarcadero.com/products/delphi/starter/free-download`.

[2] Dostupné z `https://www.rstudio.com`.

exportovány do pdf pomocí `cairo_pdf` (vše open source), neboť standardní export do pdf v RStudiu stále ještě neumí pracovat s UTF-8 (stav roku 2022). Detaily naleznete v samotných skriptech.

Programy jsou napsány tak, aby výstup byl vždy uložen na disk na nějaké předem určené standardní adrese, se kterou počítají další programy, které na něj navazují. Tato umístění jsou také popsána průběžně v tomto protokolu.

Mé skripty v R nepoužívejte jako inspiraci, jsem v tomto jazyce věčný začátečník a jde spíše o slepence kódů, u kterých je podstatné, že fungují, rozhodně však nejsou příkladem správného programování, natož aby byly elegantní. Poněkud lepší je situace s programy napsanými v Delphi, které je mým mateřským programovacím jazykem, nicméně ani zde není kód příliš estetický, neboť rostl poměrně divoce podle potřeb projektu bez možnosti plánovat dopředu, a přestože prošel několikerým refaktoringem, rozhodně není etalonem správné architektury.

C.2 Korpusy a jejich příprava

Korpusy jsou popsány a jejich tvůrci řádně ocitováni v úvodní kapitole, zde naleznete pouze technické poznámky, jak se k nim dostat a jak byly zpracovány.

Nejprve byly všechny soubory převedeny do stejného formátu. Vzhledem k efektivitě jsem vybral jako společný formát vertikálu (stejného nebo obdobného stylu, jako se používá v korpusech Českého národního korpusu). Korpusy ČNK tedy nebylo nutné nijak upravovat. Vertikála ČNK je jakýmsi hybridem mezi XML a TSV (tab separated values) formátem, kdy metainformace a různé formátovací značky jsou kódovány pomocí tagů formálně podobných XML, ovšem o XML se nejedná, už proto, že konce řádků a tabelátory neztrácejí význam, naopak hrají důležitou roli. Takovýto popis může znít sice trochu zmateně, nicméně s vertikálami je velmi jednoduché a efektivní pracovat.

Pokud budete chtít studii replikovat na svých vlastních datech, je tedy nejprve potřeba převést je do vertikály. Vzhledem k jednoduchosti tohoto formátu by to ovšem neměl být problém. Fakticky se jedná o tabulku, kde každý řádek obsahuje jeden token, přičemž první sloupec je obsazen slovními formami, druhý lemmaty a třetí morfologickými tagy. Hranice textů jsou označeny pomocí pseudo-XML tagů `<doc >`, které jako atributy mohou nést další metadata, např. `<doc author="Nezval,Vítězslav">`. Vertikála může obsahovat další pseudo-XML tagy, například `<p>` pro hranice odstavce nebo `<s>` pro hranice věty, ty jsou ovšem v naší studii nepotřebné, a tedy jsou ignorovány.

V jazycích, které rozlišují kapitálky, zachovávám velké písmeno pouze tehdy, když je obsaženo i v lemmatu, tedy například velká písmena na začátku věty jsou změněna systematicky na malá.

SYN2015

Jakožto zaměstnanec Ústavu Českého národního korpusu na Univerzitě Karlově jsem měl možnost využít exkluzivního přístupu přímo k celému korpusu v původní podobě. Tento korpus ovšem v této původní podobě není veřejně dostupný, pouze je možné se dostat k tzv. zashufflované verzi, jejíž odstavce byly v rámci textů náhodně zpřeházeny, aby při redistribuci nedocházelo k porušení copyrightu.[3] Nicméně není možné očekávat, že na takto zashufflovaném korpusu získáte stejné výsledky, jako jsou v této knize. Pokud tedy nemáte přístup k vertikálám ČNK jakožto zaměstnanci ÚČNK nebo spřátelených ústavů, nezbývá vám než použít již naměřené a předpočítané metriky, které dávám k dispozici na již zmíněné adrese `http://milicka.cz/habilitace_komplet.zip`, nebo najít nějaký dostupný srovnatelný korpus s volnými právy pro distribuci. V současné době (říjen 2022) takový ovšem neexistuje.

Vertikála je dále očekávána jako soubor `habilitace\vertikalaSyn2015.txt`. Pomocí programu `FilterGenres` je z ní vydělen subkorpus obsahující pouze beletrii `habilitace\vertikalaSyn2015FIC.txt`.

Pokud není výslovně uvedeno jinak, tak při měření indexů lexikální diverzity z korpusu odstraňuji interpunkci (tedy tokeny, které mají tag roven `Z:`) a tokeny, které obsahují číslice nebo které mají příslušné tagy[4] (`C=` pro arabské číslice a `C}` pro římské číslice).

Za autosémantika jsem zjednodušeně považoval slova, která mají tag začínající na `N`, `A`, `V` nebo `D`, ale nemají lemma *být* nebo *mít*.

BNC

Prakticky volně přístupná je použitá verze Britského národního korpusu, konkrétně The British National Corpus, version 3 (BNC XML Edition) z roku 2007, která je dostupná po přihlášení z `https://cqpweb.lancs.ac.uk`.

Korpus je, jak už název napovídá, původně distribuován ve formátu XML, který jsem ovšem převedl do vertikály pomocí programu `BNCToVertical.exe`, který naleznete ve složce `Lingvisticke moduly\InterfacesToCorporaAndDictionaries\BNC`. Vertikálu další programy očekávají jako soubor `habilitace\vertikalaBNC.txt`.

Pomocí programu `FilterGenres` je z ní vydělen subkorpus obsahující

[3] Tuto verzi je možné stáhnout na adrese `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1593`. Uživatelé korpusu pak k němu standardně přistupují přes Kontext nebo jiné online rozhraní, takže je při běžné vědecké práci omezení copyrightem nemusí trápit.

[4] Popis morfologického značkování korpusů řady SYN lze nalézt na adrese `https://wiki.korpus.cz/doku.php/seznamy:tagy`.

psané texty	habilitace\vertikalaBNCwritten.txt
psané beletristické texty	habilitace\vertikalaBNCwrittenFIC.txt
mluvené texty	habilitace\vertikalaBNCspoken.txt

Pokud není výslovně uvedeno jinak, při měření indexů lexikální diverzity z korpusu odstraňuji interpunkci (tedy tokeny, které mají tag roven PUN) a tokeny obsahující číslice (zde není možné využít anotaci, protože tag CRD nerozlišuje mezi číslovkou vyjádřenou slovem a číslicí).

Autosémantika jsem rozlišoval jako slova, které mají tag, který začíná na AJ, NN, VV nebo AV0.[5]

CLAUDia

Podobně jako korpusy z rodiny SYN, ani CLAUDia není veřejně dostupná. Měl jsem k ní přístup díky svému angažmá na Ústavu srovnávací jazykovědy Filozofické fakulty Univerzity Karlovy. Existuje více veřejně dostupných korpusů diachronní arabštiny,[6] nicméně i když pomineme rozdíly v kvalitě, abychom docílili srovnatelných výsledků, bylo by nutné daný korpus zformátovat do vertikály a přiřadit mu metadata, která v této studii používám.

Další programy, podobně jako v ostatních případech, vertikálu očekávají na adrese habilitace\vertikalaCLAUDia.txt.

Devokalizovaná verze, ze které byly odstraněny samohlásky, kašídy a sjednoceno psaní alifů, je pak standardně na adrese habilitace\vertikalaCLAUDiaDevoc.txt.

Pokud není výslovně uvedeno jinak, při měření indexů lexikální diverzity z korpusu odstraňuji interpunkci (tedy tokeny, které obsahují interpunkční znaky, vzhledem k absenci anotace jsem znaky vybral ručně) a tokeny obsahující číslice.

Další jednotlivé texty

The Last of the Mohicans

Text, který je volně přístupný ze stránek projektu *Gutenberg.org*,[7] jsem pouze očistil od edičních poznámek, sjednotil velikost písmen a odstranil číslice. Text a všechny související soubory jsou ve složce habilitace\mohican.

[5] Určeno podle http://www.natcorp.ox.ac.uk/docs/c5spec.html.

[6] Například Tashkeela dostupná veřejně ze stránky https://sourceforge.net/projects/tashkeela.

[7] https://www.gutenberg.org/ebooks/940.

Temno

Text, který jsem stáhl ze stránek Městské knihovny v Praze,[8] odkud je volně dostupný, jsem pouze očistil od edičních poznámek, sjednotil velikost písmen a odstranil číslice. Text a všechny související soubory jsou ve složce `habilitace\temno`.

C.3 Předpočítání indexů lexikální diverzity

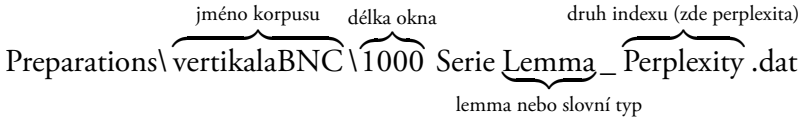
Většina prezentovaných dat vychází z předpočítaných indexů lexikální diverzity pro každý korpus. Jde o implementaci všech indexů z kapitoly 1 pro všechna okna o určitém počtu tokenů v textu. Tedy prakticky jde o použití metody délkové normalizace popsané v kapitole 2.2.1. Díky tomu je možné následně dělat rychle a efektivně vzorky z korpusů, popřípadě nějak limitovaných subkorpusů, a dál je zpracovávat.

Předpočítávání indexů je sice poněkud zdlouhavé a indexy zabírají místo, nicméně je velmi výhodné z pohledu výzkumného workflow: během práce na této studii bylo počítáno řádově více statistik, než je uvedeno v této knize, a díky předpočítaným indexům bylo možné vzorkovat miliony náhodných vzorků v podstatě instantně. Práci tak nebylo nutné přerušovat kvůli čekání na výsledky. Navíc je možné takto zpracovaná data sdílet (na rozdíl od plných verzí korpusů, kde tomu brání copyright), takže máte možnost si je stáhnout,[9] pracovat s nimi dál, zopakovat mé statistiky, popřípadě si je přizpůsobit vlastním potřebám, nebo zkusit něco úplně jiného, co v této knize není.

Indexy jsou předpočítány pomocí programu `MakeSeries`. Nejprve je nutné vytvořit slovník typů a lemmat pomocí funkce `PrepareDictionary`, které jsou nutné pro vypočítání křížové perplexity a dalších indexů, které vyžadují referenční korpus.

Tyto připravené soubory jsou ve složce `Preparations` v podsložce podle názvu vertikály jednoho každého korpusu, tedy například `Preparations\VertikalaBNC\`.

Daná složka obsahuje zejména zmíněné série pro každý index, a to jak pro lexikální typy, tak pro lemmata s následujícím způsobem pojmenování:



Při měření sérií byly respektovány hranice textů, takže pokud je velikost okna dejme tomu 1000 tokenů, tak pro prvních 999 tokenů v textu je metrika nedefinována.

Složka dále obsahuje následující soubory:

[8] https://web2.mlp.cz/koweb/00/04/48/04/92/temno.pdf.

[9] http://milicka.cz/habilitace_komplet.zip.

`Metadata.txt`	Seznam metadat jednotlivých textů
`SerieMeta.dat`	Série, která pro každé slovo určuje, v jakém je textu
`DictWord.txt`	Seznam slovních typů a jejich frekvencí v celém korpusu
`DictLemmata.txt`	Seznam lemmat a jejich frekvencí v celém korpusu

`SerieMETA.dat` je pomocný soubor, který obsahuje metadata pro jednotlivé texty (každý text na novém řádku). Podobně `SerieTokensInTextNumber.dat` je pouze pomocný soubor, který určuje pořadí daného tokenu v textu a zrychluje tím vybírání náhodného vzorku, který nejde přes hranice textů.

C.4 Výběr vzorku

Výběr náhodného vzorku N po sobě jdoucích slov, která nejdou přes hranici textu, je obtížnější, než se na první pohled zdá. Naivní metoda „vyber náhodnou sekvenci slov, a když bude obsahovat hranici textu, tak ji nepoužij a místo ní vyber jinou“ totiž preferuje delší texty. Používáme tedy algoritmus „vyber náhodnou sekvenci slov, a když obsahuje hranici textu, tak se v textu posuň dopředu o délku vzorku tak, aby začínal za poslední hranicí textu.“ Tato metoda zaručuje přibližně stejnou preferenci různých délek textů (tedy kromě těch, které jsou kratší než N) a zároveň zachovává požadavek na přirozenou kontinuitu vzorku.

C.5 Příprava grafů

C.5.1 Obrázek 1.1

Grafy vznikly ze stejných dat a pomocí stejného skriptu jako ty, jejichž vznik je popsán v podkapitole C.5.12.

C.5.2 Obrázky 1.2, 1.3, 1.4, 1.5, 1.6 a 1.7

Soubor `Statistics\<jmenokorpusu>\1000RandomSamples.dat`, který využijeme i v rámci podkapitoly C.5.9, byl přetvořen na dendrogramy pomocí skriptu `CorrelogramLarger.R`. Pokud byste měli zájem, úpravou tohoto skriptu se dá získat korelace libovolných dvou metrik, kterými se v této práci zabývám.

C.5.3 Obrázky 1.8 a 1.9

Při vytváření podkladů pro tyto grafy jsem postupoval velmi nesystematicky.

Začněme tím, že pro data byla vytvořena speciální složka `Habilitace\LengthInSyllables\`, neboť bylo potřeba vytvořit paralelní soubory jednak pro průměrnou délku slova měřenou v počtu slabik (standardně je měřena v počtu písmen), jednak

pro alternativně měřenou křížovou entropii. Obě alternativní metriky se od svých původních variant liší tím, jaký je použit slovník (`DictWord`). Tyto statistiky byly změřeny a vizualizovány pouze pro češtinu a angličtinu, neboť pro arabštinu se mi nepodařilo najít spolehlivý nástroj na segmentaci na slabiky.

Úprava slovníků proběhla napřed pomocí skriptů `SyllablesCS` a `SyllablesEN` — pro češtinu jsem přizpůsobil kód, který vytvořili Zuzana Oceláková a Tomáš Bořil (2020), zatímco pro angličtinu jsem využil knihovny `sylcount` (Schmidt, 2022). Vzhledem k tomu, že standardní příkaz `write` v R Studiu má v roce 2022 problém uložit ve Windows soubor v UTF-8 a už mi tekly nervy, byl jsem nucen akceptovat jednobytové kódování. Abych touto konverzí neztratil informace, tak jsem následně pomocí programu `TableCorrection.exe` první sloupec slovníku zkopíroval z původního souboru a vše následně uložil do UTF-8. Při té příležitosti tento program seřadí slovník od nejčastějšího slova po nejméně časté a místo frekvence přiřadí číslo využívané pro výpočet alternativní křížové entropie.

Pro výpočet průměrné délky slov měřené ve slabikách a alternativní metriky křížové entropie byl vytvořen fork procedury `PrepareData` (což opět není dvakrát čisté řešení) pod jménem `PrepareDataCrossEntropyOfRanks`, která je volaná z procedury `PrepareAlternativeLength` v programu `makeseries.exe`, který standardně vytváří série i všech ostatních metrik.

Vše bylo následně zpracováno pomocí skriptu `CorrelogramWordLenth`, jehož výsledek vypadá mnohem úhledněji než proces, jakým k němu bylo dospěno.

C.5.4 Obrázky 1.12, 1.13, 1.14, 1.15 a 1.20

Nejprve byly na textech románů *Temno* (celý text a další přidružené soubory naleznete ve složce `Habilitace\SingleTexts\Temno\`) a *The Last of the Mohican* (ve složce `Habilitace\SingleTexts\Mohican\`) pomocí programu *TypeTokener*[10] změřeny type-token relation (TTR, soubor `Concordance.txt`).

Pomocí Eureqy (Schmidt – Lipson, 2009) byly nalezeny parametry potřebných modelů. Modely jsou v souboru `SingleTexts\Mohican\modelTTRLambda.fxp`. Takto vytvořené datasety byly následně zpracovány pomocí skriptu `TypeTokenRatio.R`, a tím byly vygenerovány grafy.

C.5.5 Obrázky 1.18 a 1.19

Nejprve byly na textech románů *Temno* a *The Last of the Mohican* změřeny rank-frequency relation (RFR, soubor `M1.txt`) pomocí programu *TypeTokener*. Z takto vytvořených datasetů byly následně pomocí skriptu `RFR.R` vygenerovány grafy.

[10] Předchůdce programu LxDiversity, dostupný z `http://milicka.cz/typetokener/`.

C.5.6 Obrázky 1.10, 1.11, 1.16, 1.17 a dále 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9, 2.10, 2.11 a 2.12

Všechny tyto grafy vznikly pomocí programu `ErrorsByTextLength.exe`, který ze sérií pro okna různých délek vybral 1 000 000 vzorků, čímž vytvořil datasety, které byly následně zpracovány pomocí skriptu `ValidityOfMetricsLength.R`.

C.5.7 Obrázek 2.1

Dataset byl zpracován pomocí skriptu `ConvergencePerplexity.R`.

C.5.8 Obrázky 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12 a tabulky 3.1, 3.2 a 3.33.4

Stejně jako v případě C.5.9, pomocí procedury `GetSamples` vytvoří program `CompareSamples.exe` 3 000 náhodných sekvencí o délce 1 000 tokenů (pomocí metody popsané v C.4), které jsou následně změřeny, a výsledky metrik lexikální diverzity jsou pak uloženy do souboru `Statistics\<jmenokorpusu>\1000RandomSamples.dat`. Grafy jsou pak z těchto dat generovány pomocí skriptu `ScatterPlotsScale.R`. Stejný skript také vytváří data pro tabulky 3.1, 3.2 a 3.33.4.

C.5.9 Obrázky 4.1, 4.3, 4.4, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12 a 4.13

Program `GetAndCompareSamples.exe` vytvoří pomocí procedury `GetSamples` 3 000 náhodných sekvencí o délce 1 000 tokenů (pomocí metody popsané v C.4), které jsou změřeny, a výsledky metrik lexikální diverzity jsou následně uloženy do `Statistics\<jmenokorpusu>\1000RandomSamples.dat`. Horní dva grafy jsou následně vygenerovány pomocí skriptu `ScatterPlotsLemmata.R`.

Tentýž program také vytvoří pomocí procedury `GetSamplesAndCompareLemma` 3 000 náhodných dvojic sekvencí o délce 1 000 tokenů, které jsou následně porovnány a uloženy do `Statistics\<jmenokorpusu>\1000RandomSamplesComparison.dat`. Spodní graf je pak vygenerován skriptem `ScatterPlotsLemmataDifference.R`.

C.5.10 Obrázky 4.2 a 4.5

Program `GetAndCompareSamples.exe` pomocí procedury `CompareSamples` vybere dvakrát 100 000 000 náhodných sekvencí o délce 1 000 tokenů, tyto dvojice vzorků jsou pak následně vždy porovnány a distribuce jejich rozdílů jsou uloženy do `Statistics\<jmenokorpusu>\1000CompareSamplesLemmaVsTokenMistakes.d`

at. Grafy jsou pak generovány pomocí skriptu `DistributionLemmataDifferences.R`.

C.5.11 Obrázky 4.14 a 4.15

Oba grafy vznikly pomocí programu `ErrorsLemmatization.exe`, který ze sérií pro okna různých délek vybral 1 000 000 dvojic vzorků, porovnal je a určil chybovost, čímž vytvořil datasety uložené do souborů `Statistics\<jmenokorpusu>\ErrorsByLemmatization.dat`, které byly následně zpracovány pomocí skriptu `ValidityOfMetricsLemmatization.R`.

C.5.12 Obrázky 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8 a 5.9

Program `DiffLengthSample.exe` sampluje pomocí procedury `GetSamplesLength` 3 000 náhodných sekvencí o délce 1 000 tokenů (pomocí metody popsané v C.4), na nich změří všechny potřebné metriky a následně na stejném vzorku vypočítá klouzavý průměr těchtýž metrik pro okno 100 tokenů. Tyto výsledky následně uloží do souboru `Statistics\<jmenokorpusu>\100-1000Random3000SamplesWindowLengthDiff.dat`. Stejná procedura se provede pro náhodné sekvence velikosti 10 000 tokenů a 1 000 tokenů, přičemž výsledky se uloží do souboru, ano, hádáte správně, `Statistics\<jmenokorpusu>\1000-10000Random3000SamplesWindowLengthDiff.dat`. Grafy jsou následně vygenerovány pomocí skriptu `ScatterPlotsWinLengths.R`.

C.5.13 Tabulka 6.1

Všechny hodnoty jsou změřeny na celém románu Jamese Fenimora Coopera *The Last of the Mohicans* za použití programu LxDiversity (Příloha A).

C.5.14 Obrázky 6.3, 6.4, 6.5, 6.6, 6.7 a 6.8

Stejně jako v případě C.5.12 využíváme program `DiffLengthSamples.exe`, tentokrát ovšem pomocí procedury `GetSamplesLength` samplujeme 10 000 náhodných sekvencí o délce 100 a 1 000 tokenů, výsledky jsou uloženy v souboru `Statistics\<jmenokorpusu>\100-1000Random10000SamplesWindowLengthDiff.dat`. Tento datový soubor je pak zobrazen pomocí skriptu `Dendrograms.R`.

C.5.15 Obrázky 6.9, 6.10, 6.11, 6.12, 6.13 a 6.14

Stejně jako v předchozím případě (C.5.14) využíváme souboru `Statistics\<jmenokorpusu>\100-1000Random10000SamplesWindowLengthDiff.dat`, kde se ukrývají výsledky měření na 10 000 náhodných sekvencích o délce 100 a 1 000

tokenů, nasamplovaných a změřených pomocí programu `DiffLengthSamples.exe` a procedury `GetSamplesLength`. Tento datový soubor je pak zobrazen pomocí skriptu `PCAMetrics.R`.

Seznam obrázků

Seznam tabulek

Seznam použité literatury

ADELMAN, M. A. Comment on the 'H' concentration measure as a numbers-equivalent. *The review of economics and statistics*. 1969, s. 99–101.

BAAYEN, H. Quantitative aspects of morphological productivity. In *Yearbook of morphology 1991*. Dordrecht: Springer, 1992. s. 109–149.

BATESON, G. *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology*. Bungay Suffolk: Chandler Publishing, 1972. ISBN 9780226039053.

BENTZ, C. – CANCHO, R. Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*, s. 1–4. University of Tübingen, 2016.

BONVIN, A. – LAMBELET, A. Algorithmic and subjective measures of lexical diversity in bilingual written corpora: a discussion. *Corela. Cognition, représentation, langage*. 2017.

BRIEST, W. Kann man Verständlichkeit messen? *STUF-Language Typology and Universals*. 1974, 27, 1–3, s. 543–563.

BUCHHOLZ, S. – MARSI, E. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, s. 149–164, 2006.

ČECH, R. Text length and the lambda frequency structure of a text. In *Sequences in language and text*. Berlin: De Gruyter Mouton, 2015. s. 71–88.

CHAO, A. – CHIU, C.-H. – JOST, L. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. *Annual review of ecology, evolution, and systematics*. 2014, 45, s. 297–324.

CHEN, H. – LIU, H. A diachronic study of Chinese word length distribution. *Glottometrics*. 2014, 29, s. 81–94.

CHEN, H. – LIANG, J. – LIU, H. How does word length evolve in written Chinese? *PloS one*. 2015, 10, 9, s. e0138567.

CHO, S. et al. Automated analysis of lexical features in frontotemporal degeneration. *Cortex*. 2021, 137, s. 215–231. ISSN 0010-9452.

CHOTLOS, J. W. IV. A statistical and comparative analysis of individual written language samples. *Psychological monographs*. 1944, 56, 2, s. 75.

CONSORTIUM, B. British national corpus, 2007.

COVINGTON, M. A. CoVec: Measuring the similarity of words and the coherence of texts, 2016. Dostupné z: <www.covingtoninnovations.com/software.html>.

COVINGTON, M. A. – MCFALL, J. D. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of quantitative linguistics*. 2010, 17, 2, s. 94–100.

CVRČEK, V. et al. *Registry v češtině*. Studie z korpusové lingvistiky. NLN, 2020. ISBN 9788074227547.

CVRČEK, V. How large is the core of language. In *Proceedings from the sixth international corpus linguistics conference*, 2011.

CVRČEK, V. – CHLUMSKÁ, L. Simplification in translated Czech: a new approach to type-token ratio. *Russian linguistics*. 2015, 39, 3, s. 309–325.

CVRČEK, V. – ČERMÁKOVÁ, A. – KŘEN, M. Nová koncepce synchronních korpus psané češtiny. *Slovo a slovesnost*. 2016, 77, 2.

DALLER, H. – VAN HOUT, R. – TREFFERS-DALLER, J. Lexical richness in the spontaneous speech of bilinguals. *Applied linguistics*. 2003, 24, 2, s. 197–222.

DURÁN, P. et al. Developmental trends in lexical diversity. *Applied linguistics*. 2004, 25, 2, s. 220–242.

ELLERMAN, D. Logical information theory: new logical foundations for information theory. *Logic journal of the IGPL*. 2017, 25, 5, s. 806–835.

EVERT, S. – WANKERL, S. – NÖTH, E. Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch. In *Proceedings of the Corpus linguistics 2017 conference, Birmingham, UK*, 2017.

FAIRBANKS, H. II. The quantitative differentiation of samples of spoken language. *Psychological monographs*. 1944, 56, 2, s. 17.

CANCHO, R. – BENTZ, C. – SEGUIN, C. Optimal coding and the origins of Zipfian laws. *Journal of quantitative linguistics*. 2022, 29, 2, s. 165–194.

FRIEDMAN, W. The index of coincidence and its applications in cryptography. Publication No. 22, 1922.

FUCKS, W. Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen. In *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen*. Berlin: Springer, 1955. s. 5–110.

GELMAN, A. – LOKEN, E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*. 2013, 348.

GINI, C. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]*. Tipogr. di P. Cuppini, 1912.

GLEICK, J. *The information: A history, a theory, a flood*. Vintage, 2011.

GOLDFARB, R. *Ethics: A case study from fluency*. Plural Publishing, 2005.

GOOD, I. Diversity as a concept and its measurement: comment. *Journal of the American statistical association*. 1982, 77, 379, s. 561–563.

GOOD, I. J. The population frequencies of species and the estimation of population parameters. *Biometrika*. 1953, 40, 3–4, s. 237–264.

GUIRAUD, P. *Les caractères statistiques du vocabulaire: essai de méthodologie*. Presses universitaires de France, 1954.

HARTLEY, R. V. Transmission of information 1. *Bell System technical journal*. 1928, 7, 3, s. 535–563.

HEAPS, H. S. *Information retrieval, computational and theoretical aspects*. Academic Press, 1978.

HERDAN, G. The hapax legomenon: A real or apparent phenomenon? *Language and speech*. 1959, 2, 1, s. 26–36.

HERDAN, G. *Type-token mathematics*. 4. Mouton, 1960.

HERFINDAHL, O. C. *Concentration in the U.S. steel industry*. Disertace, Columbia University, 1950.

HESS, C. W. – HAUG, H. T. – LANDRY, R. G. The reliability of type-token ratios for the oral language of school age children. *Journal of speech, language, and hearing research*. 1989, 32, 3, s. 536–540.

HILL, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology*. 1973, 54, 2, s. 427–432.

HIRSCHMAN, A. O. *National power and the structure of foreign trade*. University of California Press, 1945.

HUFFMAN, D. A. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*. 1952, 40, 9, s. 1098–1101.

JARVIS, S. Capturing the diversity in lexical diversity. *Language learning*. 2013, 63, s. 87–106.

JARVIS, S. Grounding lexical diversity in human judgments. *Language testing*. 2017, 34, 4, s. 537–553.

JARVIS, S. – HAZHANGMOTO, B. J. How operationalizations of word types affect measures of lexical diversity. *International journal of learner corpus research*. 2021, 7, 1, s. 163–194.

JELÍNEK, T. Nové značkování v Českém národním korpusu. *Naše řeč*. 2008, 91, 1, s. 13–20.

JELÍNEK, T. et al. SYN2020: A new corpus of Czech with an innovated annotation. In EKŠTEIN, K. – PÁRTL, F. – KONOPÍK, M. (Ed.) *Text, speech, and dialogue*, s. 48–59, Cham, 2021. Springer International Publishing. ISBN 978-3-030-83527-9.

JOHNSON, W. Studies in language behavior: A program of research. *Psychological monographs*. 1944, 56, 2, s. 1–15.

JOST, L. Partitioning diversity into independent alpha and beta components. *Ecology*. 2007, 88, 10, s. 2427–2439.

JUOLA, P. Measuring linguistic complexity: The morphological tier. *Journal of quantitative linguistics*. 1998, 5, 3, s. 206–213. doi: 10.1080/09296179808590128.

JUOLA, P. Assessing linguistic complexity. *Language complexity: Typology, contact, change*. 2008, s. 89–108.

KEYNES, J. M. *A treatise on probability*. Macmillan and Company, limited, 1921.

KÖHLER, R. – GALLE, M. Dynamic aspects of text characteristics. *Quantitative text analysis*. 1993, s. 46–53.

KOJIMA, M. – YAMASHITA, J. Reliability of lexical richness measures based on word lists in short second language productions. *System*. 2014, 42, s. 23–33.

KOLMOGOROV, A. Three approaches to the quantitative definition of information. *Problems of information transmission*. 1965, 1, 1, s. 1–7. ISSN 0032-9460.

KOVÁŘÍKOVÁ, D. et al. Lexicographer's lacunas or how to deal with missing representative dictionary forms on the example of Czech. *International journal of lexicography*. 2020, 33, 1, s. 90–103.

KŘEN, M. et al. SYN2015: reprezentativní korpusů psané češtiny, 2015.

KRUPA, V. – GENZOR, J. *Písma sveta*. Obzor, 1989.

KUBÁT, M. Moving window type-token ratio and text length. *Empirical approaches to text and language analysis. Lüdenscheid: RAM*. 2014, s. 105–113.

KUBÁT, M. – ČECH, R. Thematic concentration and vocabulary richness. In *Issues in quantitative linguistics 4 (dedicated to Reinhard Köhler on the occasion of his 65th birthday)*, s. 150–159, 2016.

KUBÁT, M. – MILIČKA, J. Vocabulary richness measure in genres. *Journal of quantitative linguistics*. 2013, 20, 4, s. 339–349.

KULLBACK, S. *Statistical methods in cryptanalysis*. Aegean Press, 1976.

KULLBACK, S. – LEIBLER, R. A. On information and sufficiency. *The annals of mathematical statistics*. 1951, 22, 1, s. 79–86.

KYLE, K. – JARVIS, S. – CROSSLEY, S. The tool for the automatic analysis of lexical diversity (TAALED), 2018. Dostupné z: <https://www.linguisticanalysistools.org/taaled.html>.

KYLE, K. Measuring lexical richness. In *The Routledge handbook of vocabulary studies*. London: Routledge, 2019. s. 454–476.

KYLE, K. – CROSSLEY, S. A. – JARVIS, S. Assessing the validity of lexical diversity indices using direct judgements. *Language assessment quarterly*. 2021, 18, 2, s. 154–170.

LAUTENBACHER, S. – MOLTNER, A. – STRIAN, F. Psychophysical features of the transition from pure heat perception to heat pain perception. *Perception & psychophysics*. 1992, 52, 6, s. 685–690.

LEVENSHTEIN, V. Binary codes capable of correcting spurious insertions and deletion of ones. *Problems of information transmission*. 1965, 1, 1, s. 8–17.

LEVENSHTEIN, V. I. – OTHERS. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*. 1966, 10, 8, s. 707–710.

LEXIS, W. II. Ueber die Theorie der Stabilität statistischer Reihen. *Jahrbücher für Nationalökonomie und Statistik*. 1879, 32, 1, s. 60–98. ISSN 0021-4027.

LIDSTONE, G. J. Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*. 1920, 8, 182–192, s. 13.

LOZANO, A. et al. Fast calculation of entropy with Zhang's estimator. *arXiv preprint arXiv:1707.08290*. 2017.

MACARTHUR, R. H. Patterns of species diversity. *Biological reviews*. 1965, 40, 4, s. 510–533.

MACWHINNEY, B. *The CHILDES project: Tools for analyzing talk. transcription format and programs*. 2. Psychology press, 2000.

MALVERN, D. et al. *Lexical diversity and language development*. Springer, 2004.

MANN, M. B. III. The quantitative differentiation of samples of written language. *Psychological monographs*. 1944, 56, 2, s. 39.

MARNEFFE, M.-C. d. et al. Universal dependencies. *Computational linguistics*. 2021, 47, 2, s. 255–308.

MCCARTHY, P. M. – JARVIS, S. vocd: A theoretical and empirical evaluation. *Language testing*. 2007, 24, 4, s. 459–488.

MCCARTHY, P. M. – JARVIS, S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*. 2010, 42, 2, s. 381–392.

MCCASKEY, J. P. History of 'temperature': maturation of a measurement concept. *Annals of Science*. 2020, 77, 4, s. 399–444.

MICHALKE, M. et al. Package 'koRpus', 2021. Dostupné z: <www.maths.bristol.ac.uk/R/web/packages/koRpus/koRpus.pdf>.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 2013.

MILIČKA, J. *Teorie komunikace jakožto explanatorní princip přirozené víceúrovňové segmentace text*. Disertace, Ph. D. thesis, Charles University, Prague, Czech Republic, 2016.

MILIČKA, J. average word length from the diachronic perspective: the case of Arabic. *Linguistic frontiers*. 2018, 1, 2, s. 81–89.

MILIČKA, J. *Lexikální diverzita*. Habilitační práce, Univerzita Karlova, 2022.

MILIČKA, J. Rank-frequency relation & type-token relation: Two sides of the same coin. In IVAN OBRADOVIĆ, E. K. – KÖHLER, R. (Ed.) *Methods and applications of quantitative linguistics — Selected papers of the 8th International Conference on quantitative linguistics (QUALICO)*, s. 163–171. Academic mind, 2013. ISBN 978-86-7466-465-0.

MILLER, G. A. et al. Introduction to WordNet: An on-line lexical database. *International journal of lexicography*. 1990, 3, 4, s. 235–244.

MITCHELL, D. Type-token models: a comparative study. *Journal of quantitative linguistics*. 2015, 22, 1, s. 1–21.

NIVRE, J. et al. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth international conference on language resources and evaluation (LREC'16)*, s. 1659–1666, 2016.

NIVRE, J. et al. Universal dependencies 2.1, November 2017. Dostupné z: <https://hal.inria.fr/hal-01682188>. LINDAT/CLARIN digital library at the Institute of formal and applied linguistics (ÚFAL), Faculty of mathematics and physics, Charles university.

OCELÁKOVÁ, Z. – BOŘIL, T. slabikovacAR, 2020. Dostupné z: <https://fonetika.ff.cuni.cz/wp-content/uploads/sites/104/2020/07/slabikovacAR.zip>.

OTTO, H. A. The paternity of an index. *American economic review*. 1964, 54, 5, s. 761.

PATIL, G. – TAILLIE, C. Diversity as a concept and its measurement. *Journal of the American statistical association*. 1982, 77, 379, s. 548–561.

PEIRCE, C. S. *Collected papers of Charles Sanders Peirce*. Harvard University Press, 1931.

PELEGRINOVÁ, K. et al. MorfoCzech, 2021. Dostupné z: <http://hdl.handle.net/11234/1-4626>. LINDAT/CLARIAH-CZ digital library at the Institute of formal and applied linguistics (ÚFAL), Faculty of mathematics and physics, Charles university.

PENNINGTON, J. – SOCHER, R. – MANNING, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, s. 1532–1543, 2014.

PETKEVIČ, V. – OTHERS. Problémy automatické morfologické disambiguace češtiny. *Naše řeč*. 2014, 97, 4-5, s. 194–207.

PIANTADOSI, S. T. – TILY, H. – GIBSON, E. Word lengths are optimized for efficient communication. *Proceedings of the National academy of sciences*. 2011, 108, 9, s. 3526–3529.

PIOTROVSKAJA, A. A. – PIOTROVSKIJ, R. G. Matematičeskie modeli vdiachronii i tekstoobrazovanii. In *Statistika reči i avtomatičeskij analiz teksta*. Leningrad: Nauka, 1974. s. 361–400.

POIRET, R. – LIU, H. Mastering the measurement of text's frequency structure: an investigation on Lambda's reliability. *Glottometrics*. 2017, 37, s. 82–100.

POPESCU, I.-I. – MAČUTEK, J. – ALTMANN, G. Word forms, style and typology. *Glottotheory*. 2010, 3, 1, s. 89–96.

POPESCU, I.-I. – ČECH, R. – ALTMANN, G. *The lambda-structure of texts*. Ram-Verlag Lüdenscheid, 2011.

POPPER, K. R. *The logic of scientific discovery*. Taylor & Francis, 2005. ISBN 0-203-99462-0.

PROISL, T. Text complexity, 2021. Dostupné z: <https://github.com/tsproisl/textcomplexity>.

RAO, C. R. Diversity and dissimilarity coefficients: A unified approach. *Theoretical population biology*. 1982, 21, 1, s. 24–43. ISSN 0040-5809.

RAPAPORT, W. J. A history of the sentence 'Buffalo buffalo buffalo Buffalo buffalo.'. *Università di Buffalo computer science and engineering*. 2014.

REJEWSKI, M. How Polish mathematicians deciphered the Enigma. *Annals of the history of computing*. 1981, 3, 3, s. 213–234.

RÉNYI, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley symposium on mathematical statistics and probability, Volume 1: Contributions to the theory of statistics*, 4, s. 547–562. University of California Press, 1961.

ROCCHINI, D. et al. From zero to infinity: minimum to maximum diversity of the planet by spatio-parametric Rao's quadratic entropy. *bioRxiv*. 2021. doi: 10.1101/2021.01.23.427872. Dostupné z: <https://www.biorxiv.org/content/early/2021/01/25/2021.01.23.427872>.

ROUSSEAU, R. The repeat rate: from Hirschman to Stirling. *Scientometrics*. 2018, 116, 1, s. 645–653.

SCHMIDT, D. Package sylcount, 2022. Dostupné z: <https://github.com/wrathematics/sylcount>.

SCHMIDT, M. – LIPSON, H. Distilling free-form natural laws from experimental data. *Science*. 2009, 324, 5923, s. 81–85. doi: 10.1126/science.1165893.

SCHWEIKER, M. et al. Challenging the assumptions for thermal sensation scales. *Building research & information*. 2017, 45, 5, s. 572–589.

SCOTT, M. WordSmith tools help. *Liverpool: Lexical Analysis Software*. 2010.

SHANNON, C. E. A mathematical theory of communication. *Bell system technical journal*. 1948, 27, 3, s. 379–423. ISSN 1538-7305.

SHANNON, C. L. – WEAVER, W. *The mathematical theory of communication*. University of Illinois Press, 1949.

SHI, Y. – LEI, L. Lexical richness and text length: An entropy-based perspective. *Journal of quantitative linguistics*. 2022, 29, 1, s. 62–79.

SIMPSON, E. H. Measurement of diversity. *nature*. 1949, 163, 4148, s. 688–688.

SINKOV, A. *Elementary cryptanalysis — A mathematical approach*. Random House, 1968.

SOMERS, H. H. *Analyse mathématique du langage: lois générales et mesures statistiques*. 1. Editions Nauwelaerts, 1959.

SPELLERBERG, I. F. – FEDOR, P. J. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener'Index. *Global ecology and biogeography*. 2003, 12, 3, s. 177–179.

STARK, P. B. – SALTELLI, A. Cargo-cult statistics and scientific crisis. *Significance*. 2018, 15, 4, s. 40–43.

STEADMAN, R. G. A universal scale of apparent temperature. *Journal of applied meteorology and climatology*. 1984, 23, 12, s. 1674–1687.

STRAKA, M. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared task: Multilingual parsing from raw text to Universal dependencies*, s. 197–207, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-2020. Dostupné z: <https://www.aclweb.org/anthology/K18-2020>.

STRAKOVÁ, J. – STRAKA, M. – HAJIC, J. Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual meeting of the Association for computational linguistics: System demonstrations*, s. 13–18, 2014.

TALEB, N. N. Fooled by correlation: Common misinterpretations in social 'science'. *Draft*. 2019. Dostupné z: <https://www.academia.edu/39797871/Fooled_by_Correlation_Common_Misinterpretations_in_Social_Science_>.

TULDAVA, J. J. Tuldava's bibliography. *Journal of quantitative linguistics*. 1997, 4, 1-3, s. 8–12. doi: 10.1080/09296179708590073.

URE, J. Lexical density and register differentiation. *Applications of linguistics*. 1971, 443452.

ČECH, R. *Tematická koncentrace textu v češtině*. 15 / *Studies in computational and theoretical linguistics*. ÚFAL, 2016. ISBN 978-80-88132-00-4.

WETZEL, L. Types and tokens. In ZALTA, E. N. (Ed.) *The Stanford encyclopedia of philosophy*. Stanford: Metaphysics research lab, Stanford university, Fall 2018 edition, 2018.

WIENER, N. *Cybernetics or control and communication in the animal and the machine*. John Wiley and sons, 1948.

WIMMER, G. – ALTMANN, G. On vocabulary richness. *Journal of quantitative linguistics*. 1999, 6, 1, s. 1–9.

YUJIAN, L. – BO, L. A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*. 2007, 29, 6, s. 1091–1095. doi: 10.1109/TPAMI.2007.1078.

ZEMÁNEK, P. – MILIČKA, J. *Words lost and found: The diachronic dynamics of the Arabic lexicon*. RAM-Verlag, 2017.

ZHANG, Z. Entropy estimation in Turing's perspective. *Neural computation*. 2012, 24, 5, s. 1368–1389.

ZIPF, G. K. *The psycho-biology of language: An introd. to dynamic philology*. Mifflin, 1935.

Anotace na obálce

Tato monografie pojímá téma lexikální diverzity v celé jeho složitosti a z kalného kvasu metrik a přístupů snaží se vydestilovat lexikální diverzitu jako smysluplnou, intersubjektivní a dobře interpretovatelnou lingvistickou veličinu s jednoznačně definovanou jednotkou a intuitivním škálováním.

Ať už měříme lexikální diverzitu kvůli aplikovanému výzkumu, či s cílem objevovat lingvistické zákony a stavět teorie, kniha vede k pevnějšímu metodologickému ukotvení.