Univerzita Karlova

Filozofická fakulta

Ústav českého jazyka a teorie komunikace

Diplomová práce

Bc. Jan Henyš

Registrová variabilita českých internetových textů

Register Variability of Czech Internet Texts

Praha 2019

Vedoucí práce: doc. Mgr. Václav Cvrček, Ph.D.

Děkuji vedoucímu práce, doc. Mgr. Václavu Cvrčkovi, PhD., za podnětné rady a konzultace v přátelském duchu. Poděkování patří také Mgr. Martině Rybové za cenné poznámky k jazykové podobě textu.

Prohlašuji, že jsem diplomovou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 18. 7. 2019

……

Jan Henyš

Abstrakt

Cílem této práce je analyzovat registrovou variabilitu českých internetových textů. Použitá metoda je založena na multidimenzionálním přístupu, který do lingvistiky ve svých pracích uvedl Douglas Biber. Úvodní část práce popisuje výrazné rysy internetových textů a uvádí jejich příklady. Následující část se věnuje přehledu výzkumu jazykové variace a detailně popisuje postup multidimenzionální analýzy. Praktická část práce se snaží o multidimenzionální analýzu dat získaných ze vzorku českojazyčného webového korpusu za využití modelu implementace multidimenzionální analýzy na českém jazykovém materiálu. Na základě výsledků jsou charakterizovány jednotlivé internetové subregistry.

Abstract

The aim of this thesis is to analyse register variation among Czech internet texts. The method is based on multi-dimensional approach introduced by Douglas Biber. The initial part of the thesis describes various salient features of internet-based texts with their characteristics and examples. The following part offers an overall description of language variation research. The multi-dimensional analysis is then thoroughly described step-by-step. The practical part of the thesis pursues a multi-dimensional analysis of data sample obtained from a web-crawled corpus of Czech language, following the method of implementation of multi-dimensional analysis done on Czech language material. On the basis of the results, the characteristics of internet sub-registers are proposed.

Klíčová slova

registr, multidimenzionální analýza, variabilita, internet

Keywords

register, multi-dimensional analysis, variability, internet

Obsah

Úvod

Tato práce si klade za cíl zmapovat variabilitu textů nacházejících se na českém internetu pomocí korpusové metody, kterou do lingvistiky přinesl americký jazykovědec Douglas Biber. V českém diskurzu by tato metoda zvaná multidimenzionální analýza spadala do oblasti stylistiky. Zároveň se ale tradičním stylistickým měřítkům vymyká, a to zejména svým kvantitativním charakterem a velkým množstvím jazykových dat, která zpracovává pomocí statistických metod.

V teoretické části se práce zabývá internetovým jazykem a jazykovou variabilitou. Popisuje jazyková specifika internetových komunikátů, která jsou již několik desítek let předmětem lingvistických výzkumů. V kapitole o jazykové variabilitě se pojednává o jejích teoriích relevantních pro tuto práci, a nakonec nabízí popis samotného postupu multidimenzionální analýzy s ohledem na metodologická úskalí, která ji doprovázejí. Pozornost je věnována průběhu a výsledkům dvou pro tuto práci stěžejních výzkumů, a to analýze registrů textů na anglofonním internetu a prvnímu ucelenému pokusu o provedení multidimenzionální analýzy na českém jazykovém materiálu.

Praktická část se zabývá analýzou vzorku dat pocházejících z webového korpusu. Tato analýza je zamýšlena jako pilotní výzkum, může tedy posloužit jako podklad budoucího výzkumu provedeného ve větším měřítku. Každý z textů obsažených ve vzorku je v rámci anotace kategorizován. Díky hodnotám poskytnutým multidimenzionální analýzou je možné kategorie popsat novou metodou, jejíž výsledky mohou přispět k poznání jazykového prostoru internetu, nových internetových žánrů i samotného internetu jako média. Výsledky předložené v této práci také mohou pomoci ověřit funkčnost zvolené metody, případně pomoci v jejích budoucích úpravách.

1 Jazyk internetu

1.1 Internet jako médium

Internet je systém počítačových sítí s celosvětovým rozšířením, který byl vyvinut v 60. letech 20. století. Po zpřístupnění veřejnosti rapidně vzrostla jeho popularita. Zejména díky podpoře služby WWW (akronym pro *World Wide Web*, v České republice dostupná od roku 1992), která umožňuje prohlížení webových stránek pomocí webového prohlížeče, se stal zásadní platformou pro publikaci a komunikaci. Označuje se jako jedno z tzv. nových médií. Vzhledem k tomu, že internet je pro většinu běžných uživatelů poměrně abstraktní pojem,[1] označuje toto slovo ne vždy to samé. Velmi obvyklé je ztotožnění pojmu *internet* s výše zmiňovanou službou WWW, tedy službou umožňující uživatelům přístup na konkrétní webové stránky, případně s webovými stránkami samotnými.

Naopak Lelia Greene (2010, s. 3) pojímá ve své publikaci internet jako zastřešující pojem. Pod něj spadá kromě internetu jakožto sítě a samotných stránek také několik dílčích aspektů a konceptů. Prvním z nich je architektura[2] a software s různou měrou otevřenosti zdroje (autorka uvádí jako příklady internetové prohlížeče, vyhledávač a internetovou encyklopedii *Wikipedia*). Dále pak užití počítačových i přirozených jazyků zpřístupňujících internet lidem všech kultur a lidem všech stupňů způsobilosti. Poté Greene zmiňuje konkrétní služby, jako je e-mail, chat a jiné způsoby komunikace v reálném čase (tzv. *instant messaging*), blogy a sociální sítě. Neopomíjí ani počítačové hry spojené s internetem, komunity, které kolem nich vznikají, a fikční světy, které jsou jejich tvůrci a hráči vytvářeny. Konečně pak pod pojem internet zařazuje i všechny aspekty, díky kterým se komunikace zprostředkovaná internetem stává zdomácnělou a ovlivňuje lidskou každodennost.

[1] Poměrně časté je referování k internetu jakožto konceptuálnímu prostoru, jakési „nádobě na informace“. Obsah dostupný prostřednictvím sítě internet je pak označován jako jsoucí *na internetu*, dříve i *v internetu*.

[2] Způsob uspořádání technického a programového vybavení v počítači (ASCS).

1.2 Computer mediated communication

V odborném lingvistickém diskurzu má výzkum jazykových prostředků nacházejících se na internetu poměrně dlouhou tradici. Susan C. Herring (1996, s. 3) uvádí, že za první práci spadající svou tematikou do této oblasti může být považována již *The Network Nation* z konce 70. let 20. století (Hiltz–Turoff, 1978). K rozmachu výzkumu tématu, které se označuje jako *computer-mediated communication* (CMC) nebo *computer-mediated discourse* (CMD) však došlo až v letech osmdesátých, kdy byly publikovány stěžejní studie reflektující nové prostředky, jež s sebou přineslo nové médium internetové komunikace. Jak je zjevné, těžištěm těchto výzkumů, jejichž kvantita dosáhla vrcholu v devadesátých letech, jsou právě prostředky, které jsou pro internetovou komunikaci specifické a v době před internetem nebyly v jazyce přítomny. Mezi ty patří zejména různé způsoby ozvláštňování písemného projevu pomocí upravené typografie a ortografie, užívání emotikonů, akronymů charakteristických pro CMC atd. Dále se výzkumy zaměřují na slovotvorné postupy při označování nové reality spojené s počítači a internetem, dále např. na slovní zásobu členů různých komunit uživatelů nebo správců počítačových sítí, někdy je tématem i syntax, která bývá ovlivněna např. specifickým formátem některých druhů internetových komunikátů, jako jsou např. *tweety*, příspěvky na sociální síti Twitter, které jsou omezeny na 280 znaků včetně mezer. Významnou měrou jsou zastoupeny také studie, které jsou věnovány analýze internetového diskurzu pomocí „tradičních“ diskurzněanalytických metod.

Všechny tyto (a další) prostředky dohromady utvářejí samostatnou jazykovou varietu, která bývá nazývána *chatspeak* nebo *netspeak*;[3] její výzkum bývá vyčleňován tematicky jako CMC nebo CMD, někdy je ovšem vymezen i jako samostatná lingvistická disciplína zvaná *internet linguistics* (Crystal, 2011) nebo *netlinguistics* (Posteguillo, 2008). Vymezení CMC jako samostatné variety je problematické a různí lingvisté a lingvistky k němu zaujímají různé

[3] Tyto pojmy jsou ne vždy užívány synonymně, někdy jsou rozlišovány podle motivace na základě domény: *netspeak* jako varieta užívaná na internetu obecně, *chatspeak* pak jako varieta užívaná v chatu.

postoje. Podrobně se klasifikaci CMC věnuje např. Susan C. Herring ve studii *A Faceted Classification Scheme for Computer-Mediated Discourse* (2007).

Mnoho článků z oblasti CMC bylo uveřejněno i v českém lingvistickém prostředí. Příkladem je vydání popularizačně-vědeckého časopisu *Čeština doma a ve světě* (2006) s podtitulem *Čeština na internetu*. V tomto vydání byl publikován např. článek Světly Čmejrkové *E-čeština*, který se zabývá nejen specifiky internetové češtiny, ale také tím, jak se na internetu píše o češtině (Čmejrková, 2006). Prefixoid *e-* v názvu článku zkracuje adjektivum *elektronický*, e-čeština tedy označuje češtinu užívanou v elektronické komunikaci, což zhruba odpovídá definici CMC. Susan C. Herring užívá pro popis jazykových prostředků specifických pro CMC podobně utvořený termín *e-grammar* (2011).

1.3 Specifické prostředky v internetové komunikaci

Jak již bylo řečeno, specifika komunikace v internetovém prostředí zasahují všechny jazykové roviny. Na každé z nich se pak vyskytují jevy zasluhující odbornou pozornost. Následující část práce se věnuje nejvýraznějším z nich.

1.3.1 Ortografie a typografie

Fonologická rovina je v internetové komunikaci upozaděna, protože internetové komunikáty (alespoň ty, na něž se soustřeďuje tato práce) existují zpravidla v psané podobě: „V textově zaměřené CMC je fonologie do značné míry irelevantní; typografie a ortografie přejímají funkce, které má jinak zvuk." (Herring, 2011, s. 1, přel. JH)

Typografické zvláštnosti v rámci CMC spočívají zejména v možnosti využívání nepísmenných a tzv. speciálních znaků, využívání majuskulí a minuskulí v různých způsobech (psaní celého textu majuskulemi pro jeho zdůraznění nebo střídání velkých a malých písmen pro specifické expresivní zabarvení) (tamtéž, s. 2). Funkčně může být měněn řez písma (tučný nebo kurzívní) a jeho zdobení (podtržení, škrtnutí apod.) a další typografické modifikace. Významné jsou také odlišnosti práce s interpunkčními znaménky, zejména jejich nadužívání a opakování. Opakování se týká také grafémů, často

těch, které je možné prodloužit i při jejich vyslovení: samohlásek (*ahooooooj*) a souhlásek trvacích (*nevíšššš*), ale nejen jich. Pisatelé se k němu uchylují zpravidla tehdy, když chtějí svému textu dodat emocionální nádech (Jílková, 2017). V angličtině (oproti češtině) je dále poměrně rozšířená kreativní substituce číslic za písmena buď na základě zvukové podobnosti (např. *gr8* — číslice 8 — *eight*, vysloveno [eɪt] — tvoří s grafémy *gr* slovo *great*, vysloveno /greɪt/) nebo na základě podobnosti grafické, pro niž se vžil termín *leetspeak* nebo *Leet* (ze záměny slova *leet* za *1337*). Tomuto způsobu záměny je z důvodu využívání v hackerských komunitách připisována i kryptografická funkce (tamtéž). V češtině nejsou tyto prostředky časté, přesto se někdy objevují. Příklady uvádí například Lucie Hašová v článku věnovaném jazyku SMS zpráv: „podobně jazykově úsporné dokáží být i české výrazy nebo celé věty, např. *nej1dussi* (nejjednodušší), *jsem o5 z5* (jsem opět zpět), *jdeme na Pe3n* (jdeme na Petřín), běžné je zjednodušení souhláskových skupin *ks* a *kv*, např. *jaxe mas* (jak se máš), *qasnice* (kvasnice), *qetina* (květina), nacházíme různě důmyslné úsporné česko-anglické kombinace, např. *neska to nego* (dneska to nejde), *taxory* (tak sorry). [...] Jazyk obsahující zmíněné česko-anglické zkratky, emotikony apod. se někdy nazývá internetí mluva, netí mluva nebo netština.[4]“ (Hašová, 2002, s. 208)

Značnou lingvistickou pozornost vyvolalo rozšíření emotikonů, tzv. smajlíků. Ty jsou tvořeny zejména skládáním různých písmenných i nepísmenných znaků tak, aby připomínaly lidské obličeje a v nich „různé odstíny radosti, smutku, hněvu, údivu atp.“ (Hoffmanová et al., 2016, s. 113), příp. i předměty, činnosti apod. Původně byly tvořeny pouze znaky základní znakové sady ASCII, později začaly být používány i znaky rozšířených znakových sad, jako je tomu např. u emotikonu ლ(ಠ 益 ಠლ) vyjadřujícího hněv pomocí znaků gruzínské, kannadské a čínské znakové sady (relativně nový NESČ například rozšířené znakové sady opomíjí (Jílková, 2017)). Nejčastěji užívané emotikony jsou některými textovými editory a jinými programy konvertovány do grafické podoby převedením na jiný znak standardu Unicode

[4] Uvedená označení pro jazykové prostředky na internetu jsou značně problematická, např. po vyhledání přesného řetězce „netí mluva“ vrátí internetový vyhledávač Google (k 6. 6.2019) pouhé tři doklady, z nichž všechny odkazují právě ke článku Hašové.

(nebo na emoji — viz dále). Užitím emotikonu se text stává multimodálním, sémiotická heterogennost je přítomna přímo v rámci textu, nikoli jen jako jeho doprovod. Tato skutečnost vede k úvahám o sémantické náplni emotikonů a jejich roli v textu. Tradičně se jim ve studiích, které se jich tematicky dotýkají, přisuzuje pouze schopnost do jisté míry nahrazovat vizuální stránku mluvené komunikace a zprostředkovat tak informace, které nejsou účastníkům CMC dostupné. Jako jejich primární funkce je tak obvykle stanovena signalizace emocí autora textu (Čmejrková, 1997; Hašová, 2002; Hoffmannová et al., 2016, a další). Studie z novější doby však poukazují na to, že role emotikonů v textu je mnohem širší.[5] Dobrým příkladem je studie Eliho Dresnera a Susan C. Herring *Functions of the Nonverbal in CMC: Emoticons and Illocutionary Force* (2010). V té jsou analyzovány emotikony skrze teorii řečových aktů. Dresner a Herring na jejím základě poukazují na to, že emotikony mohou také sloužit jako indikátor ilokuční síly (Dresner–Herring, 2010). Kromě toho zpochybňují zkratkovité pojetí emotikonu pouze jako prostředku pro vyjádření emocí (tamtéž, s. 252).

Kolem roku 2010 vzrostla popularita tzv. emoji (v české transkripci někdy jako emodži), které mají blízko k emotikonům, ale jsou primárně reprezentovány nikoli sekvencí znaků, ale jediným znakem, který je standardizovaný (nejčastěji standardem Unicode) a v prohlížečích se zobrazuje v grafické, zpravidla barevné podobě. Výběr emoji je v současné době značný, k dispozici jich je přes tisíc.[6] Kromě obličejů s různým výrazem mohou emoji vyobrazovat i tvary rukou, dále ovšem také předměty, potraviny, rostliny, počasí, činnosti, velké množství informativních piktogramů, vlajky apod. Jejich použití je obdobné jako u emotikonů, díky své diverzitě ale umožňují komplexnější vyjádření. Jejich prostřednictvím lze vyjádřit souhlas či nesouhlas, úmysly, postoje, podpořit jimi textové vyjádření, posilovat některé konotace textového vyjádření, vyjádřit ticho nebo v krajních případech nahradit celý text (Zhou et al., 2017, s. 752).

[5] Je pravděpodobné, že způsob užívání emotikonů prošel určitým vývojem rozšiřujícím jejich použitelnost v textu.

[6] V současnosti je aktuální verze 12.0, v níž mají emoji 1273 unikátních reprezentací. (dostupné z <https://unicode.org/emoji/charts/index.html> k 9. 6. 2019)

1.3.2 Lexikon a slovotvorba

Vzhledem k tomu, že někteří lingvisté a lingvistky hovoří v souvislosti s CMC o internetovém slangu (např. Lišková, 2006), je zjevné, že jazyková specifika CMC zasahují i do rovin slovotvorby a lexika.

Lexikální rovina internetové komunikace v češtině se vyznačuje zvýšeným užíváním anglicismů, „jež jsou [...] součástí prostředí nových technologií jako takových, neboť zařízení, která elektronickou komunikaci umožňují (počítače, mobilní telefony), byla vyvinuta v anglojazyčném prostředí“ (Jílková, 2017). Lexikální přejímky někdy zůstávají v nesklonné podobě, často se však adaptují, např. pomocí produktivních afixů, např. *olajkovat* (z angl. *like*), *odenterovat* (z angl. *enter*), *rebootnout* (z angl. *reboot*), vyskytují se i přepisy zvukové podoby slov, např. *flejm*, *flejmovat* (z angl. *flame*, ve významu ‚hádka na internetové diskusi nebo chatu‘).

I v angličtině vzniká s příchodem elektronické komunikace mnoho nových slov, která jsou velkou měrou přejímána do češtiny, a to buď pomocí afixace (např. *e-*, *hyper-*, *cyber-*), blendingem, konverzí nebo sémantickým posunem (např. *spam* ve významu ‚nevyžádaná pošta nebo zpráva‘, původně označuje masovou konzervu)(Herring, 2011, s. 4).

Podstatným a frekventovaně užívaným prvkem jsou různé způsoby zkracování slov, ať už iniciálové zkratky, nebo jinak zkrácená slova. Podobně jako je tomu u lexikální roviny obecně, i zde se objevují prostředky původní (např. *jj* – ‚jo jo‘, *mmch* – ‚mimochodem‘) a zkratky přejaté z angličtiny (např. *btw* z angl. *by the way*, ‚mimochodem‘)(Jílková, 2017). Zkracování je motivováno nutností omezit rozsah textu, který je dán typem komunikátu (např. SMS zpráva nebo příspěvek na sociální síti Twitter), případně snahou účastníků komunikace ušetřit čas a námahu při psaní zprávy (např.SMS zprávy na alfanumerické klávesnici mobilního telefonu).

1.3.3 Syntax a text

V oblasti syntaxe se některé komunikáty v CMC vyznačují fragmentární syntaxí, velkým množstvím elizí, časté je připodobnění k syntaxi telegrafické studie (Herring, 2011, s. 5). K elizím dochází z týchž důvodů jako k používání

zkratek a jiných zkrácenin, ale také kvůli přibližování některých psaných internetových komunikátů mluvenému jazyku (Čmejrková, 1997; Crystal, 2006, a další).

1.4 Psanost vs. mluvenost

Zvláštní postavení v rámci textu mají performativní vyjádření signalizovaná ohraničením asterisky „*“ nebo špičatými závorkami „<>“. Těmito úseky textu se mluvčí snaží nahradit složky neverbální komunikace, které přirozeně nejsou v CMC přítomny, jako je kinesika, proxemika (Crystal, 2006, s. 42), nebo i haptika. Mluvčí k sobě v těchto výpovědích referují ve třetí osobě, případně používají deverbativní substantiva, adjektiva, někdy i syntakticky komplexnější konstrukce. Crystal (tamtéž) uvádí příklady jako <smirk> (‚úšklebek‘) nebo <laugh> (‚smích‘), Herring zmiňuje příklady *waves* (‚mává‘), *confused* (‚zmatený/zmatená‘) nebo *in a bad mood* (‚má špatnou náladu‘). Některá z těchto vyjádření jsou konvencionalizována, objevují se i zkratky typické pro CMC, např. <g> jako zkratka *grin* ‚škleb‘ nebo krypticky vyhlížející <gd&r> ve významu *grinning, ducking and running* ‚šklebí se, shýbá se a utíká‘[7]. David Crystal k těmto prostředkům ve vztahu k mluveným rysům CMC uvádí:

> „Tyto rysy netspeaku se vyvinuly jako způsob, jakým se lze vyhnout víceznačnosti a nedorozuměním, jež přicházejí, když musí psaný jazyk nést tíhu řeči. Jsou to odvážné pokusy, ale netspeak v celkovém pohledu nedisponuje žádnou opravdovou schopností vyjadřovat významy pomocí kinesiky a proxemiky, což jej také kvůli neschopnosti vyjadřovat prozodické rysy značně odlišuje od mluveného jazyka. Nepřítomné jsou i další jazykové rysy typické pro mluvený jazyk, což

[7] Podle příspěvku uživatele SiteNook z 19. 10. 2010 v on-line slovníku *Urban Dictionary* tvořeném uživateli internetu je zkratka <gd&r> použita uživatelem, který předpokládá, že po něm budou lidé po publikování příspěvku házet předměty (dostupné z <https://www.urbandictionary.com/define.php?term=<gd%26r>>, cit. 9. 6. 2019). Dochází zde tak k metaforickému přenesení publikování příspěvku ve virtuálním prostoru na promluvy v reálném prostředí a reakce na ně.

při snaze užívat jazyk na internetu stejně jako v mluvené konverzaci způsobuje ještě více obtíží.“ (Crystal, 2006, s. 42–43, překl. JH)

Společným rysům CMC, konkrétně komunikaci prostřednictvím e-mailu, se věnuje Světla Čmejrková. Ve studii *Čeština v síti: Psanost či mluvenost? (O stylu e-mailového dialogu)* (1997) systematicky vyhledává styčné plochy mluveného jazyka a jazyka textů posílaných prostřednictvím e-mailu. Dochází k závěru, že se e-mailová komunikace jeví „jako povrchově psaná a hloubkově mluvená“ (Čmejrková, 1997, s. 246). S mluveným jazykem totiž sdílí rysy dynamické procesuální interakce založené na společně sdíleném kontextu (tamtéž). Naopak rysy mluvenosti jsou potlačeny kvůli distanci mezi účastníky komunikace, což se např. projevuje obsažením informace o tom, na které sdělení z předchozí zprávy účastník komunikace reaguje, což je prostředek typický pro psané komunikáty jako je např. polemika (tamtéž, s. 247). Podobný názor na ambivalentní charakter e-mailové komunikace prezentuje Lauren Squires, která zmiňuje označení „written speech“ (‚psaná řeč‘). Ve studii *Enregistering Internet Language* (2010) pojednává o dvojakosti CMC obecně:

„Technické faktory, jimž bývá typicky připisována hraniční povaha CMC, jsou omezujícími a zpřístupňujícími aspekty textově založeného dialogu, zvláště **asynchronie** v některých aplikacích, jako je e-mail (kde je časová prodleva mezi vyměněnými zprávami), **synchronie** v jiných aplikacích, jako je instant messaging (IM), kde se účastníci snaží o výměnu zpráv rychlostí úměrnou rychlosti výměny při přirozené konverzaci tváří v tvář, a **modalita** strojově psaného textu napříč prakticky všemi užitími.“ (Squires, 2010, s. 462, překl. a zvýr. JH)

1.5 Jazyk internetu vs. jazyk na internetu

Internetovému jazyku bylo v posledních desetiletích věnováno mnoho prostoru, zaměření většiny studií, zejména v 90. letech a na přelomu tisíciletí, se ovšem omezuje pouze na jeho nejnápadnější součást, tedy soubor aspektů, jimiž se komunikáty publikované (nebo se jinak objevující) na internetu liší od

běžného psaného či mluveného jazyka. Mnohdy ale autoři a autorky zapomínají na komplexitu média, s nímž pracují. Internet je médium velmi dynamické, jeho obsah se mění v čase a je nestálý. Obsah, který byl na internetu dostupný v minulosti, už nemusí existovat, naopak každou sekundu je obsah aktualizován, obohacován či jinak měněn, nebo vzniká obsah zcela nový. Vzhledem k tomu, že se v posledních desetiletích stává internet velice důležitou platformou jak pro jazykovou percepci, tak jazykovou produkci, je přirozené, že se kromě nových internetových žánrů (které jsou specifické pro internet a jejich existence je do jisté míry definována právě platformou, na níž se nalézají), dostávají na internet i ekvivalenty žánrů z tradičních médií, jako jsou např. publicistické texty. Skutečnosti, že se internet stává zdrojem nové žánrové diverzifikace, si všímají např. Shepherd a Watters, kteří navrhují koncept nového, zastřešujícího žánru komunikátů spojených s internetem — kyberžánru (*cybergenre*): „Kyberžánr definujeme jako dvě hlavní třídy subžánrů: stávající a nové. Třída stávajících subžánrů sestává z těch, které jsou založeny na žánrech existujících na ostatních médiích, jako je papír nebo video, a které se přesunuly na toto nové médium. Třída nových subžánrů sestává z těch, které se vyvinuly v rámci tohoto nového média a na jiném médiu nemají žádný protějšek.“ (Shepherd–Watters, 1998, s. 2, překl. JH) Uvádějí také výzkum, jímž bylo zjištěno, že kolem 80 procent internetového obsahu (čímž je míněn obsah webový, tj. internetové stránky) tvoří právě žánry původní (stávající), jejichž charakter byl pro použití na internetu replikován. Samozřejmostí je jistá míra přizpůsobení médiu. Internetové komunikáty jsou adaptovány různou měrou, od nezaznamenatelného přizpůsobení až po pouhý okrajový průnik s původním žánrem (tamtéž, s. 3). Na základě této škály je vytvořena jemnější kategorizace stávajících subžánrů na žánry replikované a variantní. Replikované žánry jsou hodnověrnými reprodukcemi komunikátů v podobě, v které se objevují na zdrojovém médiu. Funkčnost se nijak zvlášť nemění. Oproti tomu variantní žánry taktéž vycházejí z již existujících žánrů, ale vyvinuly se do nové podoby za využití nových prostředků, které poskytuje nové médium. Jako příklad uvádějí text s hypertextovými odkazy, který disponuje obrazovým nebo audiovizuálním doprovodem (tamtéž).

Kromě aktualizace internetových žánrů vlivem média dochází také k jejich mísení. Bližší pozornost je tomuto tématu věnována ve studii Mariny Santini *Characterizing Genres of Web Pages: Genre Hybridism and Individualization* (2007). Nově vznikající žánry přebírají charakteristiky již existujících žánrů, poté se adaptují na nové prostředí a začnou se autonomně vyvíjet. Do tohoto procesu vstupují jiné žánry a dochází k mísení. Tento proces hybridizace je velmi jednoduchý a rychlý, k rozrůzňování a individualizaci žánrů přispívá i fakt, že webové prostředí není nijak institucionálně kontrolováno:

> „Zaprvé se náležitost k žánru vzájemně nevylučuje, různé žánry mohou splývat v jediném dokumentu a tvořit hybridní formy. Zadruhé žánry umožňují jistou svobodu ve variaci, díky čemuž mohou být individualizovány. A konečně, žánrové repertoáry jsou dynamické, tj. mění se v průběhu času, čímž podněcují změny jednotlivých žánrů a jejich evoluci.“ (Santini, 2007, s. 6, překl. JH)

Jako typický hybridní žánr Santini zmiňuje newsletter — elektronický zpravodaj obvykle rozesílaný prostřednictvím e-mailu. Text newsletteru často vykazuje prvky několika žánrů, jako je reportáž, reklama, rozhovor apod. Hybridní žánrová povaha je integrální součástí newsletteru. Jedná se totiž zpravidla o obchodní sdělení, které je psáno tak, aby působilo jako výčet aktualit, čímž je potlačena explicitnost persvaze.

Obě tyto (a mnohé další) studie poukazují na to, že obsah internetu netvoří pouze internetově specifické žánry, nýbrž i replikace žánrů z klasických médií, byť novému médiu různou měrou přizpůsobené.

Povahu internetového obsahu dobře shrnuje Eva Höflerová: „Prostředí elektronických textů sloučených v síti internetu je další komunikační dimenzí, která se chová stejně jako předchozí komunikační „epochy“ — zahrnuje všechny tvůrčí postupy a vyjadřovací prostředky epoch minulých — usiluje zobrazit svými (novými) způsoby účinnost neverbálních prostředků řeči, operativnost mluvené komunikace, uspořádanou lineárnost komunikace psané, rozsah a sílu působení epochy masové komunikace. To vše pomocí optických a zvukových informací.“ (Höflerová, 2013) Všechny tyto definiční

rysy vytvářejí unikátní a žánrově pestré médium, jehož diverzita je ve středu pozornosti této práce.

2 Jazyková variabilita

Jazyková variabilita je objektem zájmu české i světové lingvistiky a v různých obměnách jsou myšlenky o ní přítomny už od samotných počátků myšlení o jazyce. Fenomén jazykové variability je natolik komplexní, že pokládá základ některých samostatných lingvistických disciplín a pro jiné se stává bazálním konceptem.

Nejvágněji lze variabilitu[8] definovat jako *vlastnost lidského jazyka vyjádřit stejný obsah různými způsoby*. Konkrétněji se vyjadřuje např. Trask, který definuje variabilitu jako „existenci pozorovatelných rozdílů ve způsobech užití jazyka v rámci jazykové komunity“ (1999, překl. JH). Spojení pojmu variability s komunitou mluvčích vypovídá o zjevném spojení s oblastí variační sociolingvistiky, navíc spíše než variabilitu jako vlastnost jazyka popisuje existenci jazykových variet, tedy množin jazykových prvků s obdobnou sociální distribucí (Nekvapil, 2017b).

Přesněji se k variabilitě vyjadřuje např. Cvrček v definici v NESČ, který především akcentuje existenci variant bez ohledu na mluvčí: „[Variabilita je] ve statickém pohledu případ neexistence vzájemně jednoznačného zobrazení (tj. 1 : 1) mezi prvky množiny forem a prvky množiny významů/funkcí, tedy existence více forem pro jeden význam, n. více významů pro jednu formu.“ (2017)

Cvrčkovu formalizující definici je možné uplatnit univerzálně — už samotnou jazykovou diverzitu lze nazírat jako variabilitu pojatou na nejvyšším stupni obecnosti. Stejný obsah může být vyjádřen z formálního hlediska zcela odlišně. Nejen rozdíly lexikálními a gramatickými — známé jsou mezijazykové rozdíly v pragmatice, rozličná funkcionalita suprasegmentálních jevů, gestiky, mimiky apod. Striktně vzato, nemusí to být dokonce ani mluvený či psaný

[8] Pojem *variace* preferovaný některými lingvisty je považován za synonymní.

jazyk, co je použito pro vyjádření obsahů odpovídajících těm, které bývají vyjadřovány přirozeným jazykem — srov. diverzitu sémiotických modů, teorii multimodality apod., kde se zkoumá např. význam obrazových vjemů, barev, tvarů, prostorových objektů atd. (např. Kress, 2010, van Leeuwen, 2014)

Postupnou konkretizací se dostáváme od mezijazykové variability k vnitrojazykové. Klíčovou osobností v jejím výzkumu je bezesporu William Labov.

2.1 William Labov a variační sociolingvistika

Do popředí odborného zájmu se jazyková variabilita dostává v 60. letech 20. století díky pracím Williama Labova. Nový přístup k variabilitě prezentovaný Labovem bývá nezřídka označován za revoluci v lingvistice (např. Trask, 1999), a to právě díky tomu, že zdůrazňuje variabilitu jako nedílnou součást jazyka, jež sama zasluhuje odbornou pozornost jakožto středobod lingvistického výzkumu.

Sociolingvistika v Labovově pojetí (označovaná také variační sociolingvistika, v anglofonním odborném diskurzu také jako *labovian sociolinguistics* — *„labovovská sociolingvistika“*) je založena na zkoumání korelačních vztahů mezi jazykovými proměnnými a proměnnými určenými sociálními aspekty. Východiskem pro výzkum tohoto typu je předpoklad, že mluvčí (příp. komunity mluvčích) užívají různé prostředky pro vyjádření týchž významů, což je podmíněno sociálními aspekty, tedy konfigurací sociálně podmíněných proměnných. Jazykové proměnné bývají často na fonologické a morfologické rovině (Nekvapil, 2017a) (např. hláskové alternace, proteze, diftongizace), sociálně podmíněné proměnné pak především reflektují sociální stratifikaci, tj. přináležitost mluvčích do sociálních vrstev, dále tradičně věk, pohlaví a etnicitu.

2.2 M. A. K. Halliday

Klíčovou osobností ve výzkumu variability je vedle Williama Labova i britský lingvista Michael Alexander Kirkwood Halliday. Ten dělí jazykovou variabilitu na dva od sebe oddělené druhy: sociální a funkční (1989, s. 44).

Prvním z druhů je variabilita sociální. Tu předurčují podobné faktory, které nacházíme v labovovském paradigmatu. Halliday tento druh variability spojuje s pojmem *dialekt* — ten vysvětluje jako varietu, kterou mluvčí používá, protože náleží (původně i nepůvodně) do určité regionální oblasti, sociální třídy, kasty, generace, věkové skupiny, příslušnosti k pohlaví nebo jiných podstatných skupin v rámci komunity. Na konstituci dialektu se nepodílejí vždy všechny uvedené faktory najednou a stejnou měrou, ale jakákoli jejich kombinace může být relevantní. Dialekty nejsou ostře ohraničené, existují kontinuální přechody od jednoho dialektu k druhému (tamtéž). Mluvčí se s dialekty identifikují a rozpoznávají dialekty cizí. V praxi je běžný jev nazvaný *dialect-switching*, kdy si mluvčí úspěšně osvojují více než jeden dialekt. Buď jsou schopni užívat oba dialekty simultánně, nebo přecházejí od jednoho k druhému (Halliday, 1978, s. 34).

Přechod mezi dialekty není podmíněn náležitostí mluvčího do skupin definovaných výše zmiňovanými sociálními aspekty, nýbrž je motivován situačním kontextem (tamtéž). Situační kontext spadá do druhého typu variability, jenž Halliday vyděluje, a to variability funkční.

Funkční variabilita vychází ze skutečnosti, že se mluvčí během svých životů ocitají ve velkém množství komunikačních situací, které se od sebe v různých rysech značně liší. Z této heterogennosti pramení odlišnosti v promluvách, jež mluvčí realizují pomocí vědomého a cíleného výběru jazykových prostředků adekvátních daným komunikačním situacím. Například mluvčí, pro jehož dialekt jsou běžné diftongizované koncovky ve tvrdém skloňování adjektiv mužského rodu, bude používat tuto variantu v komunikačních situacích neformálního charakteru. Pro institucionální komunikaci např. v rámci zaměstnání se ale bude snažit vybírat koncovky spisovné. K výběru variantního prostředku, tj. výběru hodnoty jazykové proměnné, dochází na

základě několika faktorů. Ty Halliday označuje jako *field*, *mode* a *tenor*.[9] Jazykovou variaci podmíněnou konfigurací těchto tří faktorů pak nazývá registr (angl. *register*; analogicky k dialektu, který je podmíněn konfigurací sociálních proměnných).

Pojem *field* souhrnně označuje „událost, ve které text figuruje spolu s účelnou aktivitou mluvčího nebo pisatele“ (Halliday, 1994, s. 22).

Mode určuje funkci textu v rámci události definované jako *field*. *Mode* zahrnuje druh kanálu, kterým se informace přenáší (psaný, nebo mluvený), rysy jako připravenost nebo nepřipravenost promluvy a — což je podstatné — také žánr komunikátu, jako je narativní text, didaktický, persvazivní apod. (tamtéž).

Tenor pak referuje k druhu interakce mezi mluvčími a vztahy mezi nimi, jak z krátkodobého, tak z dlouhodobého hlediska (tamtéž).

2.3 Pojem registr

2.3.1 Registr u M. A. K. Hallidaye

Pojem registr, který ve větší míře zavedl M. A. K. Halliday v pracích, na něž navazuje mnoho vědců a vědkyň řadících se do směru systémové funkční gramatiky, není všemi používán stejně, byť stále zůstává pojmem ležícím v ohnisku popisu funkčně podmíněné variability. Mrázková (2014) mluví o dvojím chápání registru: „na jedné straně existuje statická představa registru jako soupisu specifických, zpravidla lexikálních prostředků (termínů či slangismů), na druhé straně je registr chápán jako repertoár jazykových prostředků ze všech rovin jazyka, včetně prozodie či parajazykové komunikace, a také aktivně, jako způsob využití těchto prostředků v komunikační situaci.“ (Mrázková, 2014, s. 103)

Pro koncept analýzy registru je charakteristickou vlastností jeho univerzálnost při uplatnění na jazykovém materiálu. Možnosti nastavení

[9] Tyto pojmy jsou jen obtížně přeložitelné, např. Kamila Mrázková (2014) překládá *field* jako pole, *mode* jako způsob/modus promluvy a *tenor* nechává nepřeložený, resp. přejímá jej v původní podobě.

„rozlišení“ jsou v tomto případě takřka arbitrární. Je možné mluvit o registru mluveného jazyka obecně, stejně jako je možné jej libovolně specifikovat např. na registr připravených proslovů a dále např. na registr připravených svatebních proslovů apod. Podstatné je, že pro registr neexistuje žádná správná míra specifičnosti (Biber–Conrad, 2009, s. 32). I tak lze ale registry rozdělit na obecné a specializované (tamtéž). Spíše než o diskrétní kategorie jde o dva póly škály, na které lze registry pomyslně umístit. Nejblíže pólu obecných registrů by byl např. uvedený registr mluveného jazyka. Ten je vymezen pouhým jedním rysem diskurzního modu (*mode*), a to vymezením druhu kanálu, jímž je předávána informace. Naopak na druhém pólu, pólu specializovaných registrů, by mohl být zmiňovaný registr připravených svatebních projevů, jenž zahrnuje kromě rysů druhu kanálu a dalších modálních charakteristik také podstatné aspekty pole (*field*), které je v tomto případě velmi specifické a řídí se konkrétními kulturními vzorci (tento registr by bylo možné dále rozdělit na subregistry podle kulturně-geografického hlediska), a tenoru, zde zastupujícího vztahy mezi účastníky komunikační situace, které jsou při svatebních proslovech přirozeně velmi podstatné.

Neohraničenost specifikace přináší značnou volnost, která dělá z analýzy registru hallidayovské tradice mocný nástroj pro popis funkční variability. Ze stejného důvodu je však registr často podrobován kritice. Např. David Crystal uvádí: „Kritika nekonzistence může být nejlépe ilustrována využíváním termínu *registr*. [...] Tento termín byl využíván pro označení jazykových variet až nekritickým způsobem, jako by bylo možné jej vhodně využít pro označení situačně distinktivních úseků jazyka jakéhokoliv druhu.“ (Crystal, 1973, s. 61, překl. JH)

Tato zastřešující funkce registru podle Crystala skrývá rozdíly mezi registry definovanými na různé míře obecnosti, což je podle něj „nekonzistentní, nerealistické a zmatečné“ (tamtéž, překl. JH) a trivializuje to potenciálně užitečný koncept.

2.3.2 Registr a jeho ekvivalent v českém prostředí

V českém jazykovém prostředí (na rozdíl od slovenského, srov. Mrázková, 2004, s. 102) se pojem registr používá jen zřídka (najdeme jej např. v pracích Světly Čmejrkové). Neznamená to však, že by v české lingvistice nebylo místo pro výzkum funkční variability. V českém prostředí je registrová analýza, resp. její ekvivalenty doménou funkční stylistiky, jejíž kořeny sahají až k základům české stylistiky per se.

Funkční stylistika se rodí na počátku 20. století v prostředí Pražské školy. Klíčovou osobností, jež stála u jejího zrodu byl Bohuslav Havránek. Ten jako první vymezil tzv. funkční jazyky, z nichž později vzchází teorie funkčních stylů. Funkční jazyky jsou funkčně podmíněné kategorie promluv v různých komunikačních situacích, které jsou spojené se specifickým výběrem jazykových prostředků. Tento koncept si neklade za cíl mít schopnost popsat veškeré potenciální promluvy mluvčích ve všech potenciálních komunikačních situacích, nýbrž zejména vystihnout a kategorizovat nejdůležitější funkce jazyka a popsat jejich rysy. Jednotlivé funkční jazyky vnímá jako rovnocenné, promluvy hodnotí na základě adekvátnosti: „Hodnotiti lze jazykový projev jedině podle jeho adekvátnosti k účelu, podle toho, zda vhodně plní daný úkol. Nelze hodnotit izolovaná slova, oddělená od jejich funkčního využití.“ (Havránek, 1932, s. 63; cit. z Schneiderová, 2013, s. 160)

K Havránkovým funkčním jazykům (hovorovému, pracovnímu, vědeckému a básnickému) přispívá Jan Mukařovský svým rozpracováním estetické funkce, čímž ve výsledku vzniká základ teorie funkčních stylů — konceptu českého funkčněstylistického popisu, který se ve více či méně nezměněné podobě užívá dodnes. Funkční styly jsou stále diskrétními kategoriemi a komunikáty jsou do nich zařazovány na základě převládající tendence k náležitosti k jednomu z nich.

Funkční styly jsou klíčovým konceptem k určování funkčních charakteristik textů, s hallidayovským pojmem registr se ale překrývají jen částečně. Jejich definiční schopnosti spadají do diskurzního módu, tak jak jej vyčleňuje Halliday. K zohlednění všech slohotvorných činitelů podílejících se na hallidayovském registru musíme zvážit i další faktory, které v české stylistice tvoří spolu s funkčními styly širší definiční rámec, jenž zahrnuje i ostatní

aspekty módu, ale také pole i tenor. Tím je v české stylistice koncept objektivního stylu, od jehož výzkumu je odvozen i název podoboru stylistiky: objektivní stylistika (Jelínek–Krčmová, 2017).

Objektivní styl je souhrnem objektivních stylotvorných faktorů (také zvaných slohotvorní činitelé). Ty stojí mimo komunikační subjekt, na rozdíl od jejich protějšku, subjektivních slohových faktorů, které jsou spjaty s komunikačním subjektem, „s jeho svérázností a individualitou" (Čechová et al., 1997, s. 50).

Subjektivní stylotvorné faktory zhruba odpovídají sociálně podmíněným faktorům definujících dialekty v Hallidayově pojetí. Pojetí subjektivních stylotvorných faktorů se u českých autorů nepřekrývá, konsenzuální výčet uvádí např. Smejkalová (2013). Ta mezi ně řadí pohlaví komunikačního subjektu, jeho věk, vzdělání, psychické založení produktora obecně, jeho aktuální fyzický a psychický stav, osobní sklony a zájmy, místo, kde žije nebo odkud pochází, jeho profesi a konečně příslušnost k určité ideologii, která může částečně spadat i do faktorů objektivních, a to v případě nevědomého uplatňování ideologických postojů mluvčím (Smejkalová, 2013, s. 153).

Stejně jako u subjektivních stylotvorných faktorů, není jednotné ani pojetí objektivních stylotvorných faktorů (o nejednotnosti viz např. Chromý, 2012; Prošek, 2013). Eva Minářová (1997) vyděluje tyto objektivní stylotvorné faktory:

a) základní funkce komunikátu, cíl komunikace a intence tvůrce (záměr)
b) ráz komunikátu
c) situace a prostředí komunikace
d) charakter adresáta
e) užitá forma komunikátu
f) míra připravenosti komunikace
g) užitý kód jazykové komunikace
h) téma komunikátu

(Minářová, 1997, s. 51)

Základní funkce komunikátu odpovídá funkčnímu stylu, jak byl popsán výše. Je nejzávažnějším objektivním stylotvorným faktorem, protože zásadně ovlivňuje jazykové prostředky v komunikátu, který vznikl za účelem přenosu jakéhokoli sdělení. Kromě základní funkce, již každý takovýto komunikát nese, jej mohou doprovázet další, a to „funkce odborně sdělná a vzdělávací, funkce direktivní (řídící), popř. operativní (správní), [...] funkce uvědomovací a získávací, funkce persvazivní (přesvědčovací), funkce esteticky sdělná a dílčí specifické funkce jiné“ (tamtéž).

Rázem komunikátu se rozumí míra jeho oficiálnosti. Reflektuje objektivní okolnosti komunikace, které se projevují výběrem jazykových prostředků na základě vztahu komunikátu s účastníky komunikace. Rázem se liší komunikace soukromá od veřejné (tamtéž). Jako výrazný příklad Minářová uvádí slavnostní projevy řečníků, které jsou cíleně stylizovány používáním citací, metafor, poetismů apod. (tamtéž, s. 53–54).

Dalším z objektivních stylotvorných faktorů je komunikační situace a prostředí, v němž se komunikace odehrává. Na prostředí komunikace má vliv velké množství dílčích faktorů. Podstatné je, nakolik je pro mluvčí prostředí známé, je-li veřejné nebo soukromé, případně nezasahují-li do komunikace nějaké vnější vlivy, jako je osvětlení, teplota, hluk apod. Důležitá je také vzdálenost mluvčího od adresáta nebo adresátů, případně obecně rozsah prostoru, v němž je komunikace uskutečňována (tamtéž, s. 54).

Předpokládaný adresát a jeho charakter se taktéž podílí výraznou měrou na výsledné podobě komunikátu. Neznámý adresát může nepříznivě ovlivnit duševní rozpoložení mluvčího, naopak pokud je adresát mluvčímu dobře znám, může se mluvčí přizpůsobovat výběrem jazykových prostředků, což dává lepší předpoklady pro úspěšnou komunikaci — srov. také teorii komunikační akomodace (např. Giles–Taylor, 1973). Mluvčí bere v potaz osobní a sociální charakteristiky známého adresáta. Např. mluví-li dospělý člověk na malé dítě, používá zjednodušenou varietu jazyka, jíž se snaží zjednodušit dítěti porozumění. Tato varieta bývá označována jako *baby talk* (např. Ferguson, 1977). K obdobnému zjednodušení dochází i v mluvě se staršími lidmi, zjednodušená varieta bývá nazývána *elderspeak* (např. Kemper et al., 1998). Charakter adresáta bývá zařazován jak k objektivním, tak

k subjektivním stylotvorným faktorům. Pro zařazení do první skupiny mluví to, že adresát není součástí mluvčího, na druhou stranu činí mluvčí rozhodnutí na základě vlastního vztahu k adresátovi. Minářová k tomuto uvádí: „Nejde o spornou problematiku, ale pouze o dvojí pohled na ni. Na jedné straně máme na mysli objektivně existující čtenáře nebo posluchače, k nimž jazyková komunikace směřuje, na druhé straně jde o komunikantovy subjektivní schopnosti vnímat nebo nevnímat adresáta a dovednosti usměrňovat a modifikovat komunikaci právě s ohledem na něho." (Minářová, 1997, s. 55)

Dalším z faktorů je užitá forma komunikátu. Ta postihuje důležitou kategorizaci komunikátů na psané a mluvené. Toto dichotomické dělení lze uplatnit takřka na všechny komunikáty, byť lze hovořit o komunikátech, které jsou z hlediska užité formy hybridní, jako jsou např. internetové komunikáty probíhající v reálném čase (tzv. *instant messaging*), které nesou definiční rysy psaného i mluveného jazyka. Mluvené komunikáty jsou většinou nepřipravené, výsadní je pro ně přítomnost suprasegmentálních jevů a neverbálních komunikačních prostředků, jako je gestika, mimika, kinesika, v případě komunikace tváří v tvář také proxemika, haptika a další. Psané komunikáty tyto prostředky často nahrazují graficky pomocí interpunkčních znamének, v případě internetových komunikátů také emotikony, emoji apod.

Jak již bylo naznačeno, s formou komunikátu úzce souvisí míra jeho připravenosti mluvčím. Podobně jako u vlivu adresáta, i zde může tento stylotvorný faktor hraničit se subjektivními stylotvornými faktory: „Na výsledné podobě stylu komunikátu se vždy míra připravenosti bezprostředně podílí. Určitou roli má rovněž i zkušenostní komplex tvůrce textu a jeho schopnost kulturního komunikování, i když jde o nepřipravenou výměnu myšlenek, o komunikaci bezprostřední a spontánní. V tomto případě už uvažujeme o zásahu činitele subjektivního." (Minářová, 1997, s. 58)

Užitým kódem jazykové komunikace je míněn výběr jazykových prostředků z určitého útvaru národního jazyka. Vědomý výběr prostředků mluvčím je však možné považovat spíše za subjektivní stylotvorný faktor. Objektivnost tohoto faktoru spočívá ve skutečnosti, že mluvčí vybírá prostředky útvaru nevědomky na základě okolností komunikace, jako je funkce, prostředí nebo účastníci (tamtéž).

Poslední faktor, který Minářová vyděluje, je téma komunikátu. Tento faktor je problematický a není vyčleňován všemi lingvisty a lingvistkami, zejména kvůli tomu, že téma komunikátu lze pojmout různými styly (Prošek, 2013, s. 149). Objektivnost tématu jako stylotvorného faktoru je však obhajitelná skrze jisté konvence, které vyžadují tematicky zaměřené komunikáty. Je-li tématem textu oblast spojená s vědeckým výzkumem, např. objektivní stylotvorné faktory, bude pravděpodobně komunikát, jenž o nich pojednává, vykazovat charakteristiky vědeckého stylu. Jedná se však jen o tendenci — žádným způsobem není omezena možnost komunikovat o stejném tématu v neformální situaci, a to za použití jazykových prostředků typických pro nespisovné útvary češtiny.

Schizoidní povaha některých objektivních stylotvorných faktorů je znamením toho, že není žádoucí operovat se stylotvornými faktory výhradně jako s diskrétními kategoriemi. Minářová uvádí, že „nelze jevy ve stylistice chápat izolovaně, ale v celistvosti“, např. že při výzkumu vlivu tématu na výsledný styl by měly být brány v potaz i ostatní stylotvorné faktory a jejich podíl na konstituci komunikátu jako celku.

Není pochyb, že je teorie objektivních stylotvorných faktorů vycházející z tradice Pražské školy nástrojem pro popis velmi podobné oblasti, kterou pokrývá koncept registru u Hallidaye a dalších. Mrázková k tomuto srovnání uvádí: „Dimenze kontextu a jejich další členění, které Halliday vymezuje, postihují zhruba totéž, co objektivní stylotvorné faktory.“ (Mrázková, 2014, s. 105) Největší podobnost a z ní pramenící možnost propojení těchto dvou konceptů spatřuje Mrázková v koncentraci na funkci. Ta se v obou pojetích do jisté míry liší. Podle Hallidaye jsou funkční všechny aspekty komunikace, které se podílejí na výsledku výběru jazykových prostředků mluvčím, v tradici Pražské školy je „pojem funkce vyhrazen pro obecné funkce jazyka či promluvy“ (tamtéž, s. 106). Rozdíl je také v pojetí literárních textů. V anglofonní lingvistické tradici je zvykem abstrahovat literární texty s převažující estetickou funkcí a analyzovat je právě skrze ni. Zatímco koncepty Pražské školy jsou zamýšleny jako univerzálně aplikovatelné nástroje pro popis jakýchkoli textů, v anglofonní tradici má hlavní vliv na

variaci textů stylizace, která reflektuje estetické preference jejich autorů a autorek.

2.4 Distinkce registr-styl-žánr

Ve studiích týkajících se analýzy registrové variace je registr často dáván do vztahu se dvěma jinými, příbuznými termíny, kterými jsou *žánr* a *styl*. Jejich přesné vymezení je velmi problematické, protože je rozrůzněné napříč přístupy mnohdy i jednotlivých lingvistů a lingvistek.

Jedním z největších rozdílů mezi anglofonním a českým diskurzem[10] je již zmiňované pojetí stylu. V anglofonní tradici je hledisko analýzy stylu uplatnitelné výhradně na beletristické texty, jako deskriptivní nástroj tedy není univerzální. Vyčleňuje se tak mimo registr a žánr jakožto specifický nástroj pro popis variability, která je podmíněná zejména estetickou preferencí — hodnoty jazykových proměnných jsou vybírány na základě jejich estetických hodnot. V české stylistice má styl významnější postavení, jelikož reprezentuje záměrný výběr a organizaci výrazových prostředků, které se obecně uplatňují při vzniku textu (Křístek, 2017). Styl je tak stěžejním konceptem celé české stylistiky, kdežto v anglofonní lingvistice je pouze jejím dílčím nástrojem, který se vyznačuje blízkostí k literární vědě, tedy částečnou interdisciplinaritou — styl je definován až v rámci již definovaného registru nebo žánru (Biber, 2009, s. 18).

Prvky, skrze něž je konstruován žánr textu, jsou přítomny v celém textu na jeho specifických místech, a pro jejich analýzu tedy nestačí analyzovat pouze úryvek. Tyto prvky „jsou v souladu s kulturně podmíněným způsobem konstrukce textu náležícího k dané varietě“ (tamtéž, s. 16, překl. JH). Jedná se tedy o jisté signalizační prvky, markery příslušnosti textů do určitých varietních kategorií. Biber uvádí, že se v textu obvykle vyskytují jen jednou, ale sehrávají velmi podstatnou roli v konstituci variety. Např. žánr pracovního dopisu je determinován formálními náležitostmi, jako je oslovení (*Vážená paní*

[10] Východiskem této práce je na jedné straně paradigma *corpus-based* přístupu spojeného s výzkumy Douglase Bibera, na druhé straně tradice pražské funkční stylistiky. Proto jsou srovnávány především tato dvě hlediska.

Holanová, příp. titul adresáta či adresátky), následované textem dopisu a formálním zakončením (*S pozdravem* apod.) (tamtéž, s. 17). Identifikace žánru na základě těchto signalizačních prvků vyžaduje čtení celého textu. Podobně např. žánr pohádky determinují konvencionalizované obraty jako *byl/a jednou jeden/jedna*, *bylo nebylo* jako prototypické úvodní fráze nebo *žili spolu šťastně až do smrti*; *pokud nezemřeli, tak tam žijí dodnes* apod. jako zakončení. Žánrové markery se netýkají jen konkrétních konvencionalizovaných jazykových prostředků, ale také specifických aspektů horizontálního nebo vertikálního členění textu. V českém prostředí je žánr ztotožnitelný se slohovým útvarem (Cvrček, 2018, s. 293). Osvojení náležitostí slohových útvarů je předmětem školní výuky jakožto dílčí schopnost výstavby textu. Vyučovány jsou zejména administrativní útvary, jako je žádost, životopis, oznámení nebo inzerát (Krčmová, 2017). Žánrové charakteristiky textu jsou považovány za charakteristiky vnětextové (Cvrček, 2018, s. 293). Díky nim lze text kategorizovat z vnějšího hlediska — specifické vnější charakteristiky dávají dostatečnou motivaci k zařazení komunikátu do některého z žánrů. Např. internetový text lze zařadit k žánru příspěvku diskusního fóra na základě vnějších charakteristik, s kterými je spjat, jako je datum a čas odeslání příspěvku, uživatelské jméno autora či autorky apod. Samozřejmostí je při kategorizaci i vliv vnitrotextových charakteristik, které nelze od těch vnětextových spolehlivě oddělit (tamtéž, s. 294).

2.5 Intratextuální přístup k registrové analýze

Moderní přístup k popisu registru reprezentuje osobnost amerického lingvisty Douglase Bibera. Jeho výzkumy vycházejí z korpusového přístupu k jazykovým datům a jejich kvantitativního zpracování.

Biberovo pojetí registru navazuje na hallidayovskou tradici. Registr definuje jako varietu, která je spojená s určitým užitím jazyka (Biber, 2009, s. 6), tj. s jeho funkcí. Klíčové pro Biberovu definici je rozlišení tří složek popisu registru: „situační kontext, jazykové rysy a funkční vztahy mezi prvními dvěma složkami.“ (tamtéž, překl. JH) Situační kontext pak Biber považuje za základní složku registru, která je nadřazená složce jazykových rysů — funkční

východiska produkce komunikátu nejsou odvoditelná, kdežto jazykové rysy vyprodukovaných komunikátů jsou odvoditelné právě ze situačního kontextu a jeho dílčích charakteristik (tamtéž). Ve výzkumech analyzujících registr je důležité věnovat se oběma těmto složkám a explicitně je popsat, jelikož funkční interpretace je založena právě na srovnání jazykových rysů s rysy situačními.

Biber upozorňuje na rozdíly mezi výzkumem registrové variability a výzkumem variability dialektické. Dialektologické výzkumy se zpravidla nesoustřeďují na významovou složku komunikace, nýbrž povětšinou na fonologické rysy, jako je redukce hlásky /r/ v anglických slovech jako *park* (tamtéž, s. 11) nebo ztráta jotace po labiálách na Litomyšlsku ve slovech *pet, pekná* (místo *pět, pěkná*) (Bachmannová, 2017), a na další lexikální a gramatické rozdíly, jako jsou dvojité negativy, synonymní výrazy pro nealkoholický nápoj *soda* a *pop* (Biber, 2009, s. 11) nebo slezský výraz *krepla* pro koblihu (ČJA2, 1997, s. 207). Tyto příklady variace signalizují příslušnost mluvčích ke geografické oblasti, sociální vrstvě nebo jinak specifikují skupinu ze sociálního hlediska. Mezi jednotlivými variantami není výrazný sémantický rozdíl, variace tohoto typu není primárně variací funkce. Z toho vyplývá, že registrová variabilita stojí v opozici vůči variabilitě dialektické (Biber, 2009, s. 11). Rozdílné jsou i formální aspekty těchto dvou druhů variace:

> „Jazykové proměnné v dialektologických studiích jsou téměř vždy výběrem mezi dvěma jazykovými variantami. Skóre variability je proporce ukazující relativní preferenci jedné, nebo druhé varianty. Na druhou stranu, jazykové proměnné ve studiích zabývajících se registrem jsou mírou výskytu jazykového rysu, přičemž vyšší míra výskytu je interpretována jako odraz zvýšené potřeby funkcí, které jsou spojeny s daným rysem.“ (tamtéž, s. 11–12, překl. JH)

Důraz na jazykové rysy reflektující funkční variabilitu textů je těžištěm Biberových výzkumů registru. Kvantitativní analýzou jazykových rysů a propojením s vnětextovou charakteristikou komunikátů, v kterých jsou přítomny, vzniká nový model popisu, který je poměrně odlišný od tradičních modelů, a to svým kvantitativním, korpusově orientovaným základem.

2.6 Multidimenzionální analýza registrové variability

Multidimenzionální analýza (MDA) je corpus-based metoda, jejímž hlavním cílem je popsat variabilitu v korpusu textů, potažmo v celém jazyce, jejž korpus reprezentuje. Síla této metody spočívá ve zohlednění velkého počtu jazykových proměnných, tj. rysů, které mohou nabývat většího množství hodnot, a tím jsou považovány za nositele variability v textu. Z velkého množství hodnot je pak pomocí faktorové analýzy vymodelován prostor s významně sníženým počtem dimenzí (zpravidla do deseti). V tomto multidimenzionálním prostoru se nachází každý z analyzovaných textů. Dimenze jsou poté lingvisticky interpretovány a zpravidla označeny názvy, jež označují výrazné hodnoty skalárních opozic. Na základě poloh textů v rámci jednotlivých dimenzí je následně možné texty charakterizovat z hledisek, která jsou determinována pouze jejich vnitrotextovými aspekty, a nikoli prominentními žánrovými markery. MDA tak nabízí alternativní přístup stylistického výzkumu, který do značné míry odstraňuje možné negativní vlivy lingvistické introspekce.

Douglas Biber není prvním lingvistou, který se věnoval MDA vnitrotextových aspektů textu, sám uvádí tematicky relevantní výzkumy, které byly prováděny už od 60. let 20. století (Biber, 1988, s. 61). Přesto je možné jej považovat za průkopníka této metody, což dokazuje i skutečnost, že současné výzkumy užívající MDA vycházejí právě z jeho studií z 80. a 90. let.

Postup při výzkumu užívajícím MDA lze rozdělit do několika fází: Sestavení jazykového korpusu, výběr a operacionalizace jazykových rysů, faktorová analýza a konečná interpretace dimenzí variability.

2.6.1 Sestavení jazykového korpusu

Na počátku výzkumu je nutné vybrat či vytvořit korpus, na kterém bude analýza provedena. Nejprve je sebráno dostatečné množství jazykových dat. Tento krok se zdá celkem neproblematický, jistá úskalí však přináší aplikace MDA na specializované korpusy textů nebo na korpusy minoritních jazyků (viz např. Biber, 1995). Při sběru textů pro výchozí (nespecializované)

korpusy je kladen důraz nejen na dostatek dat, ale také co nejvyšší žánrovou diverzitu, skrze niž je dosaženo velké míry variability napříč texty.

Z textů jsou pak utvořeny vzorky o podobné délce,[11] jednak zkrácením textů v případech, kdy je jejich délka větší než stanovený rozsah, a jednak konkatenací několika textů v případech, kdy je jejich délka kratší. Nedostatečná délka textů není závadou, nýbrž přirozenou vlastností některých žánrů. Díky konkatenaci je tak možné dosáhnout ještě vyšší žánrové diverzity. Délka textového vzorku je poměrně netriviální problém, který může mít v případě nedůslednosti vliv na výsledky analýzy. Přílišná krátkost textových vzorků byla také předmětem kritiky metodologie MDA v jejích počátcích. Douglas Biber tomuto problému věnuje bližší pozornost ve studii *Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation* (1990), kde reaguje na kritiku svého dřívějšího výzkumu (Biber, 1988). V tom užil textové vzorky o velikosti 2000 a 5000 tokenů. Dostatečnost délek kontroluje rozdělením vzorků na dílčí vzorky o velikosti 1000 tokenů, analýzu provádí na nich a výsledky porovnává vzájemně mezi dílčími vzorky uvnitř vzorků původních. Výsledky poukazují už na spolehlivost vzorku o velikosti 1000 tokenů, což přirozeně implikuje i spolehlivost vzorků většího rozsahu (Biber, 1990, s. 261).

Dalším z problémů je otázka velikosti celého korpusu a množství vzorků v rámci vymezených vnětextově definovaných kategorií. Biber ve výzkumu z roku 1988 pracuje s korpusem o přibližné velikosti 1 milion tokenů, který sestává z 481 textů. Adekvátnost velikosti korpusu i vliv diverzity testuje podobným způsobem jako velikost textových vzorků — vytvořením menších částí a analýzou provedenou na nich. Z původního milionového korpusu bylo extrahováno sedm subkorpusů, z nichž šest reprezentovalo diverzitu původního korpusu. U sedmého subkorpusu byla míra variace cíleně omezena. Na každém ze subkorpusů byla následně provedena faktorová analýza a vymodelovány dimenze. Na základě podobností ve výsledcích Biber dokazuje, že velikost korpusu o 1 milionu tokenů je dostatečná, protože už subkorpusy velikostně odpovídající jedné čtvrtině původního korpusu

[11] Podobnou délkou je zde míněn alespoň řádově stejný rozsah textových vzorků.

vykazují podobné výsledky faktorové analýzy. Toto testování poukázalo i na skutečnost, že diverzita vstupních textů je velmi podstatná, jelikož subkorpus s omezenou diverzitou zdaleka neodpovídal výchozímu korpusu, byť byl rozsahem větší než zbylých šest subkorpusů.

Poslední výraznější problém vyvstává při samotné klasifikaci textů. Vzorky jsou totiž do korpusu přidávány na základě vnětextových charakteristik — zastupují tedy žánry. Otázkou je, zdali je vnětextové hledisko dostatečně spolehlivé pro reprezentaci vnitrotextové variability. Texty jsou zpravidla čerpány z prostředí, která jsou pro ně typická — novinový článek bude čerpán z novin, odborný článek z odborného časopisu apod. Nebylo však jisté, je-li sebráním dostatečného vzorku spadajícího do vnětextové kategorie zajištěno, že texty dostatečně reprezentují daný žánr. Kritika směřuje především k otázce, zda je koncept žánru dostatečně dobře definován (tamtéž, s. 261–262). O jeho nedostatečné definici má svědčit existence subžánrů, tedy vnitřní strukturace, která není při výběru textu zachycena. Biber podrobuje i tento problém empirickému testování, a to srovnáváním textů v rámci jednotlivých žánrů. Z výsledků je zjevné, že již vzorky o deseti textech stejného žánru jsou schopny žánr dostatečně reprezentovat v korpusu a zároveň je patrné, že je žánr směrodatným konceptem pro reprezentaci diverzity v korpusech určených pro MDA (tamtéž).

Nároky na velikost korpusu pro MDA jsou tedy relativně nízké. Priority při jeho tvorbě dobře vystihují např. Cvrček et al. (2018b): „Základním požadavkem na [...] složení [korpusu] není reprezentativnost ve smyslu vyvážení proporcí textů tak, aby odrážely poměry textů v jazykové realitě okolo nás, ale reprezentativnost ve smyslu maximální pestrosti, zohledňující co nejvíc žánrových odlišností.“ (Cvrček et al., 2018b, s. 295)

2.6.2 Výběr a operacionalizace jazykových rysů

Druhým krokem je výběr jazykových rysů, které jsou nositeli variability. Funkční variabilita se v jazyce dotýká veškerých jazykových rovin. Od fonologie přes morfologii, lexikologii, syntax až k obecnějším, nadvětným textovým charakteristikám. MDA zahrnuje všechny z nich, a to do co největších podrobností. „[...] každý rys, který je spojený s určitými komunikačními

funkcemi nebo je různou měrou užitý v různých textových varietách, je do výzkumu zařazen.“ (Biber, 1995, s. 93, překl JH.) Každému jazyku je přirozeně vlastní jiná sada jazykových rysů, což ztěžuje replikaci výzkumů na jiných jazycích a částečně omezuje možnosti mezijazykového srovnávání variability. Seznamy jazykových rysů obvykle vycházejí z dosavadních lingvistických poznatků o zkoumaném jazyce, z gramatik a dalších zdrojů. Potenciální komplikace mohou nastat při výzkumu ze strukturního hlediska málo popsaných jazyků. Biber, který MDA aplikoval na větší množství jazyků, shrnuje povahu jazykových rysů takto:

> „[...] existují jisté základní gramatické, diskurzní a komunikační funkce, které jsou reprezentovány v každém lingvistickém inventáři. Ty zahrnují rysy, které značí:
>
> strukturální zpracování a komplexitu;
> lexikální zpracování, komplexitu a specificitu;
> místní a časovou referenci;
> temporální organizaci diskurzu (čas a vid);
> referenční kohezi a explicitnost;
> evidenciálnost, afektuálnost a postojovost;
> hedging a emfáze;
> interaktivnost a zapojenost.“
>
> (Biber, 1995, s. 104, překl. JH)

Všechny vybrané rysy je následně nutné operacionalizovat, a tím umožnit jejich měření v korpusu. K tomu je nutná dostatečná příprava korpusu, tj. jeho automatická tokenizace, lemmatizace, morfologické značkování, případné syntaktické značkování apod. Ruční značkování by bylo příliš obtížné a časově náročné kvůli rozsahu korpusu a velkému množství sledovaných rysů. Pro každý rys tak musí být vytvořena jeho formální reprezentace, díky které je možné jej v textu identifikovat. Toho může být dosaženo vytvořením dotazu v dotazovacím jazyku, naprogramováním skriptu nebo programu, který může pracovat s vytvořeným seznamem apod. Součástí tohoto kroku by měla být i kontrola a revize operacionalizace (viz např. Cvrček et al., 2018b, s. 299). Ne

každý rys je možné spolehlivě pojmout v celé jeho šíři, někdy jej není možné operacionalizovat vůbec. V mnoha případech je nutné přistoupit k heuristickým řešením, která jazykový rys neoperacionalizují dokonale, avšak pouze do uspokojivé míry.

V závěru této fáze jsou naměřeny hodnoty všech jazykových rysů pro každý z textových vzorků v korpusu. Vzniká tak matice hodnot velkého rozsahu o počtu řádků odpovídajícím počtu textových vzorků a počtu sloupců odpovídajícím počtu sledovaných jazykových rysů.[12] Tabulky, resp. datasety tohoto typu se označují jako *multivariační*, *multifaktoriální* (*multifaktorové*) nebo *multidimenzionální*.[13]

2.6.3 Faktorová analýza

Multidimenzionální datasety obsahují příliš informací na to, aby z nich bylo možné vyčíst podstatné informace pouhým okem, jako je možné např. u malých datasetů zobrazených v kontingenčních tabulkách. Cílem zkoumání multidimenzionálních datasetů je zpravidla odhalení skrytých vnitřních vztahů v datech, k čemuž jsou užívány různé statistické metody založené na odhalování korelací a kookurencí, regresních modelech nebo clusteringu — např. analýza hlavních komponent (*Principal Component Analysis, PCA*) nebo různé druhy korespondenčních analýz. Konkrétní metody jsou voleny s ohledem na hypotézy a na to, zda dataset obsahuje nominální nebo ordinální (příp. jiný) druh proměnných, popřípadě jejich kombinace.

Pro výzkum registrové variability se osvědčila faktorová analýza,[14] původně užívaná zejména v psychologii, sociologii nebo ekonomii. Faktorová analýza je založena na skutečnosti, že korelační vztahy uvnitř multidimenzionálních datasetů jsou spojeny s jistou mírou redundance, díky níž je možné redukovat proměnné zodpovědné za variabilitu datasetu na obecnější ukazatele, tzv. faktory, jejichž počet je výrazně nižší než původní počet pozorovaných

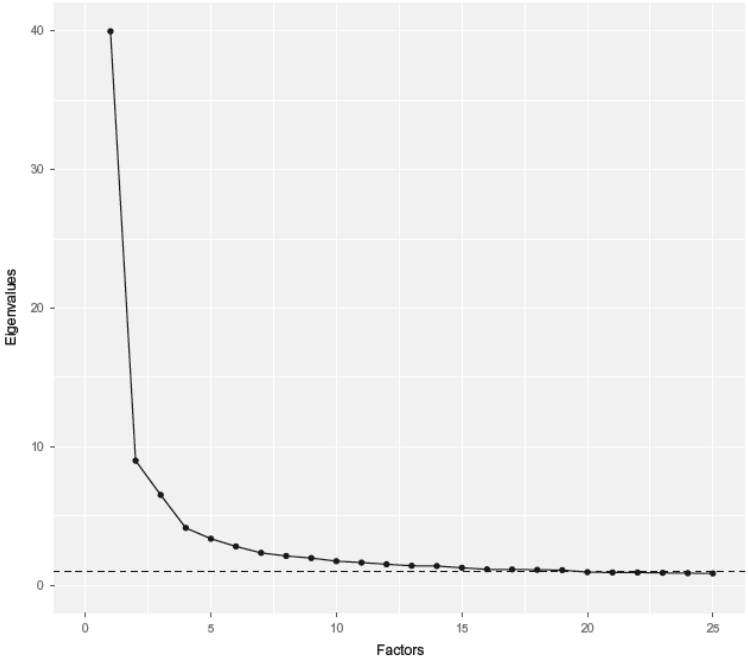
[12] Podle obecného rozvržení tradičně užívaného ve vědě, kde řádky zobrazují jednotlivá pozorování a sloupce reprezentují sledované proměnné.

[13] Užívání těchto termínů ve vědeckých pracích není rovnoměrné. Např. Guillaume Desagulier užívá adjektivum *multidimenzionální* pro označení datasetu a adjektivum *multifaktoriální (multifaktorový)* pro označení statistických metod aplikovaných na dataset (Desagulier, 2017, s. 240).

[14] Faktorovou analýzou je v této práci míněna výhradně explorační faktorová analýza.

proměnných. Např. jsou-li v matici přítomny podmnožiny s vysokou mírou korelace, je možné, že je za hodnoty jejich prvků zodpovědný jeden, latentní (tj. reálně nepozorovaný) faktor. Pokud lze v datasetu nalézt množství podobných korelujících podmnožin, pak je možné reprezentovat tyto podmnožiny malým množstvím faktorů (Rencher, 2002, s. 408). Tyto faktory jsou ve výzkumu jazykové variace interpretovány jako dimenze registrové variace.

Důležitým krokem při faktorové analýze je určení počtu výsledných faktorů. Toto rozhodnutí je netriviální a jeho pojetí se liší napříč výzkumy. Příliš malý počet faktorů není schopen pokrýt dostatečný objem variace. Snižováním počtu faktorů se tak snižuje koeficient determinace (R^2) reprezentující velikost části popsané variace všemi faktory v datasetu. Naopak příliš velký počet faktorů znesnadňuje jejich následnou interpretaci. S navyšováním faktorů nad ideální hranici se snižuje jejich přidaná informační hodnota, protože každý přidaný faktor přispívá k celkovému popisu variability méně než předchozí. Jedním z nejjednodušších způsobů je grafická reprezentace vztahu počtu faktorů a hodnoty tzv. vlastního čísla (angl. *eigenvalue*) korelační



Graf 1: Příklad sutinového grafu (Cvrček et al., 2018a, s. 13).

matice. Vlastní číslo je matematickou reprezentací rozptylu faktoru, což odpovídá míře variace datasetu, za kterou je konkrétní faktor zodpovědný.

Toto zobrazení se nazývá sutinový graf (z angl. *scree plot*) (např. Škaloudová, 2010) nebo loketní diagram (např. Harman, 2017)[15]. Na ose *x* tohoto grafu jsou vyznačeny jednotlivé faktory a na ose *y* jejich vlastní čísla. Křivka tohoto grafu je klesající a její sklon je nepřímo úměrný pořadí faktorů (viz příklad v grafu 1).

Výsledný počet faktorů užitý při analýze je určen na základě „charakteristického zlomu, který vytyčuje bod, od kterého přinášejí přidané faktory pro celkovou analýzu jen málo.“ (Biber, 1988, s. 82, překl. JH) Neexaktní metody výběru počtu faktorů, jako je tato, bývají často zdrojem kritiky faktorové analýzy.

Následujícím atributem, který se volí při provádění faktorové analýzy, je tzv. rotace faktorů. Jedná se o matematickou transformaci dat, jejímž cílem je učinit interpretaci dat co nejsnazší. Rotace zajišťuje, aby sledované proměnné měly vysokou faktorovou zátěž v co nejmenším počtu faktorů a aby byly hodnoty faktorové zátěže co nejkontrastnější (tj. s absolutní hodnotou blízkou 0 nebo 1)(Baayen, 2008, s. 127). Volba rotace je přizpůsobena povaze dat. Za předpokladu korelace mezi faktory se obvykle volí rotace *promax*. Pokud se korelace nepředpokládá, je možné zvolit rotaci *varimax* (tamtéž, s. 128).

Výstupem faktorové analýzy jsou hodnoty tzv. faktorové zátěže (angl. *factor loading*) pro každý ze sledovaných jazykových rysů. Tyto hodnoty určují, do jaké míry jsou jednotlivé faktory reprezentovány konkrétními jazykovými rysy. Faktorová zátěž může nabývat hodnoty od -1 do 1. Hodnoty od -0,3 do 0,3 zpravidla nejsou považovány za podstatné (bez ohledu na jejich statistickou signifikanci)(Biber, 1988, s. 85). Záporné hodnoty faktorové zátěže neznamenají nižší důležitost rysu — tu by bylo možné určit výpočtem absolutní hodnoty faktorové zátěže. Pozitivní faktorová zátěž jazykových rysů v rámci faktoru odráží skutečnost, že rysy kookurují v textech častěji než rysy s nulovou faktorovou zátěží. Naopak rysy spojené negativní faktorovou zátěží jsou v textech podreprezentovány, což je poznatek stejně podstatný jako

[15] Užívanější termín *sutinový graf* (podle internetového vyhledávače Google k 8. 7. 2019) je poněkud neprůhledný, původní *scree* totiž neodkazuje k sutinám jako materiálu, nýbrž k tzv. kamennému osypu, který vzniká na úpatí hor. Loketní diagram odkazuje k charakteristickému ohbí křivky. V textech je možné se setkat i s česky skloňovaným *scree plotem*.

pozitivní hodnota faktorové zátěže (tamtéž, s. 87–88). Pro každý faktor vznikají množiny sledovaných proměnných s prominentní faktorovou zátěží, které jsou ve výzkumu registrové variability interpretovány jako soubory jazykových rysů konstituující dimenze variability.

Každý textový vzorek získává hodnoty tzv. faktorového skóre, kterým je určováno umístění textu v rámci dimenzí. Tímto způsobem je možné každý z textových úryvků reprezentovat bodem v multidimenzionálním prostoru, jehož počet dimenzí odpovídá počtu faktorů.

2.6.4 Interpretace faktorů

Finální fází MDA je interpretace faktorů. Předpokládá se, že sdružení skupin jazykových rysů s výrazně vyšší či nižší faktorovou zátěží v rámci jednotlivých faktorů je funkčně podmíněné, tj. že skupiny rysů slouží společné funkci v textu (Biber, 1988, s. 91). Výsledné faktory je tak možné interpretovat jako dimenze, podél nichž texty variují. Jednotlivým dimenzím lze na základě obezřetné lingvistické interpretace přiřknout názvy, a tím stručně charakterizovat, čím se liší texty nacházející se v protilehlých oblastech této dimenze.[16] Tato fáze je náročná, protože je nutné faktory analyzovat z několika úhlů pohledu.

Primárním ukazatelem jsou zde právě extrémní absolutní hodnoty faktorové zátěže, jimž je přisuzována funkční platnost, a korelace mezi nimi. V interpretaci také výrazně pomáhá vnětextová klasifikace textových vzorků. Díky té je možné texty sdružovat do vnětextově definovaných kategorií a sledovat jejich distribuci v multidimenzionálním prostoru. Doplňkovou analýzou je pak zkoumání konkrétních textových vzorků, které dosahují z různých hledisek významných hodnot faktorové zátěže. Do interpretace vstupují i tzv. inertní jazykové rysy, tj. rysy, které v rámci faktorů nedosahují stanoveného prahu podstatnosti a jejich hodnoty se blíží nule. U některých faktorů může totiž být směrodatná i informace o tom, že se některý z rysů na konstituci faktoru vůbec nepodílí (Cvrček et al., 2018b, s. 301).

Z výsledků užívajících MDA za účelem zachycení registrové variability jazyků vyplývá, že pro každý jazyk jsou charakteristické jiné dimenze

[16] Každou z dimenzí je možné reprezentovat jednorozměrným skalárním grafem.

variability. Z mezijazykového srovnání Douglase Bibera vyplývá, že některé dimenze mají své odpovídající protějšky ve všech (zatím zkoumaných) jazycích, a tedy se jeví jako univerzální, některé jsou mezijazykově sdílené jen částečně a některé mohou být jazykově specifické (Biber, 1995, s. 237). Nejblíže univerzální platnosti jsou podle Bibera dvě dimenze, a to dimenze rozlišující psané a mluvené komunikáty (*oral* vs. *literate*) a dimenze, v níž se odráží míra narativnosti textu (*narrative* vs. *non-narrative*)(tamtéž). Vysokou míru jazykové specificity mohou vykazovat dimenze reflektující idiosynkratické rysy např. z oblasti pragmatiky, jako je honorifikace v korejštině (tamtéž).

2.6.5 Aplikace MDA na internetové texty

MDA je metodou, kterou lze uplatnit i pro popis různě vymezených oblastí jazyka, není tedy pouze nástrojem pro popis registrové variability jazyka v celé jeho šíři. Příkladem jsou výzkumy variability akademického jazyka, jako je výzkum Douglase Bibera (2006) nebo analýza Susan Conrad, jejíž studie se zabývá akademickými texty z oblasti biologie (1996).

Registrové variabilitě elektronické komunikace a internetového jazyka byla v anglofonním prostředí věnována relativně velká pozornost. Za jednu z prvních prací na toto téma lze jistě považovat studii Mileny Colot a Nancy Belmore *Electronic Language: A New Variety of English* (1996),[17] která aplikuje Biberovu MDA na korpus ELC (*Electronic Language Corpus*) a snaží se identifikovat situační rysy, které se významně podílejí na konstituci textů v ELC. Výzkum konkrétních internetově specifických žánrů pak reprezentuje např. MDA internetových blogů provedená Jackem Grievem et al. (2011). Výsledkem tohoto výzkumu bylo popsání čtyř dimenzí variace (informační vs. osobní zaměření textu, menší vs. větší zaměření textu na adresáta, míra tematické variace a dimenze týkající se narativního stylu). Z těchto výsledků byla pomocí clusterové analýzy vytvořena kategorizace blogů založená na lingvistických datech.

[17] Poprvé byla tato studie publikována už v roce 1993.

Pravděpodobně nejdůležitějším a nejobsáhlejším příspěvkem k výzkumu registrové variability internetových textů je studie *Register Variation on Searchable Web – A Multi-Dimensional Analysis* (2016), kterou publikovali Douglas Biber a Jesse Egbert. Jejím cílem je popsat pomocí MDA celý prostor prohledávatelného, „povrchového" webu, tj. internetových stránek, které jsou přístupné běžným uživatelům, aniž by museli zadávat přihlašovací údaje, případně za přístup platit apod. Do korpusu se tak nedostaly příspěvky na sociálních sítích a další soukromé dokumenty. Stejně tak v korpusu přirozeně chybí obsah tzv. deep webu, tj. stránek, které není možné vyhledat běžnými internetovými vyhledávači. (O možnostech získávání textů z internetu viz Ide et al., 2002 nebo starší studii Dewe et al., 1998) Data byla sbírána způsobem blízkým náhodnému výběru: obsah korpusu byl vytvořen excerpcí webových stránek, které byly vyhledány pomocí seznamu nejfrekventovanějších trigramů v korpusu současné americké angličtiny (COCA). Každý trigram byl automaticky zadán do vyhledávače, načež bylo staženo 800–1 000 odkazů na webové stránky, které byly následně staženy do počítače. Výsledný korpus obsahuje 43 685 textových dokumentů o přibližné celkové velikosti 52 665 000 tokenů.

Těžištěm tohoto výzkumu je anotace textových vzorků koncovými uživateli. Ti byli najati prostřednictvím crowdsourcingové sítě *Mechanical Turk*. Pracovníci anotovali texty na dvou úrovních obecnosti. Vniklo tak 8 obecných kategorií a 27 specifických registrových kategorií.

Obecné kategorie registru zahrnují narativní texty, informačně-popisné texty, názorové texty, interaktivní diskuse, návody / instruktážní texty, informačně-persvazivní texty, básně / písňové texty a přepisy mluvených textů. Dále Biber a Egbert vymezují hybridní registry, které vycházejí z neshody mezi anotátory. Každý z textů byl hodnocen čtyřmi anotátory, existuje tedy pět typů rozložení hodnocení: absolutní anotátorská shoda, kdy se všichni čtyři anotátoři shodnou na kategorii textu. Druhým rozdělením je nedokonalá shoda, při které tři anotátoři zařadí text do stejné kategorie, a jeden se odlišuje. V tomto případě je lišící se anotace zanedbána a text zařazen do kategorie, na níž se shodla většina anotátorů. Z třetího a čtvrtého rozdělení vznikají hybridní registry. Dvoučlenný hybridní registr vzniká

v situaci, kdy se dva anotátoři shodují na jedné kategorii a dva na jiné, trojčlenný hybridní registr v situaci, kdy se dva anotátoři shodují na jedné kategorii a zbylí dva volí dvě odlišné kategorie. V přibližně pěti procentech případů docházelo k absolutní neshodě, tj. případu, kdy každý z anotátorů volil jinou kategorii.

Kromě osmi obecných registrů byly texty zařazeny i do subregistrů, které nabízejí jemnější rozdělení a povahou zhruba odpovídají webovým žánrům. Jejich výčet a přiřazení k obecným registrům uvádí tabulka 1.

Registr	**Subregistr**
Narativní texty	Zpravodajský článek/blog
	Sportovní reportáž
	Osobní blog
	Historický článek
	Umělecká próza
	Cestovatelský blog
Informačně-deskriptivní texty	Popis předmětu
	Encyklopedický článek
	Výzkumný článek
	Popis osoby
	Informační blog
	Nejčastěji kladené dotazy (FAQs)
Názorové texty	Recenze
	Názorový blog
	Náboženský blog / kázání
	Poradenský text
Interaktivní diskuse	Diskusní fórum
	Fórum typu *question–answer*
Návody / instruktážní texty	Návod typu *jak na to*, instruktáž
	Recept
Informačně-persvazivní texty	Popis s účelem prodeje
	Editorial

Básně / písňové texty	Básně
	Písňové texty
Přepisy mluvených textů	Interview
	Formální projev
	Přepis mluveného textu z televize

Tabulka 1: Souhrn subregistrů (Biber–Egbert, 2016, s. 105–106, překl. JH).

U textů byly naměřeny hodnoty 57 jazykových rysů (na počátku autoři pracovali s více než 150 rysy, seznam byl ale výrazně redukován), a následně byla provedena faktorová analýza. Z ní vzešlo 9 následujících dimenzí variability[18]:

(1) Mluvenost/zapojenost vs. psanost/informativnost

Mezi rysy s pozitivní faktorovu zátěží zde patří nedokonavý vid, přítomné a budoucí slovesné časy, zájmena pro první a druhou osobu a různé markery postojovosti, negativní zátěž vykazují určité členy, předložkové vazby, pasivum nebo vztažné věty. Pozitivní faktorové skóre získávaly dle očekávání registry básní / písňových textů, přepisů mluvených textů nebo interaktivní diskuse, což značí, že tyto texty nesou rysy mluvenosti / zapojenosti, naopak negativní faktorové skóre vykazují informačně-popisné texty a texty narativní (jako jsou např. historické články).

(2) Zpracování mluveného textu

Tato dimenze odráží podobně jako první dimenze rysy mluvenosti. Jazykové rysy s pozitivní faktorovou zátěží tak zahrnují slovesa existence nebo myšlení, různé slovesné vazby, elize *that* apod. Lehce negativní faktorovou zátěž mají dějová slovesa, vlastní jména a hodnoty type-token ratio. Vysokých faktorových skóre dosahují básně / písňové texty, interaktivní diskuse a přepisy mluvených textů, negativní pak informačně-popisné texty a narativní texty, tedy podobně, jako je tomu u první dimenze.

(3) Mluveně-narativní vs. psané informace

[18] Charakteristiky dimenzí v této práci nezahrnují veškeré rysy s prominentní faktorovou zátěží, ani všechny textové třídy spojené s výraznými hodnotami faktorového skóre. Pro souhrn byly vybrány nejdůležitější aspekty konstituce dimenzí. (Pro podrobnou deskripci viz Biber–Egbert, 2016, a Cvrček et al., 2018a, nebo 2018b.)

I třetí dimenze je spojena s opozicí psanosti a mluvenosti, tentokrát je akcentován rozdíl mezi narativními a popisnými texty. Pozitivní faktorová zátěž je kromě sloves spojena s příslovci a příslovečnými vazbami nebo zájmeny první osoby, negativní pak s dlouhými slovy či frekvencí substantiv. Kladné faktorové skóre je zde spojeno s narativními registry, jako jsou básně / písňové texty, interaktivní diskuse, přepisy mluvených textů a konečně narativní texty (např. umělecká próza), negativní naopak s registry informačními, jako jsou informačně-persvazivní texty a informačně-popisné texty.

(4) Nepřímá komunikace, „hlášení“

Pozitivní faktorová zátěž čtvrté dimenze je spojená zejména se subregistrem zpravodajských článků nebo blogů a odráží jejich specifika. Ta spočívají především ve způsobech užívání sloves v hlavních větách a jejich roli při uvozování vět vedlejších. Naopak negativní faktorovou zátěž vykazují nominalizace. Kromě zpravodajských (tedy narativních) textů jsou vysoká skóre typická i pro přepisy mluvených textů. Na druhém konci škály se pak nacházejí informačně-popisné texty, návody / instruktážní texty a informačně-persvazivní texty.

(5) Nerealizovanost (irrealis) vs. informační narativita

Pátá dimenze byla při interpretaci spojena se způsobem referování k událostem, které popisuje analyzovaný komunikát. Rozdíly tkví především v rozdílném užití slovesných způsobů, které značí, zda popisovaný děj již proběhl. Realizovanost je reprezentována nejčastěji indikativem, nerealizovanost např. kondicionálem. Pozitivní faktorová zátěž je zde spojena s modálními slovesy, kondicionálem, jinými časy, než je čas přítomný, nebo zájmeny pro druhou osobu. Negativní faktorovou zátěž vykazují slovesa v minulém čase či předložkové vazby. Nejvyšších faktorových skóre dosahují umělecké texty (konkrétně písňové texty), návody a instruktážní texty, dále např. přepisy mluvených textů. Na druhém pólu stojí dle očekávání texty informačně-popisné a narativní, nejvýrazněji subregistry popisující minulé události, jako je cestovní blog nebo historický článek.

(6) Vysvětlení/popis procedury

Šestá dimenze zachycuje rozdíly ve strategiích vysvětlování, kladnou faktorovou zátěž vykazují kauzativa, dokonavá slovesa či substantiva označující procesy, negativní pak vlastní jména. Vysokých pozitivních faktorových skóre dosahují návody a instruktážní texty a informačně-popisné texty, negativních přepisy mluvených textů nebo básně / písňové texty.

(7) Postojovost

V sedmé dimenzi je nejpříznakovějším subregistrem vědecký článek. Důvodem je vysoký výskyt substantiv značících postoj (např. substantiva označující kognitivní procesy a stavy) a jejich užití např. v předložkových frázích. Vysoké faktorové skóre je typické kromě informačně-popisných textů také pro texty názorové.

(8) Popis osob

Osmá dimenze odráží jazykové prostředky užívané k popisu osob, které jsou realizovány specifickými substantivy, zájmeny pro třetí osobu, vztažnými větami nebo neurčitými členy, které mají pozitivní hodnoty faktorové zátěže. Výrazných pozitivních faktorových skóre dosahují narativní texty (zejména umělecká próza) a informačně-persvazivní texty.

(9) Popis (nikoli technický)

Poslední, devátá dimenze je spojena s rozdíly v popisných textech, a to zejména v použití substantiv a nominalizací. Substantiva (zejména konkréta) zde mají pozitivní faktorovou zátěž, nominalizace negativní. Nejvyšších faktorových skóre dosahují návody / instruktážní texty (zejména recepty), nízkých naopak texty informačně-persvazivní a básně / písňové texty

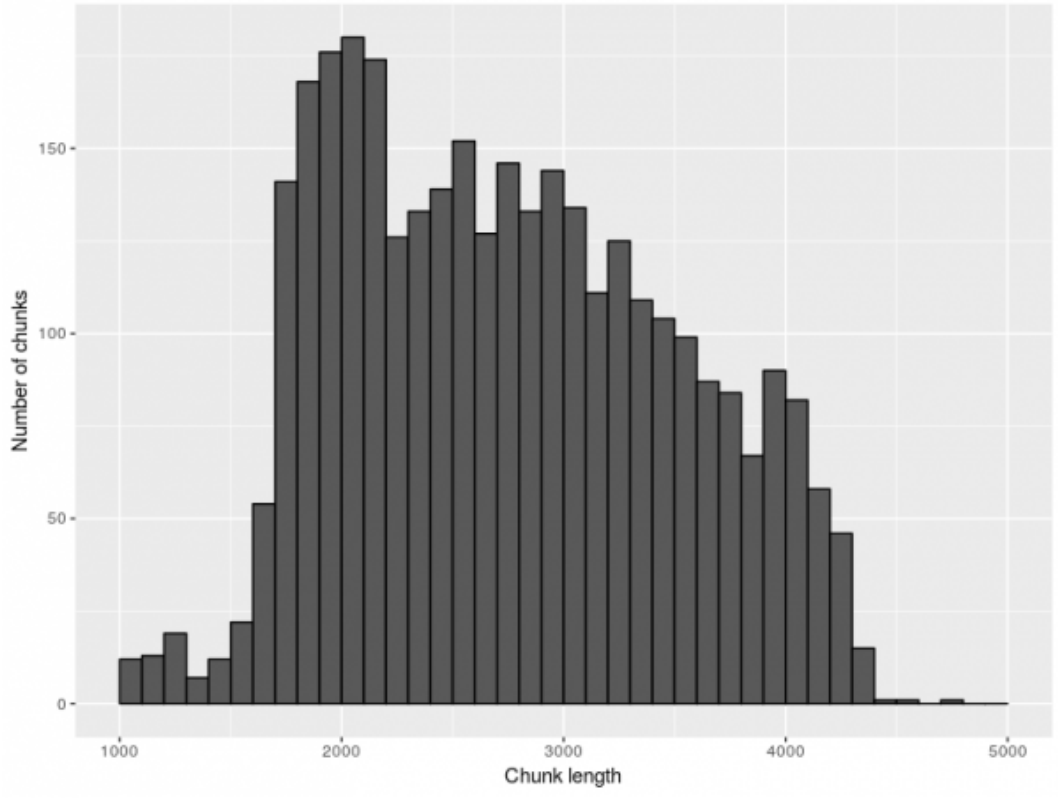
Multidimenzionální analýza internetových textů přispívá k důkazům o univerzálních dimenzích variability, přítomny jsou jak opozice mluvených a psaných textů, tak opozice v míře jejich narativnosti. Význam studie však spočívá zejména v představení nového přístupu k popisu registrové variability internetových textů, náročnosti manuální anotace a konečně důsledném provedení MDA, z níž vzchází mnoho mimořádných poznatků.

2.6.6 Aplikace MDA na český materiál

Na českém materiálu byla MDA registrové variability provedena v návaznosti na nepublikovaný výzkum Viléma Kodýtka, který vycházel z prací

Douglase Bibera a byl prvním, kdo se o MDA češtiny pokusil. Výzkum provedený pracovníky a pracovnicemi Ústavu Českého národního korpusu FF UK byl publikován v roce 2018 (Cvrček et al., 2018a, 2018b).

Materiálem pro analýzu byly texty z korpusu Koditex,[19] který byl pro tento účel vytvořen. Korpus obsahuje 9 mil. tokenů (s interpunkcí 10,9 mil. tokenů), které tvoří 3 334 textových vzorků (Cvrček et al., 2018b, s. 296). Většina zdrojových dat pochází z korpusů Českého národního korpusu, ostatní data, jako např. internetové texty, byla sebrána z jiných zdrojů (podrobně viz tamtéž, s. 296–297). Velikost vzorků v korpusu se pohybuje od 2 000–5 000 slov. Některé druhy textů typicky nedosahují ani spodní hranice tohoto rozsahu (např. příspěvky na sociálních sítích), proto byly jednotlivé texty konkatenovány. Naopak delší texty musely být na vyhovující velikost vzorkovány. Distribuce délek textových vzorků je uvedena v grafu 2.



Graf 2: Histogram distribuce délek textových vzorků v korpusu Koditex.[20]

[19] Název Koditex je akronymem názvu *Korpus diverzifikovaných textů* a odkazuje ke jménu Viléma Kodýtka.

[20] Osa *y* reprezentuje počet textových vzorků v korpusu, osa *x* délku textu v tokenech. Graf je dostupný on-line na <http://wiki.korpus.cz/doku.php/cnk:koditex> (cit. 11. 7. 2019).

Primárním cílem při sběru dat bylo zajistit co nejvyšší diverzitu vzorků, nikoli dosáhnout reprezentativnosti korpusu v tradičním slova smyslu. Tato snaha se odráží v bohaté vnětextové kategorizaci. Pro detailní rozlišení textů je vytvořena čtyřúrovňová kategorizace. Základní klasifikace vymezuje tři módy, které se dále dělí na divize, divize na nadtřídy a nadtřídy na třídy. Tři hlavní módy reprezentují mluvenou komunikaci, internetovou komunikaci a psanou komunikaci. Tato distinkce je v souladu s představami o internetovém jazyku jako útvaru nacházejícím se mezi mluveným a psaným jazykem. Divize rozlišují mluvenou komunikaci na interaktivní a neinteraktivní, internetovou komunikaci na jednosměrnou a mnohosměrnou a psanou komunikaci na beletrii, oborovou literaturu, publicistiku a soukromou komunikaci (korespondenci). Podrobnější dělení na nadtřídy a třídy je uvedeno v příloze 1.

Seznam jazykových rysů má základ v českých mluvnicích a stylistických příručkách, dotvořen je samotnými autory a autorkami. V prvotní fázi se operovalo až se 160 rysy, postupně byl však tento počet redukován až na konečných 122 rysů, které vstupovaly do faktorové analýzy. Rysy se dotýkají všech relevantních jazykových rovin, přítomny jsou rysy fonologické, lexikální, slovotvorné, morfologické a syntaktické. Hodnoceny jsou i ukazatele spadající do pragmatiky. Poslední skupinu tvoří obecnější textové charakteristiky, jako je poměr typů a tokenů v textu, hodnoty tematické koncentrace nebo koeficient lexikální repetitivnosti. Úplný seznam jazykových rysů použitých v tomto výzkumu je uveden v příloze 2. Hodnoty všech jazykových rysů byly relativizovány pomocí míry zTTR (Cvrček–Chlumská, 2015), která standardizuje hodnoty tak, aby byly nezávislé na délce textů, z nichž jsou naměřeny.

Pro faktorovou analýzu byla zvolena kosá rotace *promax* a počet faktorů byl stanoven na osm. Určení finálního počtu faktorů bylo výsledkem poměrně složitého procesu, který je detailně popsán ve studii *From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA* (Cvrček et al., 2018a). Tento proces je založen na srovnávání potenciálně vhodných řešení s hypotetickým řešením ideálním, a to na základě metrik z oblasti teorie

informace. Osmifaktorový model vysvětluje 56 % variace v datasetu, což je hodnota přibližně srovnatelná s Biberovými výzkumy (Cvrček, 2018b, s. 302).

Po analýze vysokých absolutních faktorových zátěží i inertních rysů, zvážení metainformací, kontrole konkrétních textových vzorků atd. (podrobně viz tamtéž, s. 301–302) docházejí autoři a autorky k následujícím interpretacím osmi dimenzí registrové variability:

(1) Dynamický vs. statický

První dimenze vysvětluje největší podíl variability v korpusu. Popisuje zejména poměr užívání sloves a jmen. Pozitivní faktorové zátěže dosahují rysy spojené se slovesy — jejich frekvence obecně, užívání indikativu, první osoby, negace či slovesa myšlení a mluvení, dále pak použití zájmen (zejména pro 1. a 3. osobu), spojek a adverbií. Negativní faktorovou zátěž vykazují substantiva a adjektiva obecně, dlouhá slova, sekundární předložky, pasivum a vyšší hodnoty tematické koncentrace. Pozitivních faktorových skóre dosahují především narativní texty, jako jsou romány, soukromá korespondence nebo webová fóra. Nízká skóre pak spojují texty administrativní, vědecké a encyklopedické.

(2) Spontánní vs. připravený

K poznatkům o funkci komunikátu ovlivněného situačním kontextem přispívají hodnoty druhé dimenze. Spontaneita textu je reprezentována korelujícími jazykovými rysy dosahujícími pozitivních hodnot faktorové zátěže, jako jsou osobní zájmena, kontaktové výrazy, nestandardní hláskoslovné varianty, výplňková slova či vysoká míra lexikální repetitivnosti. Negativní faktorovou zátěž vykazují věty s interogativními a vztažnými adverbii, delší slova, množství substantiv nebo lexikální bohatost reprezentovaná hodnotou zTTR.

(3) Vyšší vs. nižší stupeň koheze

Texty, lišící se v hodnotách třetí dimenze, jsou rozdílné především různým užíváním spojovacích a odkazovacích prostředků. Pozitivních hodnot faktorové zátěže dosahují vztažné věty se zájmenem *který*, korelativa, přivlastňovací zájmena, jmenné přísudky se substantivem nebo slovesa v přítomném čase a kondicionálu. Nejvyšších faktorových skóre dosahují předem připravené projevy, vědecké texty z oblasti společenských věd

a diskusní pořady. Na opačné straně pak stojí encyklopedické texty nebo běžná konverzace.

(4) Polytematický vs. monotematický

Hodnoty čtvrté dimenze charakterizují lexikální bohatost komunikátu. Pozitivně zde koreluje diverzita unigramů a bigramů, šíře inventáře předložek nebo zájmen, negativně naopak tematická koncentrace, lexikální repetitivnost, verbální substantiva, pasivum, abstrakta nebo sekundární přeložky. Polytematické texty zastupují především časopisy, ať už lifestylové nebo zájmové, za monotematické můžeme považovat texty vědecké a administrativní.

(5) Vyšší vs. nižší míra explicitní adresnosti

Pozitivní hodnoty faktorové zátěže páté dimenze odrážejí zvýšenou frekvenci vykřičníků a otazníků, slovesa v 2. osobě, budoucí čas, vokativ nebo imperativ. Tyto rysy reprezentují vyšší míru explicitní adresnosti komunikátu. Negativní faktorovou zátěž zde jeví pouze délka věty v tokenech. Pozitivní faktorová skóre jsou spojena s uměleckými texty (scénáři a romány různých druhů) a s mluvenou komunikací.

(6) Obecný vs. konkrétní

V šesté dimenzi pozitivně korelují sémanticky vyprázdněná adjektiva, koordinace a konjunkce, negativně pak toponyma, antroponyma, sekvence substantiv nebo číslovky. Distinkce, kterou nabízí tato dimenze, neodlišuje vnětextové kategorie příliš výrazně, přesto jsou z textů zjevné rozdíly mezi texty, které odpovídají protikladu obecnosti a konkrétnosti. Za obecné jsou považovány např. encyklopedické texty popisného charakteru, za konkrétní pak texty s výčty osob, čísel nebo jiných údajů.

(7) Prospektivní vs. retrospektivní

Sedmá dimenze zachycuje zejména rozdíly ve slovesných kategoriích. Prospektivní texty pojednávají o přítomnosti a budoucnosti, případně nespecifikují čas vůbec, retrospektivní texty jsou naopak spojeny s odkazováním do minulosti. Kromě prézentu a futura nesou pozitivní faktorovou zátěž také rysy druhé osoby a imperativ, negativní zátěž pak třetí osoba a přivlastňovací adjektiva a zájmena. Prospektivními se ukázaly být

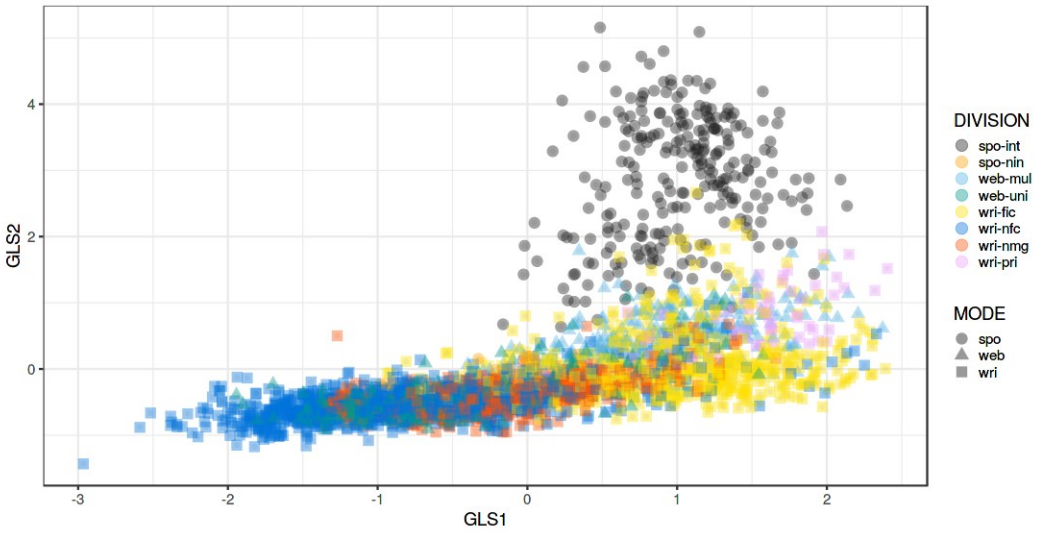
texty na webových fórech a některé texty mluvené komunikace, retrospektivními naopak texty beletristické.

(8) Postojovost vs. faktuálnost

S osmou dimenzí je spjata opozice postojovosti a faktuálnosti. Vliv na konstituci této dimenze mají především rysy částic, a to restriktorů, intenzifikátorů, strukturačních částic nebo částic oslabujících modalitu. Pozitivní faktorovou zátěž vykazují také adverbia, jejich shluky a stupňované tvary. Negativní zátěž je spojena jen s rysem koordinace. Kladné skóre získávají texty vyjadřující názory (např. mluvené texty), a záporné skóre texty s výraznou snahou o objektivitu (výčty, recepty apod.)(tamtéž).

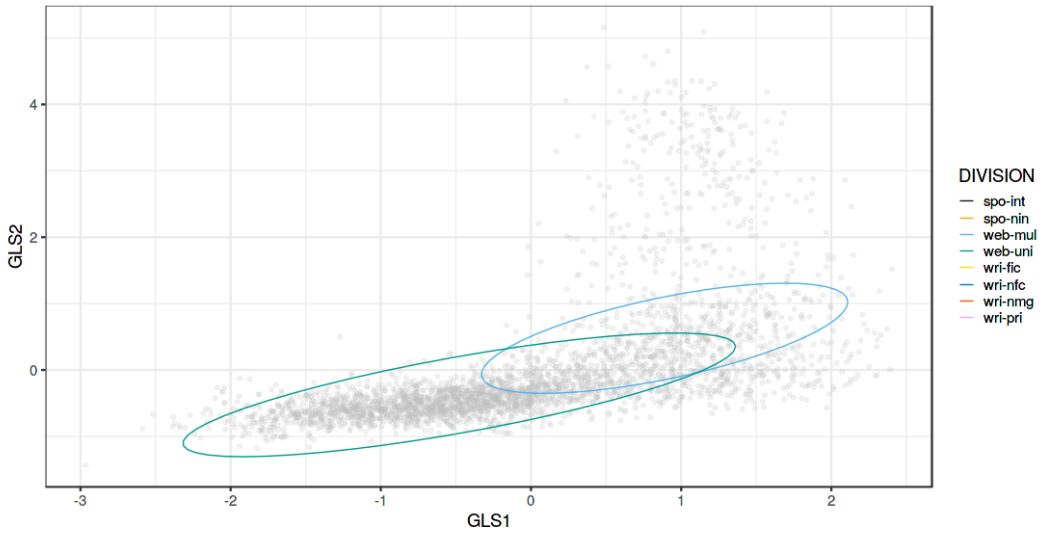
MDA češtiny publikovaná Cvrčkem et al. (2018a, 2018b) je unikátním příspěvkem k výzkumu variability češtiny, protože obohacuje dosavadní stav bádání v oblasti české stylistiky o kvantitativní, korpusově orientované pojetí. Zasazením do vícejazyčného kontextu pak přispívá k poznání variability v jazyce vůbec.

Zahrnutím různých vzorků internetových textů do MDA češtiny je možné poukázat na smíšenou povahu CMC z hlediska psanosti a mluvenosti. Tu lze demonstrovat graficky promítnutím faktorových skóre jednotlivých textů do dvourozměrného grafu. Grafy č. 3 a č. 4 vyobrazují hodnoty pro první a druhou dimenzi variability.[21] Osa *x* reprezentuje škálu dynamických (+) a statických (–) textů, osa *y* rozlišuje texty spontánní (+) vs. připravené (–).



Graf 3: Faktorová skóre textů z korpusu Koditex.

[21] Oba grafy jsou vygenerovány nástrojem MDAvis (Lukeš, 2018)(cit. 12. 7. 2019).



Graf 4: Kategorie webových textů z korpusu Koditex.

Graf č. 3 poukazuje na překryvy psaných a webových textů v druhé dimenzi (spodní část grafu). Mluvené texty se vymykají svými skóre spontánnosti. V první dimenzi se však webové texty (zvýrazněné v grafu č. 4) překrývají s texty mluvenými a psanými velmi výrazně. Mapování prostoru internetových textů, přibližně reprezentovaného grafem č. 4, se věnuje praktická část této práce.

3 Registrová variabilita textů na českém internetu

Následující analýza je pokusem o podrobné popsání registrové variability textů na českých webových stránkách. Východiskem této analýzy je propojení manuální anotace webových registrů inspirované studií Bibera a Egberta (2016) s analýzou dimenzí variability vymezených pro češtinu Cvrčkem et al. (2018a, 2018b).

3.1 Data

Pro analýzu byla zvolena data z korpusu Araneum Bohemicum Maximum. Jedná se o web-crawled korpus o velikosti 3,3 miliardy tokenů, který byl vytvořen v průběhu roku 2013 (Benko, 2014) jakožto součást srovnatelné (vícejazyčné) rodiny korpusů Aranea (tamtéž).

Sestavení web-crawled korpusu vyžaduje využití relativně velkého množství nástrojů, které stahují data z internetu, získávají čistý text (*plain*

text) z HTML souborů, sjednocují jeho kódování, identifikují jeho jazyk apod. Prostřednictvím dalších nástrojů jsou pak získané texty tokenizovány, lemmatizovány a značkovány.

Korpus Araneum Bohemicum Maximum obsahuje texty excerpované především z webových stránek s národní doménou nejvyššího řádu (*cz*), zařazeny ale byly i českojazyčné texty z generických domén nejvyššího řádu, jako je *com*, *org*, *net* apod.

Pro manuální anotaci byl vytvořen subkorpus o velikosti 1 000 textových vzorků, které distribucí délek odpovídají korpusu Koditex (viz graf č. 2 výše). Většina délek textů se pohybuje kolem 1500–4500 tokenů, což je ověřená délka vhodná pro MDA. (Biber, 1990, s. 261). Velikost subkorpusu je 3 232 975 tokenů, průměrná velikost textu tak činí přibližně 3 232 tokenů.

Pro každý z textů jsou dostupné metainformace, jako je délka textu v tokenech, datum stažení z internetu, doména nejvyššího řádu, doménové jméno nebo URL, tj. internetová adresa, na které se text v době crawlingu nalézal.

3.2 Anotace dat

Jako podklad pro vnětextovou anotaci textových vzorků sloužila subregistrová kategorizace podle Bibera a Egberta (2016, s. 105–106, viz výše tabulka 1). Autoři vymezují 27 subregistrů, jejichž označení lze na této rovině obecnosti ztotožnit s žánrovými charakteristikami,[22] a jsou tedy vhodné pro kategorizaci textových vzorků na základě rozpoznatelných vnětextových parametrů. Nadřazené registry nebyly zahrnuty, a to zejména pro jejich přílišnou obecnost, z které pramenily neshody v anotacích, jež byly později interpretovány jako důkaz o existenci hybridních registrů.

Výchozí sada 27 subregistrů byla revidována ve dvou fázích tak, aby co nejlépe vystihovala podobu českého webu a zároveň byla každá kategorie v datasetu dostatečně zastoupena. První fáze revize byla provedena ještě před započetím samotné anotace. Testování, které proběhlo na vzorku 50 textů, mělo vést ke zjištění, zda jsou kategorie dostačující a zda je možné každý

[22] Tato práce se přidržuje termínu subregistr, jak jej užívají Biber a Egbert (2016) ve smyslu souboru jazykových charakteristik žánru (viz např. Mrázková, 2017).

z textů zařadit do některé z nich. Výsledkem tohoto iniciálního testování bylo přidání dvou subregistrů, a to právních textů (*LEG*) a textů z oblasti hobby a volného času (*HOB*). První kategorie odráží vysoký výskyt stránek s obchodními podmínkami, právními podmínkami využití různých služeb apod., druhá nedostatečné možnosti kategorizace textů na on-line lifestylových magazínech a dalších zájmových webech. Kategorie náboženský blog / kázání (*REL*) byla rozšířena obecně na texty zabývající se duchovnem, přibyly do ní tedy texty filozofické a také ezoterické, parapsychologické apod. Při rychlém čtení těchto textů v rámci anotace by mnohdy bylo obtížné je rozlišit a spolehlivě zařadit do separátních kategorií. Zrušen byl subregistr editorialu, protože na webu nebývá dostatečně explicitně signalizován a do velké míry se překrývá s kategorií informačního textu,[23] se kterou byl ve výsledku smíšen.

Cílem anotačního procesu bylo přidat ke každému z textových vzorků dvě informace. První z nich určuje, zda je text v korpusu stále dostupný na URL obsaženém v metainformacích korpusu. Druhou informací je pak náležitost textu k subregistru.

Dostupnost každého z textových vzorků byla ověřena zadáním odkazu z metainformace do internetového prohlížeče. Za nedostupné byly považovány texty na adresách, po jejichž zadání prohlížeč vracel chyby různého typu, zpravidla chyby 404 (dokument nenalezen), ale také chyby 403 (přístup odepřen) nebo 500 (chyba serveru). Dále byly jako nedostupné označeny odkazy na weby napadené počítačovými viry (tj. ty, které byly zablokovány antivirovým programem) a odkazy, které vedly na stránky, které obsahovaly jiný text než text vzorku. Shodnost texu na webu a vzorku v korpusu byla kontrolována prostřednictvím zobrazení v konkordančních řádcích v rozhraní KonText[24] (Machálek, 2014). Nahlížení do textů v KonTextu sloužilo také k analýze subregistrů u textů, které byly nedostupné v prohlížeči. Anotace probíhala na přelomu června a července 2019, všechny hodnoty tak odpovídají stavu webu v tomto období.

[23] Prototypickou povahu editorialu, kde se mísí informační obsah s projevováním názoru redakce/pisatele vykazovalo pouze několik jednotek textů.

[24] Prostřednictvím konkordančních řádků bylo možné zobrazit celý text z korpusu bez omezení.

Při zařazování textových vzorků do subregistrových kategorií byly kladeny nároky především na zachycení textových i netextových žánrových markerů. Práci usnadňovalo zobrazení stránek v prohlížeči, kde bylo možné text lépe specifikovat díky kontextu na základě nadpisů, členění nebo adresy.[25] V případě smíšené povahy textů byl vybrán převládající subregistr. Nelze předpokládat naprostou správnost anotace, jelikož byla prováděna jen jednou osobou a není tak možné ověřit její konzistentnost změřením mezianotátorské shody. Analýza má však povahu pilotní studie a jejím účelem je ověřit, zda je zvolená metoda vhodná a směrodatná pro budoucí výzkum ve větším měřítku. Jako problematické se jevily texty smíšené povahy a texty marginálních žánrů, které plně neodpovídají žádnému z vymezených subregistrů, ale zároveň nejsou dostatečně početné, aby tvořily vlastní subregistrovou kategorii. Sporné texty byly zařazeny do nejbližších kategorií na základě intuice.

V druhé fázi revize definovaných subregistrů bylo přihlíženo k početnímu zastoupení textů v jednotlivých subregistrových kategoriích a k možnosti jejich sloučení za účelem vyšší směrodatnosti výsledků. Subregistr receptů byl sloučen s instruktážními texty (*HOW*), výzkumné články byly sloučeny s encyklopedickými články do jedné kategorie (*ENC*). Popisy osob byly zařazeny do kategorie popis předmětu/věci (*DET*). Písňové texty byly zařazeny pod umělecké texty (*FIC*) a přepisy mluvených textů z televize byly rozřazeny mezi umělecké texty (*FIC*) a formální projevy (*FOR*). Do subregistru básní nebyl zařazen jediný text, výsledná sada subregistrů tak obsahuje 22 položek, které jsou uvedeny v tabulce 2.

#	**Subregistr**	**Zkratka**	**Poznámka**
1	Poradenský text	ADV	
2	Popis s účelem prodeje	DES	
3	Popis předmětu	DET	+ popis osoby
4	Diskusní fórum	DIS	
5	Encyklopedický článek	ENC	+ vědecký článek
6	Nejčastěji kladené dotazy (FAQs)	FAQ	
7	Umělecká próza	FIC	+ písňové texty
8	Formální projev	FOR	
9	Historický článek	HIS	
10	Hobby a volný čas	HOB	

[25] V adresách jsou mnohdy zahrnuty informace o rubrikách, v nichž se texty nacházejí.

11	Návod typu *jak na to*, instruktáž	HOW	
12	Informační text	INF	
13	Interview	INT	
14	Právní text	LEG	
15	Zpravodajský článek/blog	NEW	
16	Názorový blog	PEO	
17	Osobní blog	PER	
18	Fórum typu *question-answer*	QAF	
19	Náboženský blog / kázání	REL	+ ezoterie a parapsychologie
20	Recenze	REV	
21	Sportovní reportáž	SPO	
22	Cestovatelský blog	TRA	

Tabulka 2: Seznam subregistrů užitých pro anotaci vzorku textů.

3.3 Jazykové rysy a faktorová analýza

Seznam jazykových rysů, způsob jejich identifikace a parametry faktorové analýzy vycházejí z výzkumu Cvrčka et al. (2018a, 2018b)(viz výše kapitola 2.6.4 a příloha 2).[26] Pro každý z textů ve vzorku byly naměřeny hodnoty faktorového skóre. Na základě jejich sdružování podle vnětextové kategorizace (viz předchozí oddíl) a zpracování a vizualizace pomocí knihoven *dplyr* (Wickham et al., 2018) a *ggplot2* (Wickham, 2016) v programu R Studio (2015)[27] jsou interpretovány výsledky výzkumu.

3.4 Výsledky

3.4.1 Dostupnost textů

Z 1 000 textů v anotovaném vzorku bylo po přibližně šesti letech od crawlingu dostupných pouze 40,5 % textů. 59,5 % textů bylo z původních adres odstraněno. Toto svědčí o dynamické povaze a proměnlivosti obsahu internetu jakožto média. S podobným problémem se potýkali i Biber a Egbert (2016, s. 100), kteří zaznamenali nedostupnost 8 % webových stránek po 7 měsících od získání jejich URL. Nedostupnost textů z korpusu Araneum

[26]

[27] R Studio (2015) je grafickým rozhraním programovacího jazyka R (2018).

Bohemicum Maximum ztížila proces anotace, protože nedostupné texty musely být prohlíženy v prostředí KonText, což zvýšilo časovou náročnost práce.

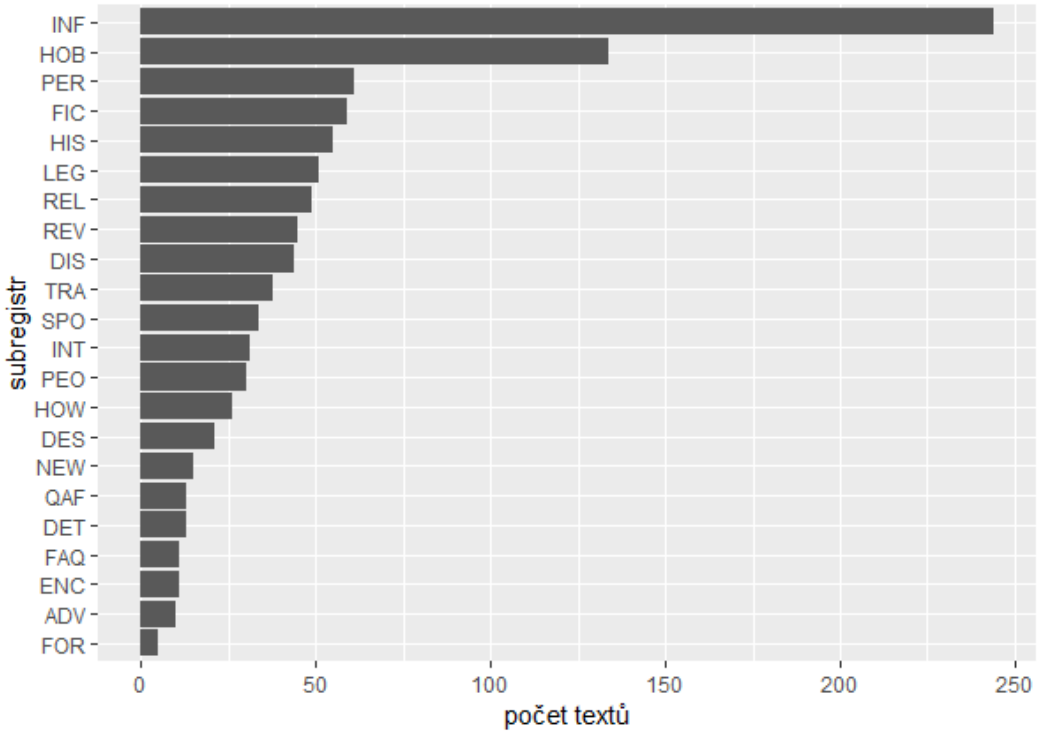
Z těchto zjištění vyvstává otázka o využitelnosti webových korpusů po delším časovém období od jejich vzniku. Analýza prezentovaná touto prací zachycuje stav českého internetu v roce 2013, pro rok 2019 ovšem ztrácí na reprezentativnosti. Korpusové výzkumy, které usilují o výsledky reprezentativní pro současnost, by měly analyzovat data z webových korpusů, které jsou co nejnovější.

3.4.2 Rozložení subregistrů

Téměř čtvrtina textů ze vzorku byla zařazena do kategorie informačních textů (*INF*). Spadá sem velké množství vzorků, jejichž obsahem jsou souhrny aktualit a informací o aktualizacích webů. Další, nezanedbatelnou část tvoří ostatní texty, jejichž primárním účelem je poskytnout informace. Toto obecné vymezení subregistru je důvodem jeho velkého obsazení. Jemnější dělení bylo žádoucí, avšak poměrně problematické, jelikož texty z této kategorie bývají rozmanité a důraz na informativnost je mnohdy jediné, co je spojuje. Díky zařazení subregistru hobby a volný čas (*HOB*) bylo možné lépe kategorizovat texty, jejichž značná část by spadala právě do informačních textů. Tato kategorie dokonce obsahuje druhý nejvyšší počet textů. Většina subregistrů je reprezentována přibližně 20–60 texty, mezi nejméně zastoupené subregistry patří formální projevy (*FOR*), poradenské texty (*ADV*), nejčastěji kladené dotazy (*FAQ*) a encyklopedické a výzkumné články (*ENC*). Podrobné informace o distribuci subregistrů jsou uvedeny v grafu č. 5 a tabulce č. 3.

subregistr	ADV	DES	DET	DIS	ENC	FAQ	FIC	FOR	HIS	HOB	HOW
poč. textů	10	21	13	44	11	11	59	5	55	134	26
	INF	INT	LEG	NEW	PEO	PER	QAF	REL	REV	SPO	TRA
	244	31	51	15	30	61	13	49	45	34	38

Tabulka 3: Distribuce subregistrů ve vzorku.

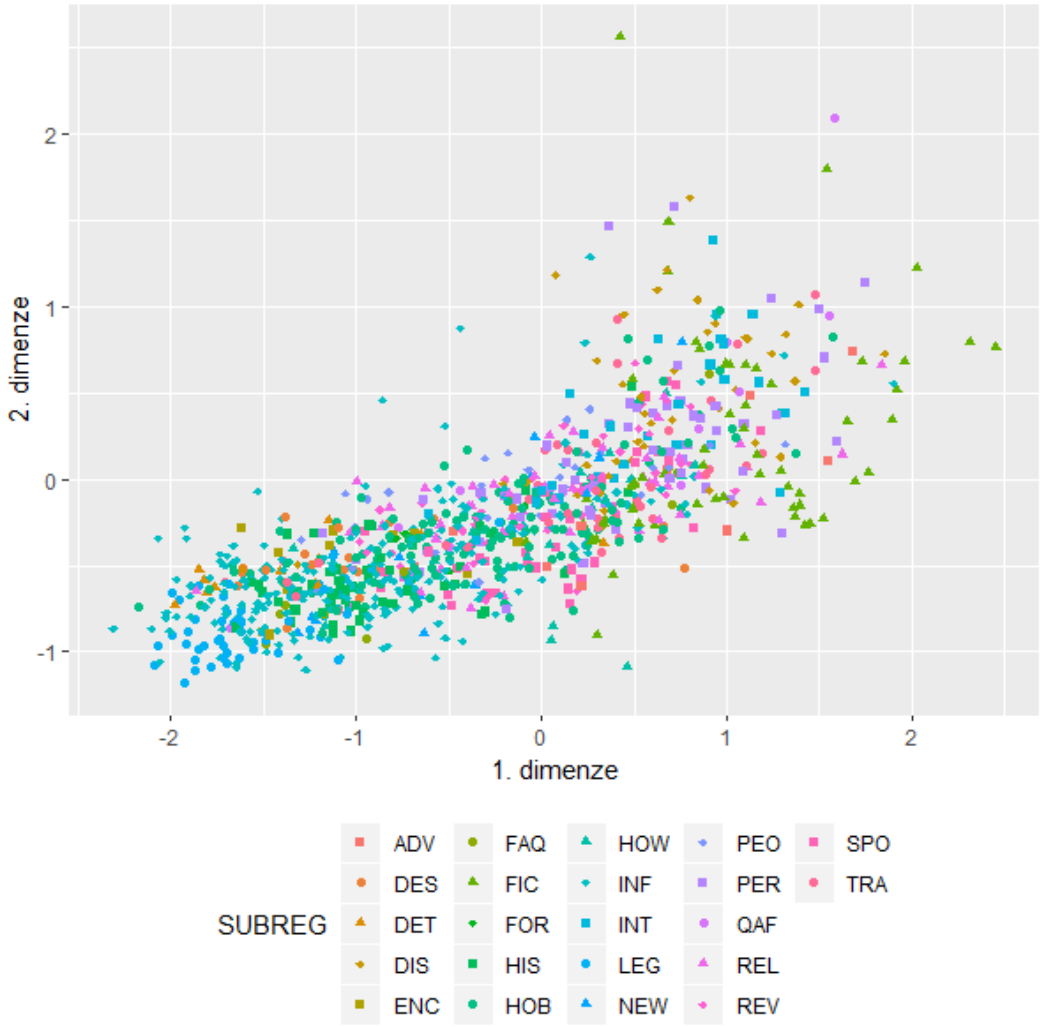


Graf 5: Distribuce subregistrů ve vzorku.

3.4.3 Celkový pohled

Promítneme-li proti sobě hodnoty první dimenze (dynamický vs. statický) a druhé dimenze (spontánní vs. připravený) do dvourozměrného grafu (graf č. 6), můžeme sledovat jasné tendence u některých subregistrů.

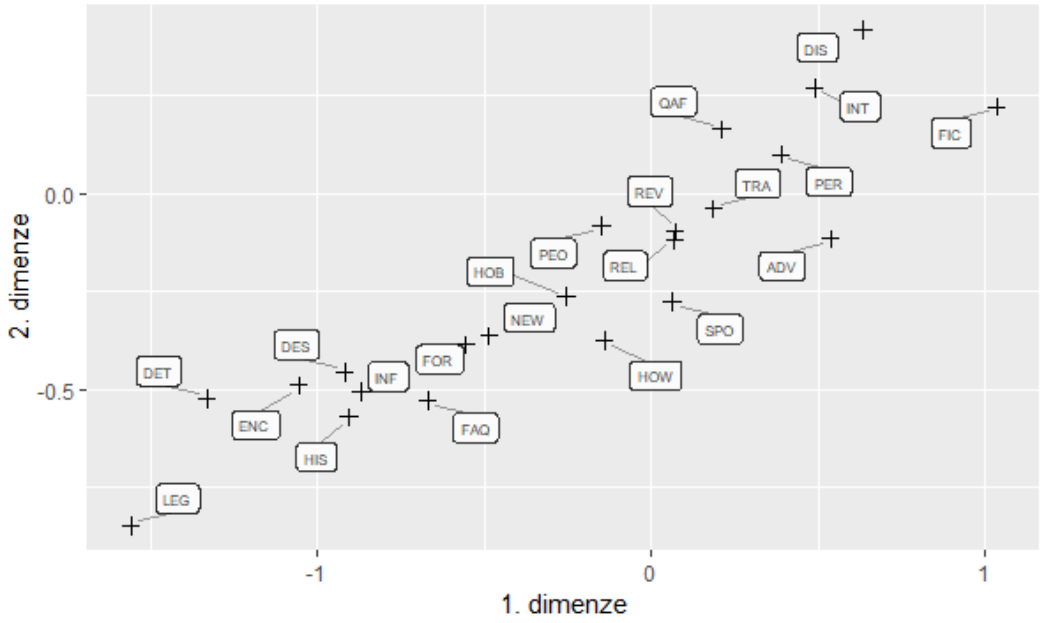
Vlevo dole jsou v grafu vyobrazeny subregistry, které jsou statické a připravené. V této oblasti je nápadný shluk právních textů (*LEG*), dále je zde znatelná převaha deskriptivních textů, jako je popis předmětu (*DET*) a popis s účelem prodeje (*DES*), výrazné je také zastoupení historických článků (*HIS*). Do středového pásma kolem nulových skóre pokračuje velký shluk informačních textů (*INF*) a textů z volnočasové oblasti (*HOB*). Vyšší dynamika a spontaneita je spojena s osobními (*PER*) nebo cestovatelskými blogy (*TRA*), nejvyšších skóre v obou dimenzích však dosahují interview (*INT*), diskusní fóra (*DIS*) a umělecké texty (*FIC*).



Graf 6: Faktorová skóre pro první a druhou dimenzi

Velký počet subregistrů ztěžuje orientaci ve dvourozměrném bodovém grafu. Pro lepší orientaci byly vypočítány centroidy jednotlivých datových clusterů[28], které vyobrazuje graf č. 7. Toto zobrazení je přehlednější, ale zcela zanedbává datový rozptyl. Možným řešením je rozložit data každého registru do samostatné části grafu. Tento složený graf je z důvodu většího rozsahu uveden jako příloha práce (příloha 3).

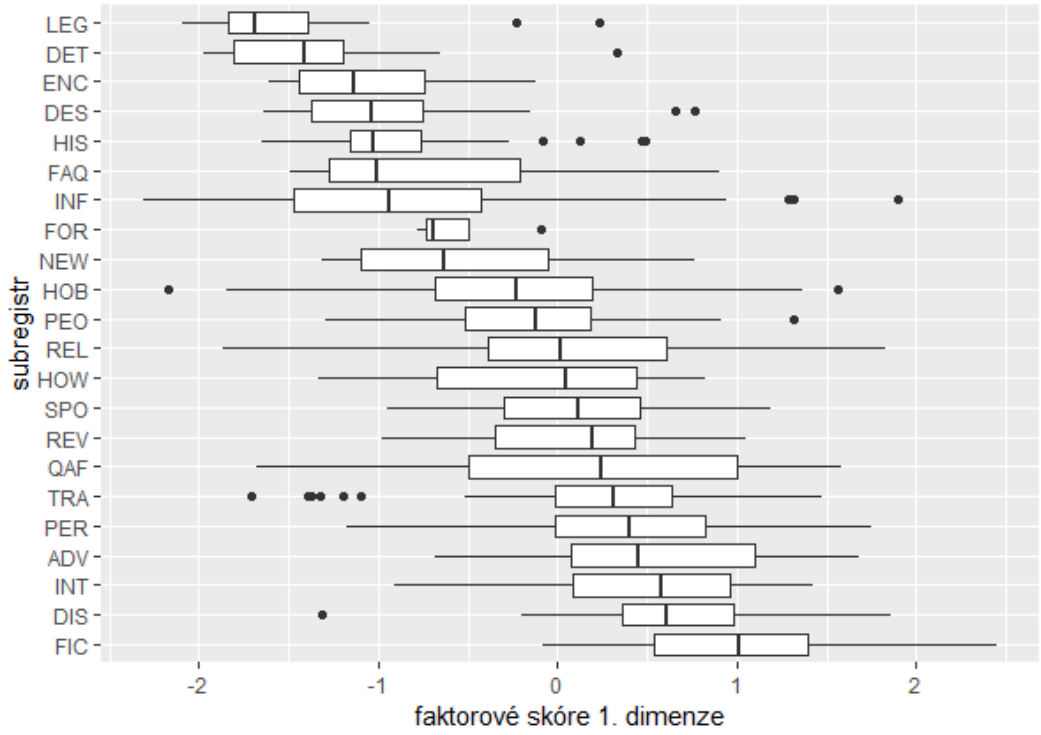
[28] Centroidy byly vypočítány podle mediánových hodnot skóre pro každou z dimenzí.



Graf 7: Centroidy subregistrů pro první a druhou dimenzi.

3.4.4 Jednotlivé dimenze variability

(1) Dynamický vs. statický



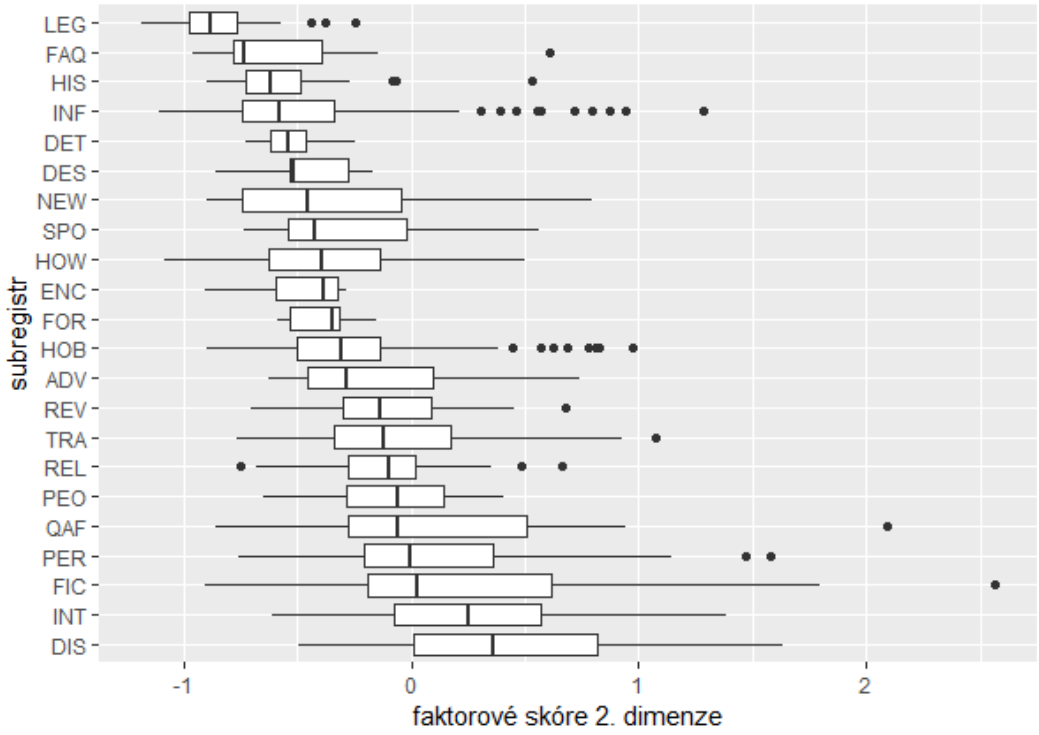
Graf 8: 1. dimenze variability: dynamický (+) vs. statický (–) text.[29]

[29] Krabicové grafy pro subregistry jsou seřazeny podle hodnot mediánu. Každý graf zobrazuje medián (tučná čára uprostřed obdélníku), mezikvartilový rozptyl (bílý obdélník) a linie reprezentující rozložení marginálnějších dat. Extrémní hodnoty jsou vyobrazeny pomocí bodů, tzv. *outliers*.

Jak bylo poznamenáno výše, nejnižších faktorových skóre dosahují právní texty (*LEG*), a jsou tedy nejstatičtější. Vytvoření subregistru právních textů se jeví být dobrým krokem, protože texty do něj zařazené by byly pravděpodobně zařazeny do informačních textů (*INF*), které ale v první dimenzi dosahují výrazně odlišných skóre (medián právních textů leží dokonce pod hranicí druhého kvartilu informačních textů). Nízká faktorová skóre vykazují také popisy předmětů (*DET*) a encyklopedické a výzkumné články (*ENC*). Právní, vědecké a encyklopedické texty jsou umístěny dle očekávání v této oblasti, podobné výsledky zaznamenávají i Cvrček et al. (2018b, s. 303).

Nejvyšší mírou dynamičnosti se vyznačují umělecké texty (*FIC*) a diskusní fóra (*DIS*), dále interview (*INT*) nebo poradenské texty (*ADV*). Umělecké texty (až na výjimky prozaické) a diskusní fóra ukazují podobné tendence jako výsledky Cvrčka et al. (tamtéž).

(2) Spontánní vs. připravený



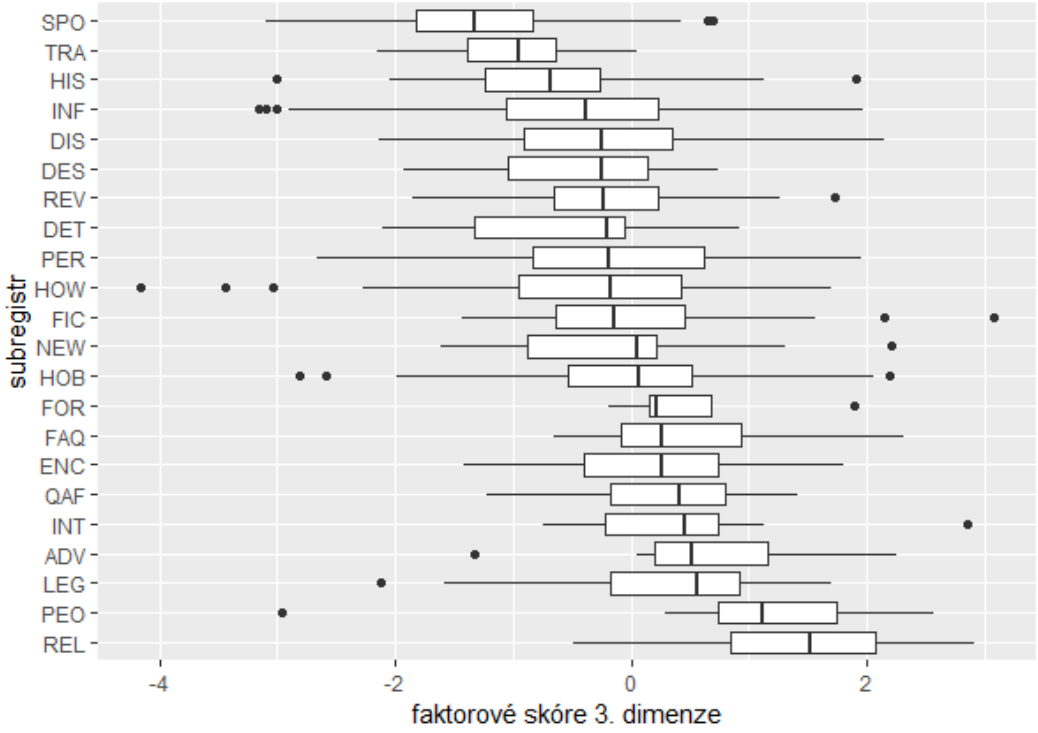
Graf 9: 2. dimenze variability: spontánní (+) vs. připravený (–) text.

Podobně jako u první dimenze stojí právní texty (*LEG*) mimo ostatní internetové komunikáty. Druhého nejnižšího skóre dosahuje subregistr nejčastěji kladených otázek (*FAQ*). Zajímavé je, že texty tohoto subregistru jsou stylizované jako dialog (otázky a odpovědi), od diskusního fóra nebo fóra

typu *question–answer* se však výrazně liší právě stylizací, která se odráží ve skóre připravenosti. Jako připravené byl také vyhodnoceny informační texty (*INF*) a popisy (*DES* a *DET*).

Nejspontánnější jsou naopak dle očekávání texty internetových diskusí (*DIS*), interview (*INT*), umělecké texty (*FIC*) a osobní blogy (*PER*).

(3) Vyšší vs. nižší stupeň koheze



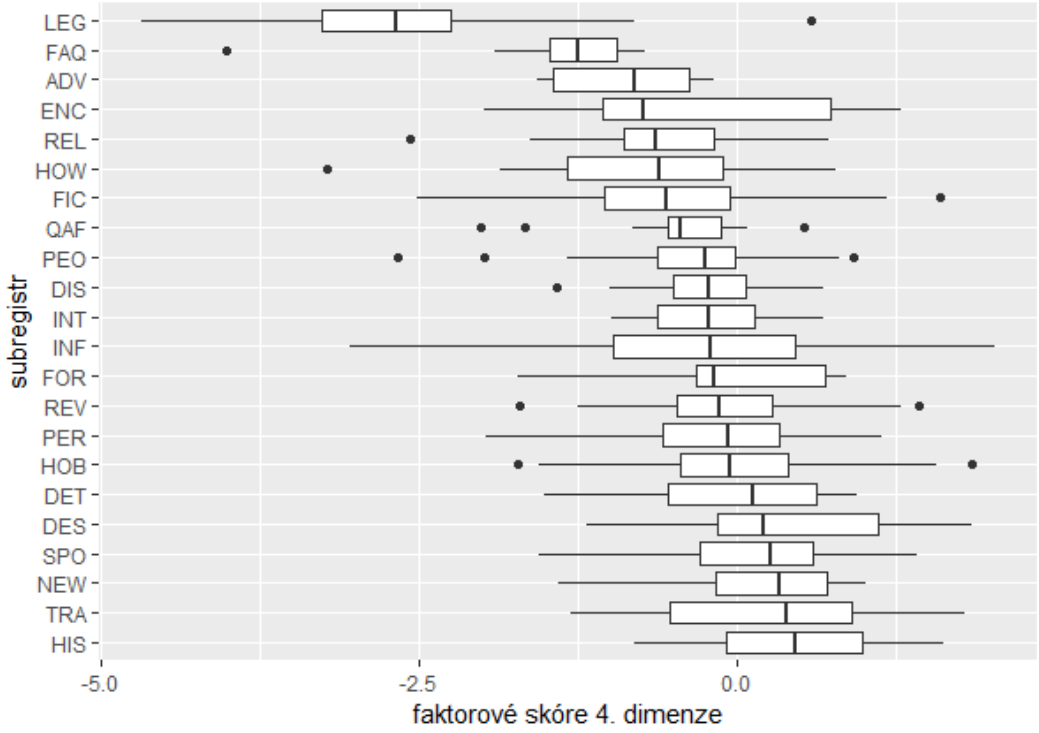
Graf 10: 3. dimenze variability: vyšší (+) vs. nižší (–) stupeň koheze textu.

Třetí dimenze odráží míru vnitrotextové reference. Crvček et al. uvádějí, že „vypovídá o tom, zda je při vytváření komunikátu pozornost zaměřena na sdělované skutečnosti, či spíše na jejich usouvztažnění“ (tamtéž, s. 306). Účelem sportovních článků (*SPO*) a reportáží je zpravidla informovat o proběhlých sportovních událostech, stejně tak subregistr historických článků (*HIS*)(obsahující velké množství textů týkajících se historie obcí nebo různých institucí) je spojený především s faktografií historických událostí. Cestovatelské blogy (*TRA*) jsou taktéž často vystavěny na pouhém sledu událostí. Není tedy překvapením, že se tyto texty ocitají v oblasti nižšího stupně koheze.

Naopak výrazně vyšší stupeň koheze vykazují dva subregistry: názorový blog (*PEO*) a náboženské, ezoterické a parapsychologické texty (*REL*).

Názorové blogy se v hodnotách skóre této dimenze znatelně odlišují od obecných osobních blogů (*PER*). Důvodem pozitivních skóre těchto dvou subregistrů je zřejmě zvýšená snaha odůvodňovat tvrzení jejich vzájemným usouvztažňováním a uváděním do kontextu, která je pro oba subregistry typická.

(4) Polytematický vs. monotematický

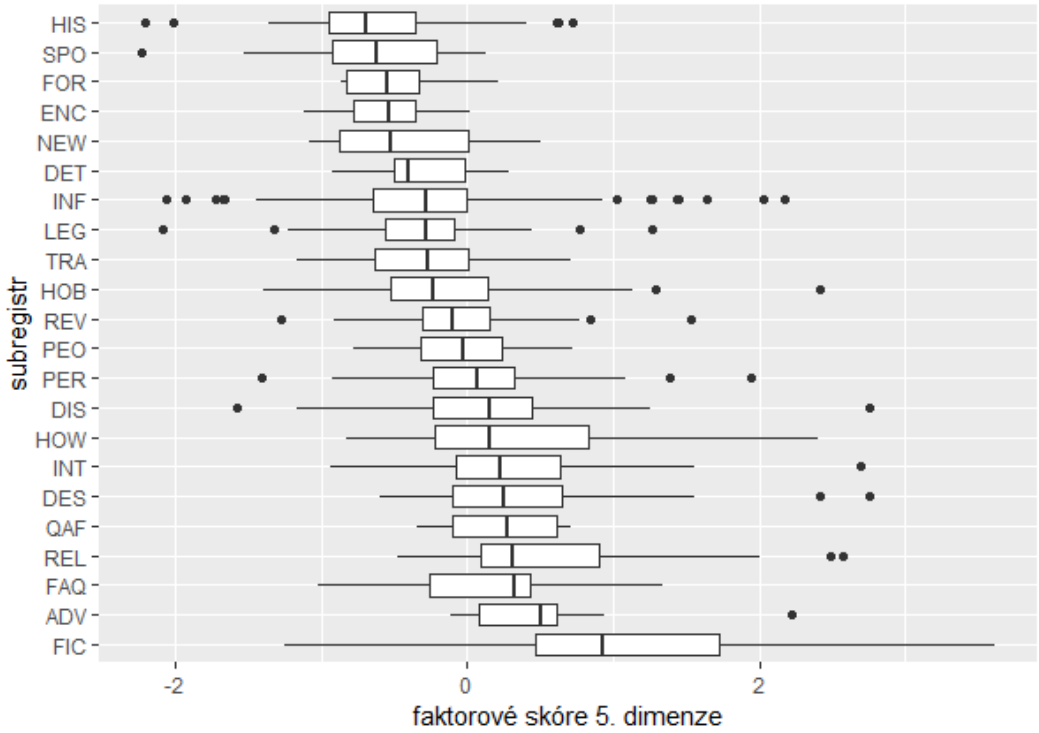


Graf 11: 4. dimenze variability: polytematický (+) vs. monotematický (–) text.

Ze všech skóre čtvrté dimenze, reprezentujících zejména míru bohatosti lexikonu, je nejvýraznější záporné skóre právních textů (*LEG*). Tyto texty jsou typické svým zaměřením na jedno téma, jistou repetitivností, omezeným lexikem a jazykovými prostředky, jako jsou sekundární předložky nebo pasivum. Tematickým vymezením jsou typické i nejčastěji kladené otázky (*FAQ*) či poradenské texty (*ADV*). Negativních hodnot této dimenze nabývá i většina encyklopedických a vědeckých článků (*ENC*) (srov. tamtéž, s. 306).

Jako polytematické byly vyhodnoceny historické články (*HIS*), cestovatelské blogy (*TRA*) nebo zpravodajské články (*NEW*).

(5) Vyšší vs. nižší míra explicitní adresnosti

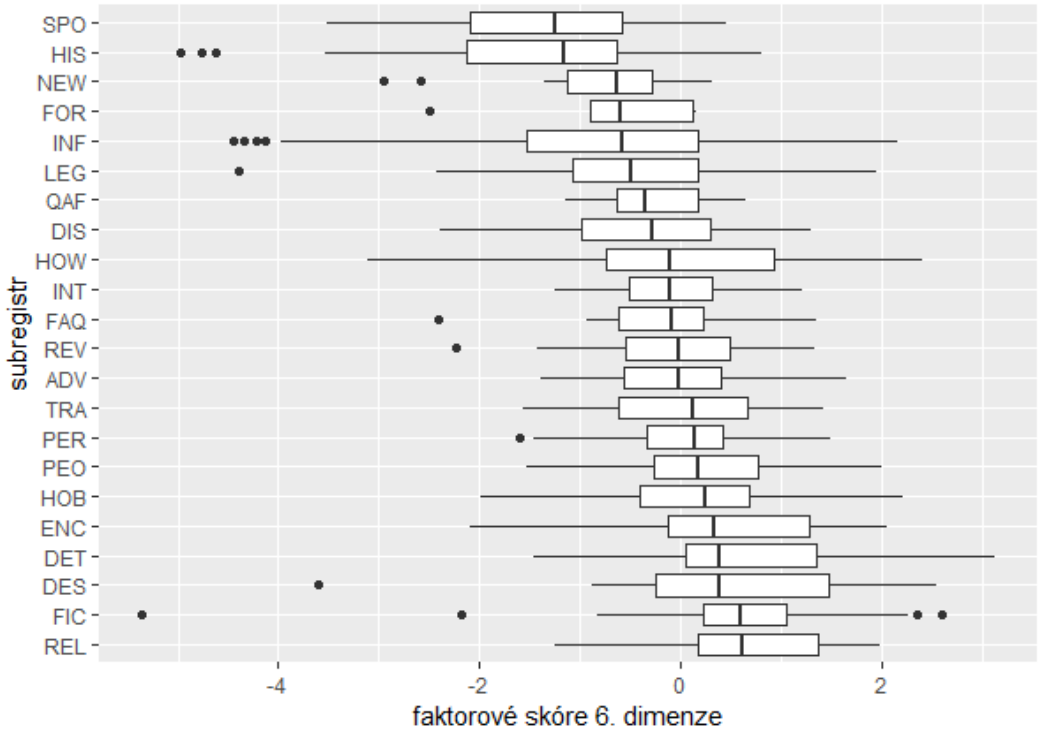


Graf 12: 5. dimenze variability: vyšší (+) vs. nižší (–) míra explicitní adresnosti textu.

Pátá dimenze je spojena s kódováním adresáta a její hodnoty jsou ovlivněny zejména přítomností komunikačního partnera v rámci komunikačního aktu. Vysokou míru adresnosti vykazují texty spadající do subregistru uměleckých textů (*FIC*), a to pravděpodobně díky obsahu přímé řeči, která v internetových textech nebývá častá. Na adresáta se přímo obracejí autoři poradenských textů (*ADV*).

V záporných hodnotách se dle očekávání nacházejí texty, v kterých adresát není oslovován vůbec, jako jsou historické články (*HIS*), sportovní reportáže a články (*SPO*) nebo encyklopedické a výzkumné články (*ENC*) (srov. tamtéž, s. 307).

(6) Obecný vs. konkrétní



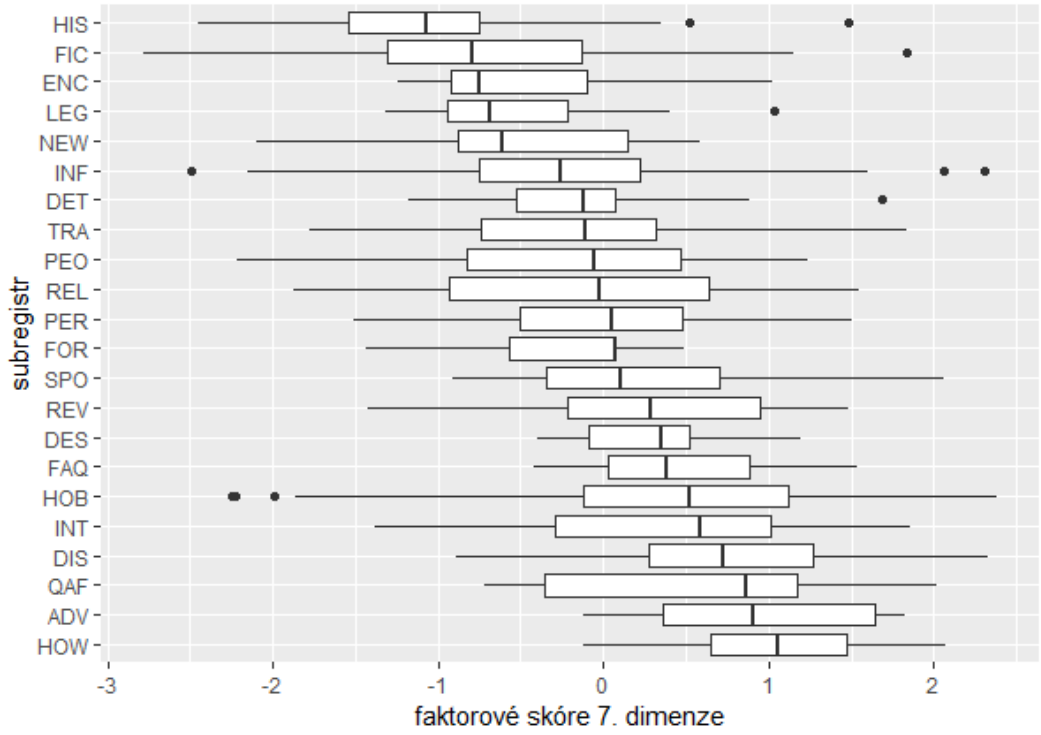
Graf 13: 6. dimenze variability: obecný (+) vs. konkrétní (–) text.

Pozitivní faktorové skóre v 6. dimenzi variability je spojeno s rysy jako jsou koordinace nebo sémanticky vyprázdněná adjektiva, které svědčí o obecném rázu textů, konkrétní texty se naopak vyznačují zvýšeným výskytem antroponym, toponym, číslovek apod.

Jako výrazně konkrétní byly vyhodnoceny texty ze subregistrů historických a sportovních článků (*HIS* a *SPO*). Oba subregistry jsou spojeny s častým užíváním vlastních jmen historických osobností nebo sportovců a sportovkyň a numerických výrazů, jako jsou letopočty nebo sportovní výsledky.

Na opačné straně škály se nachází obecné subregistry, jako jsou náboženské texty (*REL*), umělecké texty (*FIC*) a deskripce (*DET* a *DES*).

(7) Prospektivní vs. retrospektivní

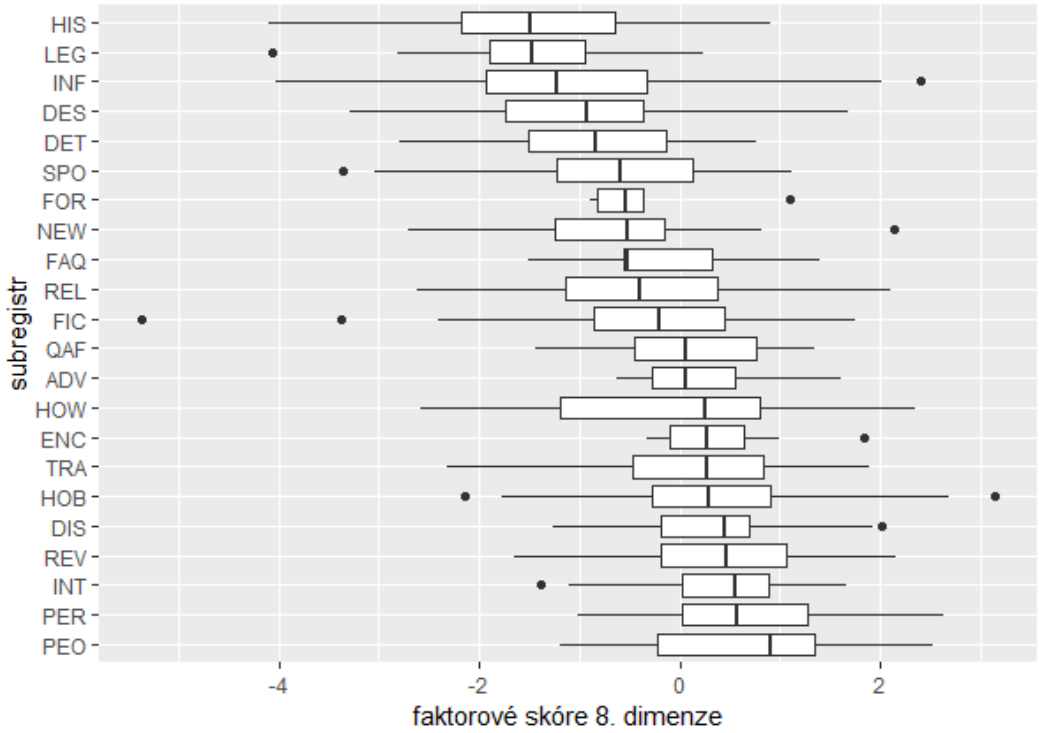


Graf 14: 7. dimenze variability: prospektivní (+) vs. retrospektivní (–) text.

Negativního skóre 7. dimenze dosahují prototypické texty retrospektivního diskurzu, a to historické články (*HIS*). Druhé místo zaujímají umělecké texty (*FIC*), třetí pak texty encyklopedické a výzkumné (*ENC*).

Jako výrazně prospektivní jsou hodnoceny návody typu *jak na to* a jiné instruktážní texty (*HOW*), které zpravidla referují k pracovním postupům prováděným v budoucnosti. Do budoucnosti taktéž odkazuje subregistr poradenských textů (*ADV*) a fórum typu *question–answer* (*QAF*). Pozitivních skóre dosahuje obecné diskusní fórum (*DIS*). Vysoká skóre těchto subregistrů způsobuje pozitivní faktorová zátěž rysů spojených s interakcí mezi účastníky komunikace (imperativ a druhá osoba).

(8) Postojovost vs. faktuálnost



Graf 15: 8. dimenze variability: postojovost (+) vs. faktuálnost (–) textu.

V osmé dimenzi získávají vysoká skóre texty výrazné svou postojovostí, jež je dána především užíváním různých druhů částic.

Negativní skóre mají subregistry faktuální, jako jsou historické články (*HIS*), právní texty (*LEG*) nebo informační texty (*INF*). Mezi faktuální lze řadit také deskriptivní subregistry (*DES* a *DET*).

Mezi subregistry typické vysokou mírou postojovosti patří diskusní fóra (*DIS*), recenze (REV), interview (*INT*), osobní a názorové blogy (*PER* a *PEO*). Nejvyšší hodnota mediánu faktorového skóre byla naměřena právě u názorových blogů (*PEO*), které byly kategorizovány právě na základě charakteristik spojených s postojovostí (nikoli však frekvenční analýzou částic). Relativně vysokého skóre dosahuje také subregistr recenzí (*REV*), jehož faktorová skóre se v rámci ostatních dimenzí pohybovala spíše kolem průměru.

3.4.5 Charakteristika subregistrů

Díky výsledkům multidimenzionální analýzy je možné stručně popsat subregistry pomocí označení vycházející z interpretace pozitivních

a negativních hodnot dimenzí variability. Tabulka č. 4 uvádí pro každý subregistr charakteristiky, které odpovídají výrazným hodnotám faktorového skóre v jednotlivých dimenzích.

Subregistr	**Zkratka**	**Poznámka**	**Charakteristika textu**
Poradenský text	ADV		monotematický, vyšší míra explicitní adresnosti, prospektivní
Popis s účelem prodeje	DES		statický
Popis předmětu	DET	+ popis osoby	statický
Diskusní fórum	DIS		dynamický, spontánní, prospektivní
Encyklopedický článek	ENC	+ vědecký článek	statický, retrospektivní
Nejčastěji kladené dotazy (FAQs)	FAQ		připravený, monotematický
Umělecká próza	FIC	+ písňové texty	dynamický, vyšší míra explicitní adresnosti, obecný, retrospektivní
Formální projev	FOR		nižší míra explicitní adresnosti
Historický článek	HIS		připravený, polytematický, nižší míra explicitní adresnosti, konkrétní, retrospektivní, faktuální
Hobby a volný čas	HOB		—
Návod typu *jak na to*, instruktáž	HOW		prospektivní
Informační text	INF		faktuální
Interview	INT		dynamický, spontánní
Právní text	LEG		statický, připravený, monotematický, faktuální
Zpravodajský článek/blog	NEW		konkrétní
Názorový blog	PEO		vyšší stupeň koheze, postojovost
Osobní blog	PER		postojovost
Fórum typu *question-answer*	QAF		prospektivní
Náboženský blog / kázání	REL	+ ezoterie a parapsych.	vyšší stupeň koheze, obecný
Recenze	REV		postojovost
Sportovní reportáž	SPO		nižší stupeň koheze, nižší míra explicitní adresnosti, konkrétní
Cestovatelský blog	TRA		polytematický

Tabulka 4: Charakteristika internetových subregistrů.

S ohledem na výsledky je zjevné, že některé subregistry by zasloužily jemnější kategorizaci. Týká se to především subregistru informačních textů (*INF*) a zpravidla z něj odštěpených textů subregistru hobby a volný čas (*HOB*). Z krabicových grafů uvedených v předchozím oddílu práce lze vyčíst, že texty těchto subregistrů mají v jednotlivých dimenzích, potažmo v celém multidimenzionálním prostoru velký rozptyl, což signalizuje jejich vnitřní diverzitu. Texty subregistru hobby a volný čas (*HOB*) byly příliš rozptýlené na to, aby je bylo možné výrazněji charakterizovat (viz tabulku 4). Na základě vnějšího pohledu na texty je možné subregistry rozdělit jen stěží. Možným řešením je podrobná analýza distribuce faktorových skóre textů určitého subregistru v jednotlivých dimenzích. Případná bimodální či multimodální distribuce zobrazená v histogramu by svědčila o tom, že uvnitř registru existují dvě či více homogenních skupin.[30] Subregistry hobby a volný čas (*HOB*) a informační texty (*INF*) byly podrobeny tomuto testu. Pro každý subregistr bylo vygenerováno osm histogramů odpovídajících osmi dimenzím. V žádném z nich však nebylo odhaleno jiné než unimodální rozdělení. V případě bimodálního nebo multimodálního rozdělení by mohly být následně porovnány texty z rozdílných skupin a na základě jejich odlišností by mohly být identifikovány podřadné kategorie.

[30] Zobrazení v krabicovém grafu bimodální či multimodální rozdělení neodhalí.

Závěr

Cílem této práce bylo popsat registrovou variabilitu českých internetových textů pomocí multidimenzionální analýzy.

V teoretické části byla představena specifika internetového jazyka na všech relevantních jazykových rovinách a základní přístupy ve výzkumu jazykové variability s důrazem na pojem registr a jeho výzkum prostřednictvím MDA. Tato metoda byla popsána krok za krokem, od sbírání jazykových rysů, přes jejich evaluaci a operacionalizaci, faktorovou analýzu a výslednou interpretaci dimenzí. Představeny byly metody a výsledky dvou pro tuto práci zásadních výzkumů, a to aplikace MDA na texty z webových korpusů (Biber – Egbert, 2016) a aplikace MDA na české texty (Cvrček et al., 2018a, 2018b).

V praktické části byl popsán vzorek dat z webového korpusu Araneum Bohemicum Maximum pomocí modelu MDA převzatého od Cvrčka et al. (tamtéž). Jako podklad pro subregistrovou kategorizaci posloužila klasifikace Bibera a Egberta (2016).

Při anotaci dat se ukázalo, že je klasifikaci nutno revidovat, některé subregistry byly přidány, jinde zas bylo více subregistrů sloučeno do jednoho. Kromě klasifikace do subregistrů byla součástí anotace i kontrola dostupnosti webů na původních adresách obsažených v metainformacích korpusu. Výsledkem této kontroly bylo zjištění, že po sedmi letech od vzniku korpusu je on-line dostupných pouhých 40,5 % textů. V případě dalšího výzkumu registrové variability by bylo žádoucí pracovat s co největším počtem textů, které jsou stále dostupné, čehož by bylo možné dosáhnout minimalizací časové prodlevy mezi získáním dat a analýzou.

Pro další výzkum je nutné znásobit počet anotátorů, jelikož k identifikaci subregistrů dochází značnou měrou na základě intuice. Vyhodnocením mezianotátorské shody by bylo možné dojít k vyšší míře správnosti, resp. adekvátnosti anotace.

Před dalším výzkumem by bylo vhodné věnovat pozornost subregistrům, do kterých spadá výrazně vyšší množství textů než do ostatních, jako je subregistr informačních textů (*INF*). Adekvátnost kategorizace by bylo možné testovat utvořením vzorku textů, jeho anotaci obecnějšími (sub)registry a následnou kontrolou distribuce faktorových skóre v histogramech pro

každou z dimenzí variability. Pomoci by mělo také prohlížení konkrétních textů, obzvlášť pak těch, které nabývají extrémních hodnot (tzv. *outliers*).

Pro faktorovu analýzu byl zvolen výchozí model vytvořený Cvrčkem et al. (2018a, 2018b). Díky tomuto přístupu je možné internetové texty lépe zařadit do celojazykového kontextu a srovnávat faktorová skóre různých textů pomocí téhož modelu. Druhým přístupem, vhodnějším pro zkoumání internetového jazyka jako ohraničené variety (bez nároku na jeho detailní usouvztažnění s ostatními varietami), je provedení faktorové analýzy výhradně na datech z webového korpusu (viz Biber – Egbert, 2016). Výsledné dimenze variability pak mohou lépe odrážet variabilitu typickou právě pro internetový jazyk. Přizpůsobit by bylo možné i seznam jazykových rysů, např. plošnou operacionalizací zkratek, emotikonů, emoji a dalších jazykových prostředků typických pro CMC.

Ve finální části práce byly srovnány subregistry z hlediska distribuce faktorových skóre v jednotlivých dimenzích. Analyzované subregistry často dosahovaly očekávaných skóre (např. retrospektivnost u historických článků (*HIS*) nebo postojovost u názorových blogů (*PEO*)), výsledky tak mohou sloužit i jako důkaz adekvátnosti interpretace dimenzí, kterou navrhují Cvrček et al. (2018a, 2018b). Pro každý subregistr pak byly vybrány charakteristiky určené výraznými hodnotami faktorových skóre v rámci dimenzí variability.

Práce přinesla díky multidimenzionální analýze nový pohled na české internetové texty a na internet jako médium. Tato metoda může svým kvantitativním korpusovým přístupem, užíváním statistických metod a vnitrotextovou orientovaností přinést nový impuls i do české stylistiky.

Literatura

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

BACHMANNOVÁ, J. (2017): Česká nářeční skupina. In: Petr Karlík – Marek Nekula – Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita. Cit. 28. 6. 2019.

BENKO, V. (2014): Aranea: Yet Another Family of (Comparable) Web Corpora. In: Petr Sojka – Aleš Horák – Ivan Kopeček – Karel Pala (eds.), *TSD 2014.* New York: Springer International Publishing, s. 257–264.

BIBER, D. (1988): *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

BIBER, D. (1990): Methodological Issues Regarding Corpus-Based Analyses of Linguistic Variation. *Literary and Linguistic Computing*, 5(4). 257–269.

BIBER, D. (1995): *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

BIBER, D. (2006): *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

BIBER, D. – CONRAD, S. (2009): *Register, Genre, and Style*. New York: Cambridge University Press.

BIBER, D. – EGBERT, J. (2016): Register Variation on the Searchable Web: A Multi-Dimensional Analysis. *Journal of English Linguistics*, 44(2). 95–137.

COLLOT, M. –BELMORE, N. (1996): Electronic Language: A New Variety of English. In: Susan Herring (ed.): *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives*. Amsterdam: John Benjamins, s. 13–28.

CONRAD, S. M. (1996): Investigating Academic Texts with Corpus-Based Techniques: An Example from Biology. *Linguistics and Education*, 8(3), s. 299–326.

CRYSTAL, D. (2006): *Language and the Internet*. Cambridge: Cambridge University Press.

CRYSTAL, D. (2011). *Internet Linguistics: A Student Guide*. New York: Routledge.

CRYSTAL, D. – DAVY, D. (1973): *Investigating English Style*. New York: Routledge.

CVRČEK, V. (2017): Variabilita. In: Petr Karlík – Marek Nekula – Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita. Cit. 16. 6. 2019.

CVRČEK V. – CHLUMSKÁ L. (2015): Simplification in Translated Czech: A New Approach to Type-Token Ratio. *Russian Linguistics*, 39(3), s. 309–325.

CVRČEK, V. – KOMRSKOVÁ, Z. – LUKEŠ, D. – POUKAROVÁ, P. – ŘEHOŘKOVÁ, A. – ZASINA, A. J. (2018a): From Extra- to Intratextual Characteristics: Charting the Space of Variation in Czech through MDA. *Corpus Linguistics and Linguistic Theory*.

CVRČEK, V. – KOMRSKOVÁ, Z. – LUKEŠ, D. – POUKAROVÁ, P. – ŘEHOŘKOVÁ, A. – ZASINA, A. J. (2018b): Variabilita češtiny: multidimenzionální analýza. *Slovo a slovesnost*, 79(3), s. 293–321.

Český jazykový atlas 2 (1997). Praha: Academia.

Čeština na internetu (2006). *Čeština doma a ve světě*, 14(1–4). Praha: FF UK.

ČMEJRKOVÁ, S. (1997): Čeština v síti: Psanost či mluvenost? (O stylu e-mailového dialogu). *Naše řeč*, 80(4), s. 225–247.

ČMEJRKOVÁ, S. (2006): E-čeština. *Čeština doma a ve světě*, 14(1–4), s. 4–15.

DESAGULIER, G. (2017): *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. New York: Springer International Publishing.

DEWE, J. – KARLGREN, J. – BRETAN, I. (1998): Assembling a Balanced Corpus from the Internet. In: Bente Maegaard (ed.), *Proceedings of the 11th Nordic Conference of Computational Linguistics (NODALIDA 1998)*. Copenhagen: University of Copenhagen, s. 100–108.

DRESNER, E. – HERRING, S. C. (2010): Functions of the Nonverbal in CMC: Emoticons and Illocutionary Force. *Communication Theory*, 20(3), s. 249–268.

FERGUSON, C. A. (1977): Baby Talk as a Simplified Register. In: Charles A. Ferguson – Catherine Snow (eds.), *Talking to Children: Language Input and Acquisition*, s. 209–235.

GILES, H. – TAYLOR, D. (1973): Towards a Theory of Interpersonal Accommodation through Language: Some Canadian Data. *Language in Society*, 2(2), s. 177–192.

GREENE, L. (2010): *The Internet: An Introduction to New Media*. New York: Berg publishers.

GRIEVE, J. – BIBER D. – FRIGINAL, E. – T. NEKRASOVA (2011): Variation Among Blogs: A Multi-Dimensional Analysis. In: Alexander Mehler – Serge Sharoff – Marina Santini (eds.), *Genres on the Web: Computational Models and Empirical Studies*, TLTB, 42, s. 303–322.

HALLIDAY, M. A. K. (1978): *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. London: Edward Arnold.

HALLIDAY, M. A. K. (1989): *Spoken and Written Language*. Oxford: Oxford University Press.

HALLIDAY, M. A. K. – HASAN, R. (1994): *Cohesion in English*. London: Longman.

HARMAN, R. (2017): Faktorová analýza. In: Petr Karlík – Marek Nekula – Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita. Cit. 8. 7. 2019.

HAŠOVÁ, L. (2002): Lásky jedné esemesky. *Naše řeč*, 85(4), s. 207–212.

HERRING, S. C. (ed.) (1996): *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. Philadelphia: John Benjamins

HERRING, S. C. (2007): A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@internet*, 4(1), s. 1-37.

HERRING, S. C. (2013): Grammar and Electronic Communication. In: Carol Chapelle (ed.), *The Encyclopedia of Applied Linguistics*. Chichester: Wiley-Blackwell.

HILTZ, S. R. – TURROF, M. (1978): *The Network Nation: Human Communication via Computer*. Boston: Addison-Wesley Publishing Company.

HOFFMANNOVÁ, J. – HOMOLÁČ, J. – CHVALOVSKÁ, E. – JÍLKOVÁ, L. – KADERKA, P. – MAREŠ, P. – MRÁZKOVÁ, K. (2016): *Stylistika mluvené a psané češtiny*. Praha: Academia.

CHROMÝ, J. (2012): Koncepce stylotvorných faktorů v československé lingvistice: rozbor, kritika, nástin řešení. *Naše řeč*, 95(2), s. 57-69.

IDE, N. – REPPEN, R. – SUDERMAN, K. (2002): The American National Corpus: More Than the Web Can Provide. In: *Proceedings of the Third Language Resources and Evaluation Conference (LREC)*. Las Palmas, Canary Islands, Spain: Citeseer, s. 839–844.

JELÍNEK, M. – KRČMOVÁ, M. (2017): Stylistika. In: Petr Karlík – Marek Nekula – Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita. Cit. 24. 6. 2019.

JÍLKOVÁ, L. (2017): Elektronická komunikace. In: Petr Karlík – Marek Nekula – Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita. Cit. 23. 6. 2019.

KEMPER, S. – FINTER-URCZYK, A. – FERRELL, P. – HARDEN, T. – BILLINGTON, C. (1998): Using Elderspeak with Older Adults. *Discourse Processes*, 25(1), s. 55–73.

KRČMOVÁ, M. (2017): Slohový postup. In: Petr Karlík – Marek Nekula – Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita. Cit. 26. 6. 2019.

KRESS, G. (2009): *Multimodality: A Social Semiotic Approach to Contemporary Communication*. Abingdon: Routledge.

LIŠKOVÁ, M. (2006): Některé komunikační a mluvnické aspekty počítačového slangu. In: Petr Pořízka – Vladimír P. Polách (eds.), *Tzv. základní výzkum v lingvistice – desideratum, nebo realis?*, s. 46–52.

MRÁZKOVÁ, K. (2014): Pojem „registr" v soustavě pojmů sociolingvistiky a funkční stylistiky. In: Jana Kesselová (ed.), *Registre jazyka a jazykovedy*. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove.

NEKVAPIL, J. (2017a): *Sociolingvistika*. In: Petr Karlík – Marek Nekula – Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita. Cit. 21. 6. 2019.

NEKVAPIL, J. (2017b): *Varieta jazyka*. In: Petr Karlík – Marek Nekula – Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Masarykova univerzita. Cit. 21. 6. 2019.

PETRÁČKOVÁ, V. – KRAUS, J. (1998): *Akademický slovník cizích slov [A-Ž]*. Praha: Academia.

POSTEGUILLO, S. (2003): *Netlinguistics: An Analytical Framework to Study Language, Discourse and Ideology in Internet*. Castellón de la Plana: Jaume I University.

PROŠEK, M. (2013): Objektivní stylotvorné faktory. In: Oldřich Uličný – Soňa Schneiderová (eds.), *Studie k moderní mluvnici češtiny: 2, Komunikační situace a styl*. Olomouc: Univerzita Palackého v Olomouci.

RENCHER, A. C. – CHRISTENSEN, W. F. (2002): *Methods of Multivariate Analysis*. New York: Wiley.

SANTINI, M. (2007): Characterizing Genres of Web Pages: Genre Hybridism and Individualization. In: *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS).* Waikoloa: IEEE.

SHEPHERD, M. – WATTERS, C. (1998): The Evolution of Cybergenres. In: *Proceedings of the 31st Hawaii International Conference on System Sciences (HICSS)*. Wailea: IEEE.

SCHNEIDEROVÁ, S. (2013): Pojem funkce a klasifikace funkčních stylů v české stylistice. In: Oldřich Uličný – Soňa Schneiderová (eds.), *Studie k moderní mluvnici češtiny: 2, Komunikační situace a styl*. Olomouc: Univerzita Palackého v Olomouci.

SMEJKALOVÁ, K. (2013): Subjektivní stylotvorné faktory. In: Oldřich Uličný – Soňa Schneiderová (eds.), *Studie k moderní mluvnici češtiny: 2, Komunikační situace a styl*. Olomouc: Univerzita Palackého v Olomouci.

SQUIRES, L. (2010): Enregistering Internet Language. *Language in Society*, 39(4), s. 457–492.

ŠKALOUDOVÁ, A. (2010): *Faktorová analýza* [online]. Cit. 8. 7. 2019. <http://kps.pedf.cuni.cz/skalouda/fa>.

TRASK, R. L. (1999): *Key Concepts in Language and Linguistics*. London: Routledge.

VAN LEEUWEN, T. (2014): *Multimodality and Multimodal Research*. In: Eric Margolis – Luc Pauwels (eds.), *The Handbook of Visual Research Methods*. London: Sage Publications, s. 549–569.

Korpusy a aplikace:

BENKO, V. (2015): *Srovnatelné webové korpusy Aranea* [online]. Praha: Ústav Českého národního korpusu FF UK. <http://www.korpus.cz>

LUKEŠ, D. (2018): *MDAvis* [online]. Praha: Ústav Českého národního korpusu FF UK. <https://jupyter.korpus.cz/shiny/lukes/mda>

MACHÁLEK, T. (2014): *KonText – aplikace pro práci s jazykovými korpusy* [online]. Praha: Ústav Českého národního korpusu FF UK. <http://kontext.korpus.cz>

R DEVELOPMENT CORE TEAM (2018). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

RSTUDIO TEAM (2015): *RStudio: Integrated Development for R*. Boston: RStudio.

WICKHAM, H. (2016): *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer International Publishing.

WICKHAM, H. – FRANÇOIS R. – HENRY, L. – MÜLLER, K. (2018): *dplyr: A Grammar of Data Manipulation*. R package, version 0.7.6.

ZASINA, A. J. – LUKEŠ, D. – KOMRSKOVÁ, Z. – POUKAROVÁ, P. – ŘEHOŘKOVÁ, A. (2018): *Koditex: korpus diverzifikovaných textů* [online]. Praha: Ústav Českého národního korpusu FF UK. <http://www.korpus.cz>.

Seznam příloh

Přílohy

Příloha 1: Vnětextová kategorizace korpusu Koditex

MODE	DIVISION	SUPERCLASS	CLASS	Tokeny	Vzorky
spo (mluvená komunikace)	int (interaktivní)		bru (nepřipravené veřejné / vysílané rozhovory)	221 812	90
			eli (formální rozhovor)	201 690	82
			inf (neformální rozhovor)	208 565	86
	nin (neinteraktivní)		wbs (připravený/čtený projev)	213 201	71
web (internetová komunikace)	mul (mnohosměrná komunikace)		dis (internetové diskuse)	197 948	87
			fcb (facebookové statusy)	199 418	91
			for (webová fóra)	200 104	85
	uni (jednosměrná komunikace)		blo (blogy)	204 356	74
			wik (wikipedie)	201 691	84
wri (psaná komunikace)	fic (beletrie)	nov (romány)	crm (detektivky)	190 026	68
			fan (fantasy)	189 432	69
			gen (bez bližšího určení)	193 667	67
			lov (milostné)	189 893	70
			scf (sci-fi)	188 703	68
			col (povídky)	195 595	70
			scr (scénáře a dramata)	182 689	76
			ver (poezie a písně)	205 837	76
	nfc (oborová literatura)	pop (populárně-naučná)	fts (formální a technické vědy)	207 607	68
			hum (humanitní vědy)	204 837	74
			nat (přírodní vědy)	204 751	71
			ssc (společenské vědy)	203 698	68
		pro (profesní literatura)	fts (formální a technické vědy)	210 010	71
			hum (humanitní vědy)	207 916	69
			nat (přírodní vědy)	209 580	70
			ssc (společenské vědy)	209 385	72
		sci (vědecká literatura)	fts (formální a technické vědy)	202 932	67
			hum (humanitní vědy)	204 300	71
			nat (přírodní vědy)	206 716	72

			ssc (společenské vědy)	205 358	67
			adm (administrativa)	203 542	82
			enc (encyklopedie)	203 957	73
			mem (auto-/biografie)	203 390	71
	nmg (noviny a časopisy)	lei (volnočasová publicistika)	hou (bydlení, zahrada, hobby)	207 499	68
			int (zajímavosti ze světa)	209 232	69
			lif (životní styl)	203 124	72
			mix (víkendové přílohy)	205 310	75
			sct (bulvár)	201 417	73
			spo (sport)	199 238	70
		new (tradiční publicistika)	com (komentáře)	205 372	68
			cul (kultura)	205 690	68
			eco (ekonomika)	211 481	70
			fre (volnočasové aktivity)	208 532	71
			pol (politika)	206 893	70
			rep (reportáže)	206 377	70
	pri (soukromá komunikace)		cor (dopisy)	96 366	68
Celkem				**9 039 137**	**3292**

Příloha 2: Seznam jazykových rysů užitý v MDA na českém materiálu

ID	Rys	Rovina	Poznámka/příklad
FEI	úžení *é* > *í* v kořeni (typ *polívka*)	fonologie	
FEIK	úžení *é* > *í* v koncovce (typ *starý město*)	fonologie	
FYEJ	přehláska *ý* > *ej* v kořeni (typ *bejt*)	fonologie	
FYEJK	přehláska *ý* > *ej* v koncovce (typ *mladej*)	fonologie	
WL	průměrná délka slova v textu (v počtu slabik)	fonologie	pouze tokeny obsahující písmena
AA	adjektiva	lexikon	všechny typy
PAJ	sémanticky vyprázdněná adjektiva	lexikon	víc než 3 významy v SSČ
DEMD	ukazovací adverbia	lexikon	
DB	adverbia	lexikon	všechny typy
DT	adverbia vyjadřující určení času	lexikon	
DP	adverbia vyjadřující určení místa či směru	lexikon	
DI	neurčitá adverbia	lexikon	
RST	restriktory (např. *hlavně, alespoň, maximálně*)	lexikon	
PAV	sémanticky vyprázdněná adverbia	lexikon	víc než 3 významy v SSČ
INT	citoslovce	lexikon	
FNW	synsémantika (zájmena, číslovky, předložky, spojky a částice)	lexikon	vč. *být* v roli pomocného slovesa v min. čase
POE	poetismy (podle příznaku v SSČ)	lexikon	bez výrazně homonymních jednotek
TIME	časové výrazy (adverbia, jména a časové entity)	lexikon	včera, pondělí, 30. 4. 2010
CONJI	inventář spojek	lexikon	normalizováno pomocí zTTR
MOD	modální slovesa, adjektiva, adverbia a částice	lexikon	
NUM	numerale	lexikon	
PREI	inventář předložek	lexikon	normalizováno pomocí zTTR
DEM	ukazovací zájmena (bez slovního tvaru "to")	lexikon	
IND	neurčitá zájmena	lexikon	
P1P	zájmena pro 1. os. (osobní i přivlastňovací)	lexikon	vč. mluvených variant
P2P	zájmena pro 2. os. (osobní i přivlastňovací)	lexikon	vč. mluvených variant
P3P	zájmena pro 3. os. (osobní i přivlastňovací)	lexikon	vč. mluvených variant
PP	pronomina	lexikon	všechny typy
PPI	inventář zájmen	lexikon	normalizováno pomocí zTTR
POSP	přivlastňovací zájmena	lexikon	viz i P1P, P2P a P3P
NN	substantiva	lexikon	všechny typy
PSB	sémanticky vyprázdněná substantiva	lexikon	víc než 3 významy v SSČ
VTS	slovesa myšlení	lexikon	
VD	slovesa mluvení	lexikon	
VB	slovesa	lexikon	všechny typy
PVB	sémanticky vyprázdněná slovesa	lexikon	víc než 3 významy v SSČ

ID	Rys	Rovina	Poznámka/příklad
ARCH	knižní a zastaralá synsémantika a vybraná adv.	lexikon	např. *rovněž, dle, též, aniž*; na základě SSJČ
GENL	lexikální genderové markery	lexikon	propria a apelativa označující ženy
ADN	jmenné tvary adjektiv	morfologie	
DAT	dativ (substantiv)	morfologie	
GEN	genitiv (substantiv)	morfologie	
INS	instrumentál (substantiv)	morfologie	
LOK	lokál (substantiv)	morfologie	
NAKK	nominativ–akuzativ (substantiv)	morfologie	
CAS	pády realizované s předložkou	morfologie	bez předložkově obligatorního lokálu, tj. gen., dat., ak. a inst.
IEK	morfologická konkurence koncovek *-é* vs. *-i*	morfologie	typ *policisté*
VOC	vokativ (substantiv)	morfologie	
SGPL	množné číslo substantiv	morfologie	
VPE1	1. slovesná osoba	morfologie	
VPE2	2. slovesná osoba	morfologie	
ICI	infinitiv na *-ci*	morfologie	
VIM	imperativ	morfologie	
VIN	indikativ	morfologie	
VTE3	slovesný čas – budoucí	morfologie	
VTE1	slovesný čas – minulý	morfologie	
VVO	slovesný rod – pasivum	morfologie	
VTE2	slovesný čas – přítomný	morfologie	
NEG1	negace větná	morfologie	slovesa s prefixem *ne-*
VF	verba finita	morfologie	
BYTS	pomocné "být" ve formě *-s* (typ *přišels*)	morfologie	
VAS	slovesný vid – perfektivum	morfologie	
VCO	kondicionál	morfologie	verba
NEG2	negace lexikální	morfologie	substantiva s prefixem *ne-*
NEG4	negace členská (*ne*, *nikoli* + varianty)	morfologie	
DROP	pronoun non-dropping	morfologie	explicitní přítomnost osobního zájmena v 1. a 2. osobě
GENG	morfologické genderové markery	morfologie	slovesná participia s ženským rodem
KONT	kontaktové výrazy	pragmatika	např. *hele, aha*
VYPW	výplňková slova	pragmatika	např. *hm, vlastně*
EXP	expresivní částice vyjadřující hodnocení, postoj, modalitu a emoce	pragmatika	např. *každopádně, vždyť*
AMP	částice zesilující význam (amplifiers/boosters)	pragmatika	např. *dokonce, vyloženě*
DOWN	částice oslabující význam (downtoners/hedges)	pragmatika	např. *poněkud, bezmála*
IT	polyfunkční "to"	pragmatika	všechny funkce a významy
PROPT	toponyma	pragmatika	

ID	Rys	Rovina	Poznámka/příklad
PROPA	antroponyma	pragmatika	
EXC	vykřičník	pragmatika	
NGR	výskyt frekventovaných 5-gramů	pragmatika	frekvenční špička ze SYN2015 a ORAL v4
GRAAA	analytické stupňování (typ *méně pěkný*)	slovotvorba	
ASIM	adjektiva podobnosti (podle přípon)	slovotvorba	adj. na *-ovský, -oidní, -ovitý*
POS	přivlastňovací adjektiva (typ *otcův* i *matčin*)	slovotvorba	
AAV	verbální adjektiva (typ *tekoucí* i *přišedší*)	slovotvorba	s vyloučením lexikalizovaných
GRAD	stupňování adjektiv (komparativ + superlativ)	slovotvorba	
GRADD	stupňování adverbií (komparativ + superlativ)	slovotvorba	
DEMI	deminutiva	slovotvorba	
NNV	verbální substantiva	slovotvorba	
ABS	abstrakta	slovotvorba	
ATA1	přívlastky shodné anteponované	syntax	
ATA21	přívlastky shodné postponované nerozvité	syntax	
ATA22	adjektivní přívlastky shodné postponované rozvité	syntax	
CLUAC	klastry pádů adjektiv	syntax	sekvence adj. se stejným pádem
CLUA	klastry adjektiv	syntax	
CLUAD	klastry adverbií	syntax	
CPD	věty s interog. a vztaž. adv.	syntax	např. *kde, kdy, jak*
CONJ	konjunkce	syntax	všechny typy
COH1	vícečlenné konektory (typ *jednak – jednak*)	syntax	
COOR	koordinace	syntax	*a, i, nebo, či, ani*
PRE	předložky	syntax	všechny typy
PRE2	sekundární předložky	syntax	
REL2	vztažné věty typu *který*	syntax	
REL3	vztažné věty typu *co*	syntax	
REL1	vztažné věty typu *jenž*	syntax	
SER	větná relativa	syntax	*což, přičemž*
ACM	komplementace adjektiva	syntax	např. *je jasné, že...*
VCM	komplementace slovesa	syntax	např. *myslím, že...*
CLUN	klastry substantiv	syntax	
CLUNC	klustry pádů substantiv	syntax	sekvence subst. se stejným pádem
STA4	subst. přívlastky neshodné postponované	syntax	
PA	jmenný přísudek adjektivní	syntax	
PN	jmenný přísudek substantivní	syntax	
CIR3	konektory přípustkové a další adverzativní	syntax	*aniž, ačkoli, přesto, třebaže*
CIR124	konektory okolnostní	syntax	podmínkové, příčinné a další okolnostní
SL	průměrná délka věty (v počtu slov)	syntax	

ID	Rys	Rovina	Poznámka/příklad
NEG3	negace vícenásobná v rámci jedné věty	syntax	
CORR	korelativa (deiktikum + konektor)	syntax	*...tak, že… / ...tím, aby...*
COH2	částice členící text	text	*dále, například*
QUE2	otázky (zjišťovací i doplňovací)	text	
QUE1	otázky doplňovací	text	
PHRA	četnost frekventovaných frazémů	text	na základě automatické frazémové anotace
REP	opakování identických slov	text	
LR	Yulův koeficient lexikální repetitivnosti	text	
TC	tematická koncentrace textu	text	
BIG	bigramy	text	zTTR dvojic slov (tvary)
UNG	unigramy	text	zTTR slov (lemmata)

Příloha 3: Distribuce dat jednotlivých subregistrů v první a druhé dimenzi