Univerzita Karlova v Praze
Filozofická fakulta
Ústav Českého národního korpusu

DISERTAČNÍ PRÁCE

Překladová čeština a její charakteristiky

Translated Czech and Its Characteristics

Mgr. Lucie Chlumská

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

Studijní program: Filologie

Studijní obor: Matematická lingvistika

Praha 2015

Prohlášení

Prohlašuji, že jsem disertační práci napsala samostatně s využitím pouze uvedených a řádně citovaných pramenů a literatury a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 19. srpna 2015 Mgr. Lucie Chlumská

Poděkování

Tato disertační práce by nemohla vzniknout bez podpory mnoha lidí a já bych zde ráda poděkovala alespoň několika z nich.

Největší dík patří mému školiteli Václavu Cvrčkovi za vstřícné vedení, nesčetné konzultace i konstruktivní kritiku. Kolegovi Michalu Křenovi vděčím za technickou pomoc při vytváření korpusu Jerome, Jiřímu Václavíkovi a Petru Trunečkovi za technickou přípravu dat k analýze. Bez pomoci Davida Lukeše by mi programování a vytváření grafů v R trvalo jistě mnohem déle. Kolegové Pavel Vondřička a Martin Vavřín mi svými radami mnohokráte usnadnili psaní v LaTeXu. Dominice Kováříkové děkuji za její neutuchající morální podporu a koučce Silvii Nedvědové za to, že mi byla téměř dva roky laskavým průvodcem při psaní této práce.

Velmi ráda bych zde také vyjádřila upřímný dík své rodině, neboť mě po celou dobu studií vytrvale podporovala. A v neposlední řadě děkuji celému Ústavu Českého národního korpusu, v němž je radost studovat i pracovat.

Abstrakt

Název práce: Překladová čeština a její charakteristiky
Autor: Mgr. Lucie Chlumská
Katedra: Ústav Českého národního korpusu
Školitel: doc. Mgr. Václav Cvrček, Ph.D.

Abstrakt:

Ačkoli překladová literatura tvoří v českém prostředí více než třetinu knižní produkce, čeština v překladech doposud nebyla podrobena systematické kvantitativní analýze. Tato korpusovětranslatologická práce si proto vzala za cíl popsat charakteristické rysy překladové češtiny ve srovnání s češtinou v původně českých, nepřekladových textech. Analýza byla provedena na základě rozsáhlého jednojazyčného srovnatelného a synchronního korpusu Jerome, který byl sestaven přímo za tímto účelem. Zahrnuje beletrii i odbornou a populárně-naučnou literaturu a svým složením z hlediska zdrojových jazyků odráží reálnou situaci v ČR (převažují překlady z angličtiny). Inspiračním zdrojem práce se stal koncept tzv. překladových univerzálií, tj. údajných typických jazykových rysů, jež by měly být překladovým textům společné. Pozornost byla věnována především simplifikaci, konvergenci a obecným frekvenčním charakteristikám, vč. porovnání výskytu slovních druhů a typických n-gramů.

Výsledky zkoumání potvrdily, že překladová čeština se ve své podobě, jak se s ní setkává průměrný čtenář, v určitých aspektech skutečně odlišuje od češtiny nepřekladové: dochází v ní k simplifikaci i konvergenci a liší se i na úrovni konkrétních lexikálních vzorců. Neméně důležitým zjištěním však je, že rozdíly mezi překlady a nepřeklady nejsou příliš výrazné, což je patrné především v porovnání s výrazněji odlišnými výsledky u obou zkoumaných textových typů (beletrie a odborné literatury), jež jsou na základě zvolených testů mnohem lépe identifikovatelné.

Na základě provedených analýz tak lze konstatovat, že překladová čeština, jak je zachycena v použitém korpusu, vykazuje jisté specifické rysy; vzhledem k dostupnosti dat a složení korpusu však nelze s jistotou hovořit o univerzálních vlastnostech překladového jazyka. Výzkum tak rovněž poukázal na přednosti i nedostatky specifických korpusů používaných v translatologii a odhalil další možnosti výzkumu v oblasti překladové češtiny.

Klíčová slova: překladový jazyk, čeština, korpusová translatologie, překladové univerzálie, simplifikace, konvergence

Abstract

Title: Translated Czech and Its Characteristics
Author: Mgr. Lucie Chlumská
Department: Institute of the Czech National Corpus
Supervisor: doc. Mgr. Václav Cvrček, Ph.D.

Abstract: Despite the fact that translated literature accounts for more than one third of all written publications in the Czech Republic, Czech in translations has not yet been systematically analyzed from a quantitative point of view. The main objective of this corpus-based dissertation is to identify characteristic features of translated Czech compared to Czech in original, i.e. non-translated texts. The analysis was based on a large monolingual comparable corpus Jerome, created for the purposes of this study. It includes both fiction and non-fiction texts and its design reflects the real Czech situation regarding the translations' source languages, i.e. translations from English prevail. The research was inspired by the theory of translation universals (typical linguistic features common to any translated text) and focused mainly on simplification, convergence and general frequency characteristics, including parts-of-speech distribution and n-gram analysis.

The findings have supported the hypothesis that translated Czech, as reflected in the Jerome corpus, is different from the non-translated Czech in terms of higher degree of simplification, convergence and distinct lexical patterning. The differences, however, are not so striking, as we may have expected, especially when compared to the distinction between fiction and non-fiction, which proved to be more prominent and recognizable based on the carried-out tests.

To conclude, translated Czech certainly has some specific features resulting probably both from an interference effect and the very process of translation. Due to the data limitations and the Jerome corpus design it is impossible to claim these features to be universal; they only apply to the translated Czech as a common reader encounters it, i.e. dominated by the translations from English. The analysis has also highlighted the advantages and disadvantages of specific types of corpora used in translation studies, and suggested the avenues for further research.

Keywords: language of translation, Czech, corpus-based translation studies, translation universals, simplification, convergence

Obsah

Kapitola 1

Úvod

1.1 Překladová literatura u nás

Překlady z cizích jazyků tvoří snad ve všech kulturách s psanou tradicí určitou, méně či více významnou část vydávané literatury, o to větší pak v malých zemích, kde se domácí produkce nemůže objemem rovnat produkci zahraniční. To je i případ českého prostředí. Ačkoli se jazyky a kultury, z nichž se nejvíce překládalo, v dějinách naší země proměňovaly, překladová literatura měla v naší kultuře vždy nezastupitelné místo. V různých obdobích plnil překlad různé funkce, za všechny zmiňme například specifickou dobu národního obrození, kdy překladová literatura suplovala nedostatek domácí literární produkce (např. rytířskou povídku nebo prózu ze šlechtického prostředí) a dokládala, že čeština je schopna vyjádřit to, co velké světové jazyky (Hrala 2002: 7–8). Dnes už čeština nic dokazovat nemusí, překlady většinou vznikají z prozaičtějších důvodů – vedle přinášení významných cizojazyčných děl do českého prostředí se řídí především poptávkou. S rozvojem domácí kultury se překlad postupně emancipoval a stal se její relativně samostatnou součástí. Tím se otevřely možnosti zkoumání překladu jako takového – od jeho dějin, tradice či norem, až po výzkum hypotézy zastarávání překladu či samotného jazyka překladu.

Úměrně se zvyšováním domácí literární produkce se dařilo a daří i překladové literatuře. Podíváme-li se na situaci v nedávných pěti letech (od roku 2008 do roku 2012), počet překladů neperiodických publikací – kam patří beletrie, populární i odborná literatura – stále narůstá. V roce 2012 bylo vydáno téměř o 98 % překladových knih více než v roce 2008. V roce 2012 to bylo celkem 5 871 titulů překladové literatury, což je více než 34 % z celkového objemu knižní produkce. Přeloženy byly knihy ze 44 jazyků. V roce 2008 pak bylo vydáno 5 546 titulů (téměř 30 % knižní produkce). Nejvíce překládanými jazyky jsou po mnoho let angličtina, ze které bylo v roce 2012 přeloženo 3 238 titulů, dále pak němčina (970 titulů) a francouzština (239 tituly). Ze slovanských jazyků se na prvním místě dlouhodobě drží slovenština se 128 přeloženými tituly, na dalších místech je polština s 86 tituly) a ruština se 78

tituly (viz statistiky NKP[1]). Podrobnější statistiky překladové literatury a to, jakým způsobem se odráží ve složení zkoumaného korpusu, lze nalézt v kapitole 3.2.1.

Je všeobecně přijímaným faktem, že recepce textů má vliv na percepci i na produkci jazyka příjemců. Uvážíme-li tedy, že více než třetinu produkce dnes tvoří překladová literatura, musí nás nutně zajímat, zda se překladový jazyk nějak neodlišuje, zdali není svébytným kódem, který má své vlastní zákonitosti a svá pravidla. Je překladová čeština jiná než čeština původních, česky psaných děl? Vykazuje jazyk překladů nějaké specifické rysy, jež jsou pro něj typické bez ohledu na jazyk, z něhož byl překlad pořízen? Je možné tyto rysy odhalit a popsat pomocí korpusových metod? Nejen na tyto otázky se pokusí odpovědět tato práce.

1.2 Téma práce

Tématem této disertace je tedy překladový jazyk, angl. *language of translation*[2], v tomto případě jazyk textů přeložených z cizího jazyka do češtiny. Cílem výzkumu je s pomocí kvantitativních metod zjistit, zda se tento jazyk nějak odlišuje od jazyka nepřekladového, a pokud ano, formulovat a popsat jeho charakteristické rysy. Za nepřekladový jazyk jsou zde považovány takové texty, které byly původně napsány i publikovány v českém jazyce pro české publikum. Mezi zkoumané texty je zařazena jak beletrie, tak i odborná literatura. Je třeba předem zdůraznit, že cílem práce není hodnotit kvalitu překladu textů, ani by to vzhledem k objemu a povaze zkoumaných dat nebylo možné (viz kapitola 3.2.2). Hlavním předmětem zájmu jsou jakékoli rysy typické pro překladový jazyk, jejich interpretace a kategorizace s přihlédnutím k dosavadnímu výzkumu v této oblasti (viz kapitola 2).

Kvantitativní pohled je v této oblasti české translatologie výjimkou (např. Kubáčková 2008), většina studií a vědeckých prací, které se zabývají překladovým jazykem a překladovými univerzáliemi, se omezuje na kvalitativní srovnání jedné nebo více verzí konkrétního překladu s původním dílem (např. Polišenská 2010, Kočová 2009, Středová 2009). Tato práce naopak využívá rozsáhlý textový korpus překladového a nepřekladového jazyka v řádu desítek milionů slov, který byl pro tyto účely sestaven na Ústavu Českého národního korpusu FF UK (viz kapitola 3.2). Výhodou tohoto přístupu je tak jeho interdisciplinarita, která plyne ze spojení metod korpusové lingvistiky s poznatky moderní translatologie.

[1]Dostupné ve formátu PDF zde: `http://www.nipos-mk.cz/wp-content/uploads/2013/05/Statistika_kultury_2012_III.KNIHOVNY_web.pdf` nebo zde: `http://text.nkp.cz/soubory/ostatni/vykaz_dd2012.pdf`.

[2]Doslovným ekvivalentem v češtině je „jazyk překladu“, který se v tomto smyslu také používá. Termín„překladový jazyk“ se však jeví jako vhodnější, především kvůli analogii s „překladovou literaturou“ a kvůli možnosti utvořit antonymum „nepřekladový jazyk“, tedy jazyk nepřeložených textů (v kontrastu k původním textům či originálům, které zpravidla označují zdrojové texty překladu).

1.3 Struktura práce

Tato práce je rozčleněna do čtyř hlavních kapitol. Po krátkém úvodu následuje kapitola 2, která shrnuje dosavadní výzkum v oblasti moderní translatologie se zaměřením na deskriptivní translatologii, která se odklonila od tradičního preskriptivního proudu a iniciovala obrat ve vnímání překladu. Zvláštní důraz je kladen na korpusový výzkum v translatologii, který se v posledních letech zaobírá především hledáním tzv. překladových univerzálií neboli charakteristických rysů překladového jazyka.

Třetí kapitola s názvem Data a metodologie nejprve stručně představuje druhy korpusů používané v korpusové translatologii a vymezuje použitou terminologii. Jejím těžištěm však jsou informace o korpusu Jerome, především pak o jeho složení, počtu textových slov (tokenů) a zastoupení textových typů a žánrů. Rovněž jsou zde vysvětlena kritéria pro výběr textů – ať už z hlediska časové roviny, autora[3] nebo překladatele. Část věnovaná metodologii představuje hlavní důvody pro využití kvantitativní analýzy a předkládá formulaci základní výzkumné hypotézy.

Kapitola 4 nazvaná Rysy překladové češtiny tvoří jádro práce. Je rozdělena do několika podkapitol, které jsou věnovány tradičně vymezovaným překladovým univerzáliím a dalším jazykovým jevům, které jsou zde testovány. Každá podkapitola zahrnuje seznam dílčích hypotéz, popis statistických metod využitých pro jejich zkoumání a vyhodnocení výsledků, vč. jejich přehledného zpracování do grafů a tabulek.

Závěrečná kapitola 5 pak shrnuje výsledky celé práce a poukazuje na možnosti výzkumu v oblasti překladového jazyka a překladových univerzálií.

Práci doplňuje kompletní seznam literatury a seznamy vyobrazených tabulek a obrázků.

Formální zásady

V práci je uplatňována citační norma APA 6 (používaná např. v prestižním oborovém časopise *International Journal of Corpus Linguistics*). Citáty z anglicky psané zahraniční literatury jsou v celé práci ponechány v originále s přesným odkazem na příslušný text. Jména autorek jsou v textu v souladu s českým územ důsledně přechylována, avšak v odkazech a bibliografii je pro snazší dohledání ponecháno jméno vždy v původní nepřechýlené podobě.

[3]Není-li uvedeno jinak, jsou mužské tvary substantiv, např. autor či překladatel, v celé práci užívány genericky, odkazují tedy i k ženským zástupkyním, např. autorkám a překladatelkám. V případě, že byl genderový aspekt zohledněn při analýze, je tato informace vždy výslovně uvedena.

Kapitola 2

Přehled dosavadního výzkumu

Cílem následujícího přehledu není vyčerpávajícím způsobem popsat hlavní teorie překladu – k tomu slouží kromě zásadních prací konkrétních translatologů i bezpočet odborných publikací a popularizačních příruček (např. Venuti 2000, Munday 2008) –, nýbrž pro účely této práce velmi stručně představit vývoj translatologie, který na počátku jedenadvacátého století vyústil v systematické zkoumání jazyka překladu a jeho charakteristických rysů. Na výzkum tzv. překladových univerzálií měl nepochybně největší vliv příklon k deskriptivní translatologii na konci sedmdesátých let dvacátého století a později nástup rozsáhlých jazykových korpusů v lingvistice. Abychom však porozuměli tomu, jak zásadní změnu deskriptivní přístup v translatologii představoval, je třeba si nejprve krátce připomenout, jak se oblasti zájmu translatologů proměňovaly před tímto obratem.

2.1 Klíčové koncepty v dějinách translatologie

2.1.1 Věrný a volný překlad

Ústředním tématem západní translatologie[1] bylo po dlouhá staletí **dilema mezi „volným“ a „věrným“ překladem**, tedy otázka, zda má překladatel překládat doslovně, slovo od slova, nebo převádět především významy bez ohledu na jejich konkrétní lexikální podobu. Slavnou disputaci na toto téma vedl ve svých pracích už svatý Jeroným, překladatel Bible do latiny, dnes považovaný za patrona všech překladatelů. Tato problematická dichotomie, jejíž ozvuky najdeme v různých podobách v díle mnoha translatologů dodnes, se spolu s hlavním tématem překladu Bible držela v centru pozornosti teoretiků prakticky až do poloviny dvacátého století, kdy se do popředí začaly dostávat další klíčové pojmy a koncepty.

[1] Termín „translatologie“ je používaným ekvivalentem anglického spojení *translation studies*. V této době však ještě neoznačuje samostatnou disciplínu, kterou pod tímto názvem známe dnes. Ke skutečnému zrodu a institucionalizaci této disciplíny došlo až v 70. letech (viz 2.2).

2.1.2 Otázka ekvivalence

Po věčných debatách o volném a věrném překladu se od padesátých a šedesátých let dvacátého století badatelé začali pokoušet o systematičtější analýzu překladu. Translatologie se začala obracet k lingvistice a zkoumat klíčové pojmy, jako je význam a především **ekvivalence**. Problému ekvivalence mezi dvěma významovými jednotkami se poté věnovalo mnoho teoretiků překladu i lingvistů, vč. Romana Jakobsona (1959/2000: 114). Mezi nejvýznamnější představitele tohoto období translatologie patří Eugene Nida, J. C. Catford a Peter Newmark.

Eugene Nida, sám překladatel Bible, jež se stala jeho hlavní motivací a předmětem výzkumu, byl jedním z prvních teoretiků, kteří usilovali o to, aby se translatologie stala skutečnou vědou o překladu. Zapojil do své teorie překladu (Nida 1964, Nida & Taber 1969) nejen sociolingvistické a pragmatické koncepty, ale využíval také teorii syntaktických struktur Noama Chomského. Nida ve své práci zdůrazňoval komunikační charakter překladu a domníval se, že je nezbytné zohlednit příjemce překladu a mít na zřeteli rozdílnost jazyků a především kultur. Podíváme-li se na Nidovu teorii prizmatem diskuzí předchozích desetiletí, lze říct, že jeho koncept *dynamické* ekvivalence, jejímž cílem je ekvivalentní *účinek* na příjemce překladu, nikoli ekvivalentní forma a obsah slova (jako u ekvivalence *formální*), stojí v protikladu k doslovnému překladu. Nida si za svou teorii vysloužil i kritiku; odpůrci jako Lefevere neb Larose namítali, že účinek překladu ve dvou různých kulturách a časových obdobích být ekvivalentní ani nemůže, nehledě na to, že jej nelze pořádně změřit (Munday 2008: 43). Nesouhlasné reakce vzbuzoval i jeho systém transformací jádrových struktur, který postrádal jasnější definici.

Podobné rozdělení jako Nida uplatnil o něco málo později ve své koncepci i translatolog a brněnský rodák Peter Newmark, jehož díla jsou dodnes oblíbeným a hojně využívaným učebním materiálem na překladatelských kurzech (Newmark 1981, 1988). Newmark rozlišoval mezi *komunikativním* a *sémantickým* překladem, přičemž komunikativní překlad má podle Newmarka za cíl vyvolat u čtenáře překladu co možná nejpodobnější účinek, jako měl původní text na své čtenáře. Cílem překladu sémantického je pak převést přesný kontextový význam originálu (Newmark 1981: 39). Navzdory podobnosti s Nidovou teorií se však Newmark proti celému konceptu ekvivalence vymezoval a své termíny podrobně definoval z hlediska různých parametrů: kultura, vztah ke zdrojovému textu, čas a původ atd. (1981: 39–69). Newmarkovi bývá vytýkán jeho silně preskriptivní přístup, pro svůj důraz na praktické příklady a ukázky je však jeho dílo stále velmi populární mezi začínajícími překladateli.

2.1.3 Posuny v překladu

Dalším teoretikem překladu, který svou teorii překladu otevřeně nazval lingvistickou, je J. C. Catford (*A Linguistics Theory of Translation*, 1965). Catford vychází z Firthova a Hallidayova lingvistického modelu (Catford 1965: 1). Na překlad nahlíží jako na jazykovou operaci, při níž je zdrojový text nahrazen ekvivalentním překladem, a snaží se definovat podstatu a podmínky této ekvivalence na různých

jazykových rovinách (slovo, věta, text). Kromě toho zavádí do translatologie pojem „překladový posun“[2] (*translation shift*)(Catford 1965: 73), jenž se stal klíčovým termínem i v československé translatologii a který do velké míry souvisí i s tématem překladových univerzálií. Ačkoli Catford byl prvním, kdo ve své práci použil termín posun, ke kategorizaci obdobného jevu již směřovali i Vinay a Darbelnet o několik let dříve ve své komparativní stylistice angličtiny a francouzštiny (1958).

Téma jazykových posunů mezi originálem (zdrojovým textem) a překladem (cílovým textem) se odrazilo i v díle českého teoretika překladu Jiřího Levého, který spolu s Antonem Popovičem a Františkem Mikem patří mezi představitele stylisticky orientované translatologie. Levý hovoří v souvislosti s posuny o „stylistickém ochuzování slovníku“ a rozlišuje tři typy:

1. užití obecného pojmu místo konkrétního přesného označení
2. užití stylisticky neutrálního slova místo citově zabarveného
3. malé využití synonym k obměňování výrazu. (Levý 1983: 138)

K prvnímu typu Levý na jiném místě dodává:

> „Příčina tohoto jevu je zřejmá. V rámci skupiny významově příbuzných výrazů se jednotlivá slova objevují v běžném užití s různými průměrnými frekvencemi, a mají tudíž různé stupně prediktability; slova s větší průměrnou frekvencí se při překladatelově hledání vhodného výrazu vynoří první.“ (Levý 1971: 149)

V tomto konstatování můžeme spatřovat první náznaky jevu, který později začal být označován souhrnným názvem „simplifikace“ v překladu (viz kapitola 4.2). Kromě předzvěsti této univerzálie se v díle Jiřího Levého dá najít i odkaz na „explicitaci“, příp. „normalizaci“, ačkoli Levý tento výraz ještě nepoužívá v daném smyslu (o normalizaci blíže viz kapitola 4.4.1):

> „[...] překladatel při konstruování vět směřuje k vysvětlení logických vztahů mezi myšlenkami i tam, kde nejsou vyjádřeny v textu původním, k vysvětlení všech zlomů v myšlení nebo změn v perspektivě, k „normalizaci“ výrazu.“ (Levý 1971: 149)

Levý tedy popisuje tendence k posunům jako napětí „mezi obecným a specifickým pojmenováním, mezi pojmenováním stylisticky neutrálním a pojmenováním expresívním, mezi opakováním a obměnami slovního označení“ (Levý 1983: 144). Tvrdí, že překladatelé mají spíše „sklon k zobecňování, neutralizaci a opakování“ (Levý 1983: 144). Ačkoli Levý k těmto posunům přistupuje spíše jako ke změnám negativního charakteru (viz termín „ochuzování“), připouští, že „mnohé z těchto ztrát jsou při překladu nevyhnutelné“ a „překladatel by měl ztráty kompenzovat tím, že vyzdvihne stylistické hodnoty, které jsou v textu obsaženy latentně, a využít

[2] Často též „překladatelský posun“, ale vzhledem k významu považuji za přesnější ekvivalent „překladový posun“, příp. „posun v překladu“.

výhod českého jazyka“ (Levý 1983: 144). Všechny tyto posuny jsou nazírány z perspektivy zdrojového textu vůči cílovému, v současné terminologii bychom je proto popisovali v rámci tzv. S-univerzálií (viz 2.3.4).

Posuny a jejich kategorizací se ve svém díle zabýval i slovenský translatolog, literární historik a teoretik Anton Popovič (*Teória umeleckého prekladu*, 1975), který ve své teorii vycházel mj. z Mikovy výrazové soustavy (důraz na rovinu stylu). Funkčnost překladu tak Popovič měří na základě výrazové ekvivalence, jíž rozumí stylovou rovnocennost prvků, a zkoumá, nakolik může výrazová soustava sloužit jako východisko pro charakteristiku změn realizovaných v překladu (Popovič 1975: 112). Jádro textu, které je zachováno jak v originálu, tak v překladu, pak označuje za invariant. Posun nechápe jako nedostatečnou věrnost vzhledem k originálu, nýbrž jako skutečnost, že se z textu něco realizuje a něco jiného vypustí. Hovoří o posunech konstitutivních, které jsou objektivně nevyhnutelné (důsledkem rozdílnosti jazyků nebo dobových konvencí) a individuálních (jež jsou vyvolány překladatelem) (Popovič 1975: 132). Mezi ty individuální spadají i tendence překladatele k explicitaci, konkretizaci, stereotypizaci či retardaci, které mohou z dnešního pohledu opět odkazovat k překladovým univerzáliím.

2.1.4 Funkcionalistické teorie

Ačkoli někde již od sedmdesátých let začaly převažovat deskriptivní přístupy k překladu (viz 2.2), v Německu byla sedmdesátá a osmdesátá léta ve znamení odklonu od statických lingvistických teorií v translatologii a nástupu funkcionalistické školy a vlivné teorie skoposu. Mezi hlavní představitele patří Katharina Reissová, Hans J. Vermeer a Christiane Nordová. Teorie Kathariny Reissové z počátku sedmdesátých let sice navazuje na koncept ekvivalence, ale za základní jednotku ekvivalence považuje **celý text**, nikoli slovo nebo větu. Ve svém funkčním přístupu Reissová propojuje jazykové funkce, typy textu a překladatelské strategie. Vychází z Bühlerových funkcí jazyka a na jejich základě vyděluje tři různé textové typy: informativní, expresivní a operativní, kterým přisuzuje odlišné překladatelské strategie (Munday 2008: 72). Ačkoli s sebou tento přístup přinesl nespornou výhodu v podobě zkoumání vyšší roviny jazyka (textu), Reissová si za kategorizaci textových typů vysloužila i kritiku (často se v ní objevovaly výtky, proč jsou jen tři?, skutečně se od sebe tolik liší?). Katharina Reissová je rovněž spoluautorkou zásadní publikace *Grundlegung einer allgemeine Translationstheorie* (Vermeer & Reiss 1984), která tvoří základ teorie skoposu (*Skopostheorie*).

Termín „skopos“ pochází z řečtiny a znamená „účel“. Do translatologie jej zavedl Hans J. Vermeer jako označení účelu překladu, který má určující vliv na překladatelské metody a strategie. Ve své knize se Vermeer a Reissová pokusili o obecnou teorii překladu, jež by zahrnovala všechny texty. První část knihy je tak věnována obecné teorii, tedy podrobnému vysvětlení teorie skoposu, a druhá část pojmenovaná Speciální teorie je založena na textových typech podle Reissové. Skopos bychom mohli charakterizovat jako účel překladu pro dosažení nějakého cíle/záměru iniciátora/zadavatele překladu u příjemce překladu v cílové kultuře. Zásadní roli tak v teorii skoposu hraje zadání, z něhož by měl jasně vyplynout skopos překladu

a odtud i vhodná překladatelská metoda a strategie. Aby však osud textu neležel pouze v rukou zadavatele, stanovují autoři teorie dvě pravidla: pravidlo koherence a pravidlo věrnosti. Z pravidla vnitrotextové koherence, které je druhému pravidlu nadřazeno, vyplývá, že překlad musí mít vnitřní soudržnost, laicky řečeno, musí cílovým čtenářům dávat smysl. Pravidlo věrnosti neboli mezitextové koherence pak říká, že přeložený text by měl být odpovídat originálu.

V teorii skoposu je tak zdrojový text „sesazen z trůnu“ a jeho místo zaujímá skopos. Tento důraz na cílový text a jeho podobu souvisí s postupným obratem k deskriptivnímu pojetí překladu, kdy jsou překladové texty chápány jako samostatné jednotky (nikoli pouhé odvozeniny od textů zdrojových) a součást cílové kultury (viz 2.2.2.). Kdybychom chtěli teorii skoposu uplatnit při hledání překladových univerzálií, nejspíš by to nebylo možné, neboť podle této teorie je každý akt překladu svým způsobem unikátní a případné změny či odchylky v cílovém textu se mohou odvíjet od konkrétního skoposu. Podle této teorie tak prakticky nelze hledat rysy, které jsou univerzální všem překladům.

Teorie skoposu bývá často kritizována za příliš účelový postoj k překladu a překládání. Bývá jí vytýkáno, že ji lze aplikovat prakticky jen na neliterární texty, u literárních překladů totiž klasické zadání v naprosté většině případů nenajdeme. Další výtkou je nedostatek pozornosti věnované nižším jazykovým rovinám – u některých překladů sice může dojít k naplnění skoposu, ale na dílčí úrovni segmentů není po stránce stylistické či sémantické překlad adekvátní. V reakci na tuto kritiku představila Christiane Nordová (Nord 1988) svůj model **textové analýzy** pro překladatele, která se zabývá zdrojovým textem na úrovni věty a vyšší. Jejím cílem bylo poskytnout studentům překladatelství model analýzy zdrojového textu, který bude univerzálně použitelný na všechny textové typy a překladatelské situace. Její model je založen na pochopení funkčních rysů zdrojového textu a na správném výběru vhodné strategie. Stejně jako její předchůdci i Nordová zohledňuje účel překladu a zachovává funkční hledisko, ale klade mnohem větší důraz na konstituční rysy zdrojového textu.

Kromě deskriptivní translatologie, které je věnována celá následující podkapitola, se v 70. až 90. letech začaly objevovat i teorie, které se pokoušely propojit translatologii se stále populárnější analýzou diskurzu nebo žánrovou analýzou (*Register Analysis*) nebo navázat na poznatky sociolingvistiky a pragmatiky. Za všechny uveďme Juliane Houseovou, jejíž práce *Translation Quality Assessment: A Model Revisited* z roku 1997 využívá hallidayovský model jazyka a diskurzu, nebo Basila Hatima a Iana Masona zahrnující do své teorie sémiotiku a pragmatiku (podrobnější informace např. Munday 2008: 89–106).

2.2 Nástup deskriptivní translatologie

Na konci sedmdesátých let se translatologie od preskriptivních a normativních teorií odklání. V centru zájmu už není ekvivalence ve smyslu stejného účinku na příjemce překladu, nýbrž cílový text jako takový. Překlad už není vnímán jako pouhá odvozenina či náhražka zdrojového textu, kterou je třeba pečlivě porovnávat s ori-

ginálem a posuzovat, kde jsou její nedostatky (Kruger 2002: 77). Funguje sám za sebe a hlavní kritérium pro jeho hodnocení už nevyplývá z původního textu. Tento trend se odrazil již ve výše zmíněné teorii skoposu a především pak v teorii polysystému a manipulační škole.

Důležitost existence samostatné vědy o překladu a význam deskriptivního pohledu zdůraznil jako první James S. Holmes v roce 1972 ve své zásadní stati „The name and nature of translation studies“, která byla přednesena na Třetí mezinárodní konferenci aplikované lingvistiky v Kodani a Holmesovi vynesla označení „zakladatel translatologie“[3]. Holmes se v ní zabývá postavením translatologie mezi ostatními disciplínami a předkládá rámec oblastí, které by translatologie měla zahrnovat. Tuto „mapu“ translatologie od něj poté převzal a ve své knize rozpracoval významný deskriptivní translatolog Gideon Toury (1995: 10). Translatologie je na Holmesově mapě rozdělena na „čistou“ a „aplikovanou“, přičemž čistá se dále dělí na teoretickou a deskriptivní. Cílem deskriptivní větve je podle Holmese popis překladových jevů, zatímco ta teoretická se má zabývat vytvořením obecných principů, které by tyto jevy dokázaly vysvětlit či předpovídat.

Od sedmdesátých let se tak postupně objevovaly různé teorie, které zdůrazňovaly tu či onu větev Holmesovy mapy. V Německu převažovala lingvisticky orientovaná věda o překladu a teorie textových typů a skoposu, ve Velké Británii a Austrálii se dařilo teoriím vycházejícím z hallidayovské systémové funkční gramatiky a analýzy diskurzu. V Tel-Avivu se pak zrodila deskriptivní teorie nazvaná teorie polysystému, která měla velký vliv na další vývoj v translatologii.

2.2.1 Itamar Even-Zohár a teorie polysystému

Tel-Aviv se stal průkopnickým centrem deskriptivní teorie na konci 70. let především zásluhou Itamara Even-Zohára. Even-Zohár si vypůjčil myšlenky ruských formalistů z dvacátých let, kteří se věnovali literární historiografii. Na literární dílo nenahlíželi jako na izolovaný text, nýbrž jako na součást celého systému literatury. Právě „systém“ se stal pro Even-Zohára klíčovým pojmem. Podle Even-Zohára není překladová literatura nedůležitou či podřadnou součástí systému, nýbrž představuje sama svébytný systém. Even-Zohár tak přichází se zastřešujícím termínem **polysystém** a zkoumá, jak spolu všechny tyto systémy (kultura, literatura, překladová literatura atd.) interagují a jak se mění jejich postavení v rámci určité hierarchie. Pokud jsou například v určitém historickém momentu v centru pozornosti inovativní literární díla, je pravděpodobné, že konzervativní díla se přesunou na periferii, a naopak. Důležitým a inovativním prvkem této teorie je i fakt, že za překlad označuje ty texty, které cílová/přijímající kultura za překlad *považuje* (angl. termín *assumed translation*), tedy texty, jež odpovídají normám, kterými se překladové texty v dané době a dané kultuře řídí.

Překladová literatura tak v cílové kultuře může mít buď primární, nebo sekundární postavení, jinými slovy posouvá se do centra nebo na periferii poly-

[3]Vzhledem k zavedenému českému ekvivalentu je poněkud paradoxní, že Holmes ve své stati odmítl název „translatology“ a zavedl termín „translation studies“, který se v anglofonním světě používá dodnes.

systému podle dobových konvencí a norem. Za „normální“ považuje Even-Zohár spíše sekundární postavení překladové literatury v rámci polysystému. Samotná překladová literatura se však dále stratifikuje a překlady z různých jazyků nebo kultur mohou mít různě významné postavení.

Proč měla teorie polysystému tak zásadní vliv na vývoj moderní translatologie? Shrnutí nejdůležitějších důvodů najdeme u Kennyové (2001: 49) nebo Krugerové (2002: 78). Zaprvé, díky teorii polysystému se o překladové literatuře začalo uvažovat jako o svébytném systému, jenž stojí za samostatný výzkum. Zadruhé, překladovým textům se připisují určité specifické rysy, díky kterým lze tyto texty zkoumat v rámci jednoho koherentního celku, např. korpusu. A konečně zatřetí, uvážíme-li, že překladová literatura funguje jako systém v *cílové* kultuře, pak máme důvod zkoumat tyto texty *ve srovnání s* nepřekladovou literaturou cílové kultury. Tato východiska spolu se zákony Gideona Touryho tak na konci 20. století inspirovala hledání univerzálních rysů překladového jazyka (viz 2.3.4).

2.2.2 Gideon Toury: první krok k zákonům překladu

Cílem Gideona Touryho bylo vytvořit obecnou teorii překladu (*In Search of a Theory of Translation*, 1980). Ve své druhé knize *Descriptive Translation Studies – And Beyond* (1995) se inspiroval Holmesovým rozdělením translatologie a teorií polysystému svého kolegy z Tel-Avivu Even-Zohára a podrobně rozpracoval metodologii pro deskriptivní translatologii („descriptive translation studies“, DTS). Profesionální překlad v Touryho pojetí má být takový, jaký cílová kultura očekává, tedy ve shodě s normami a konvencemi cílové kultury. U Touryho najdeme i mnohokrát zkoumaný pojem ekvivalence, ale v jiném smyslu než u předchozích teoretiků. Ekvivalence je v DTS vlastně pojmenováním vztahu mezi originálem a překladem a kulturně a historicky podmíněnou veličinou (v různých dobách i kulturách může být za ekvivalentní překlad považováno něco zcela jiného).

Toury ve své knize předkládá případové studie (využívající především překlady z a do hebrejštiny), jejichž cílem je pomocí srovnání segmentů zdrojového a cílového textu odhalit tendence v překladu, vyvodit z nich závěry o rozhodovacím procesu překladatele, poté na tomto základě „rekonstruovat“ normy, které překlad ovlivnily, a formulovat hypotézy, které budou moci být testovány v rámci dalších deskriptivních překladových studií. Propracovaný systém norem tak tvoří důležitou součást Touryho konceptu (podrobněji o normách viz Toury 1995: 55 a dál).

Toury se domnívá, že identifikace těchto norem nám umožní formulaci probabilistických „zákonů“ překladu (Munday 2008: 114). Podle Touryho (1995: 259) za zákony nelze považovat pouhé výčty možností, ke kterým může při překladu dojít, ani poučky, protože ty i přes častý normativní charakter nemusejí vůbec odrážet pravidelné a skutečně se vyskytující jevy. Každý objevený a správně formulovaný zákon podle něj musí mít povahu podmínky typu: *jestliže platí X, pak s větší/menší pravděpodobností bude platit Y, přičemž Y je pozorovaný výsledek konkrétního chování a X je podmiňující faktor* (1995: 265). Kumulace těchto probabilistických tvrzení pak může vést ke skutečně deskriptivní a explikativní teorii překladu:

> „Proceeding this way, translation theory will ultimately become a series of truly interconnected hypotheses, which is the only kind of theory which would offer a possibility of supplementing exhaustive descriptions and viable explanations with justifiable predictions.“ (Toury 1995: 267)

Touryho zákony

Sám předkládá dva možné překladové **zákony**: zákon rostoucí standardizace (*law of growing standardization*) a zákon interference (*law of interference*). Zákon rostoucí standardizace by se dal obecně formulovat takto[4]:

> „In translation, textual relations obtaining in the original are often modified, sometimes to the point of being totally ignored, in favour of [more] habitual options offered by a target repertoire.“ (Toury 1995: 268)

Podle Touryho má překladatel tendenci v překladu oproti originálu více používat obvyklejší a stereotypnější prostředky, čímž dochází k menší variantnosti v textu nebo přinejmenším k většímu přizpůsobení cílovým normám. K standardizaci dochází podle Touryho především tehdy, zaujímá-li překlad v cílové kultuře (polysystému) periferní postavení.

Druhým zákonem, který Toury představuje, je zákon interference (1995: 274–9). Interferencí se myslí přenášení prvků zdrojového textu (především na lexikální a syntaktické rovině) do textu cílového, ať už v „negativním“ slova smyslu (tyto prvky nejsou pro cílovou kulturu typické a bývají pociťovány jako nestandardní a cizorodé), nebo v „pozitivním“ (prvky, které v cílovém jazyce existují a nejsou vnímány jako nepřirozené). Zde je nutno dodat, že termín interference v translatologii obecně má spíše pejorativní nádech a označuje většinou negativní přenos; Touryho pojetí „pozitivní interference“ je tak poměrně ojedinělé. Toury navíc považuje interferenci za nedílnou součást překladu:

> „[...] interference is kind of a *default*, so that the establishment of an interference-free output (or even of an output where interference has been relegated to less disturbing domains) necessitaes special conditions and/or special efforts on the translator's part.“

To, jak překlad se známkami interference přijme cílové publikum, závisí podle Touryho na tom, z jaké kultury text pochází. Texty z velkého a prestižního jazyka či kultury mají tendenci být v cílové kultuře, zvláště je-li tato kultura v určitém smyslu menší nebo minoritní, lépe tolerovány (Toury 1995: 278).

Touryho dílo vzbudilo po právu velkou pozornost. Gentzler (1993: 133–4) spatřuje Touryho hlavní přínos – samozřejmě kromě vlastního založení DTS – především v zapojení literárních tradic cílové kultury během překladu, v destabilizaci pojetí originálu jakožto nedotknutelné a neměnné entity a také v Touryho pojetí ekvivalence.

[4] Je poněkud překvapivé, že Toury při formulaci svých zákonů, byť v obecné rovině, nedbá vlastních zásad pro jejich formální znění ve tvaru podmínky, jež prosazuje o pár stran dříve ve své práci.

Gentzler ovšem také zmiňuje několik kritických výtek vůči teorii polysystému a DTS. Předně upozorňuje na riziko přílišné generalizace při formulování „univerzálních zákonů“, které bývají založeny na malém množství důkazů (několik případových studií na jednom jazyce ještě nepotvrzuje, že jde skutečně o zákon překladu). Je také otázkou, nakolik je možné Touryho systém zákonů a norem ve skutečnosti aplikovat, když jeho normy lze popsat pouze zpětně na základě překladu a odůvodnit pouze pomocí odhadnutých vzorců chování překladatele, které tyto normy údajně ovlivnily. Hermans (1999: 92) se pak ptá, zdali je vůbec možné identifikovat všechny proměnné, které jsou pro překlad relevantní a které mají na tyto zákony vliv.

Munday (2008: 116) také upozorňuje na to, že navrhované dva zákony si v jistém smyslu protiřečí, resp. ukazují opačným směrem: zákon rostoucí standardizace poukazuje na normy cílového jazyka, zatímco zákon interference je zaměřen na zdrojový jazyk. Podobné rozdělení zaznamenal A. Chesterman i u navrhovaných překladových univerzálií (viz kapitola 2.3.4). Pym (2008: 321) však oponuje, že u Touryho zákonů nutně nemusí docházet k rozporu:

> „The main point is that, thanks to these probabilistic formulations, it becomes quite reasonable to have contradictory tendencies on the level of linguistic variables. If social conditions A apply, then we might expect more standardization. If social conditions B are in evidence, expect interference. And there is no necessary contradiction involved.“

Ačkoli Toury nazývá své zákony „univerzálními“, není zastáncem termínu univerzálie (*universal*). Preferuje výraz zákon (*law*), a to z následujícího důvodu:

> „The reason why I prefer *laws* is not merely because, unlike *universals*, this notion has the possibility of exception built into it (which is important from the probabilistic point of view because no probability is ever 1), but mainly because it should always be possible to explain away [seeming] exceptions to a law with the help of *another* law, operating on *another* level.“ (Toury 2004b: 29)

Podrobněji o celém konceptu univerzálií a nejnovějších výzkumech i vědeckých názorech na ně pojednává kapitola 2.3.4 a patřičné oddíly kapitoly 4.

2.3 Korpusový výzkum v translatologii

Dalo by se říct, že propojení deskriptivní translatologie a korpusové lingvistiky bylo prakticky nevyhnutelné; obě disciplíny totiž zastávají obdobný pohled na jazyk. Jak korpusová lingvistika, tak deskriptivní translatologie vychází z empirické perspektivy a zkoumá jazyk na základě reálných dat, nikoli na základě intuitivních předpokladů. Výběr textů ke zkoumání se v obou případech neřídí nějakou předem či obecně danou představou „vhodných“ textů, nýbrž cílem popsat jazyk (resp. překladový jazyk) takový, jaký je. Na rozdíl od Touryho deskriptivní translatologie

se však korpusová lingvistika zpravidla nezajímá o mimojazykové zdroje dat (historické informace, recenze, kritiky nebo rozhovory s autory), jež Toury hojně využívá k objevení norem, které řídí chování překladatele.

Ačkoli už Toury hovoří o využití „korpusu“ textů, nejedná se o tentýž pojem, jak jej zná a definuje současná korpusová lingvistika. V Touryho pojetí jde o relativně malý soubor textů, jenž je sestaven i prohledáván ručně a může zahrnovat například texty konkrétního překladatele nebo texty z určitého časového období (Laviosa 2002: 12). Skutečnou průkopnicí korpusových nástrojů v translatologii a zakladatelkou disciplíny, kterou zde budeme nazývat „korpusová translatologie“ (*corpus-based translation studies*, příp. *corpus translation studies*), je Mona Bakerová, která na počátku devadesátých let svým článkem (Baker 1993) zásadním způsobem ovlivnila posledních dvacet let translatologie a odstartovala honbu za překladovými univerzáliemi.

2.3.1 Několik terminologických poznámek

Ještě před tím, než se blíže podíváme na vývoj korpusové translatologie, je třeba věnovat několik poznámek terminologii. Vzhledem k tomu, že tato translatologická disciplína využívá metodologii korpusové lingvistiky a korpusové nástroje, je logické, že přejala i „korpusovou“ terminologii. Ovšem ne zcela, a to může způsobovat potenciální nedorozumění mezi korpusovými lingvisty a translatology. Cílem následujících několika odstavců je tedy vyjasnit některé terminologické nesrovnalosti a navrhnout vhodné české ekvivalenty k používaným termínům.

Význam spojení *corpus-based*

V anglickém názvu korpusové translatologie se objevuje zdánlivě jednoduché a přímočaré spojení *corpus-based*, jehož použití se však může v obou disciplínách lišit. V korpusové lingvistice se totiž tradičně rozlišuje mezi dvěma metodologickými přístupy (Tognini-Bonelli 2001: 65), pro něž bohužel v češtině neexistuje ustálený ekvivalent (uvedené přibližné překlady a definice jsou převzaty z článku Cvrček & Kováříková 2011: 122). Na jedné straně je to přístup, v němž se postupuje od introspektivně vybudované hypotézy směrem k jejímu ověřování na rozsáhlých datech, tzv. přístup *corpus-based*, tedy „na korpusu založený“. Do protikladu k němu bývá dáván tzv. přístup *corpus-driven*, „korpusem řízený“, který označuje postup, v němž sice badatel vychází od určité své hypotézy či představy (což je ostatně nevyhnutelné u všech typů výzkumu), ovšem je připraven ji na základě dat zcela přeformulovat tak, aby odpovídala reálné situaci; data zde hrají skutečně klíčovou roli. Mohli bychom tedy na přístup *corpus-driven* nahlížet jako na pokračování nebo určitou nástavbu přístupu *corpus-based*.

Je třeba dodat, že vůči této neo-firthiánské[5] dichotomii se dnes mnozí lingvisté ohrazují. Dělení na *corpus-based* a *corpus-driven* totiž úzce souvisí s klíčovou

[5]J. R. Firth byl britský lingvista, který bývá označován za „otce pojmu kolokace“ a jeho dílo zůstává velkým inspiračním zdrojem pro mnohé korpusové lingvisty.

otázkou, která rozděluje korpusovou lingvistiku na dva tábory: je korpusová lingvistika pouhou metodou, nebo představuje i teorii? K přístupu *corpus-linguistics-as-method* (McEnery & Hardie 2012: 150) bychom mohli vztáhnout metodologii *corpus-based*, zatímco přístup *corpus-linguistics-as-theory* by využíval metodologii *corpus-driven*. McEnery a Hardie však zastávají názor, že ve skutečnosti nejde o opozici:

> „The implication of corpus-based versus corpus-driven is that the *primary* difference between the two is the degree to which empirical data from a corpus is relied on. [...] But in fact, respect for the empirical evidence of the corpus is probably one of the closest points of agreement between the two traditions of corpus linguistics."

Domnívají se, že namísto polarizujícího přístupu *buď-anebo*, bychom měli hovořit spíše o škále přístupů:

> „Moreover, the corpus-based versus corpus-driven distinction implies a dichotomy where there is actually a sliding scale. [...] Within what would be dubbed corpus-based linguistics, we see an entire range of roles for corpus, from providing (at most) a series of examples to illustrate a grammatical theory developed independently of corpus linguistics [...] to being the source of most of the claims made [...]."

Kromě výše diskutovaných přístupů bývá v korpusové literatuře zmiňován i přístup označovaný jako *corpus-assisted* („korpusem podporovaný", který by v obecném pojetí McEneryho a Hardieho spadal také mezi *corpus-based*). Používá se především v souvislosti s moderní korpusovou analýzou diskurzu[6] (např. Partington 2010, Duguid 2010, Baker 2006), která na základě analýzy a srovnání klíčových slov v různých textech (např. publicistických článcích za posledních dvacet let) zkoumá proměnu významu určitých pojmů či konceptů.

A konečně výraz *corpus-informed* („korpusem poučený") někdy nalezneme u studií, které sice korpus okrajově využívají, ale nelze je ještě označit za *corpus-based*; v korpusové translatologii se s nimi však nesetkáváme.

Korpusová translatologie

Ačkoli Bakerová ve svém článku hovoří o využití metodologie *corpus-driven* (Baker 1993: 242), pro tento typ výzkumu v translatologii se později vžil obecnější název *corpus-based translation studies*, jak dokladují četné názvy sborníků či konferencí. Termín *corpus-driven translation studies* se nepoužívá, ačkoli v názvech některých korpusově-translatologických prací se objevuje (např. Goethals 2007, Wang 2006). Jako nejvhodnější český ekvivalent celé této disciplíny se tak jeví spojení korpusová translatologie, ačkoli nese bezesporu obecnější význam. To však může být i výhodou, chceme-li pod korpusovou translatologii zahrnout i ty studie, které propagují přístup

[6] ang. *Corpus-Assisted Discourse Studies* (*CADS*), příp. *Modern Diachronic Corpus-Assisted Discourse Studies* (*MD-CADS*)

corpus-driven nebo se pohybují na široké škále korpusových metod v McEneryho pojetí. V angličtině by podobnou neutrální úlohu zastávalo spojení *corpus translation studies*, které na konci devadesátých letech použila např. Maria Tymoczková (1998), ale nejnovější tendence vedou, jak jsme viděli, k obecnému užití výrazu *corpus-based*.

Podobná nekonzistentnost v používání termínů korpusové lingvistiky se projevuje i u rozlišení typů korpusů a označení originálních/nepřekladových textů (viz kapitola 3.1).

2.3.2 Počátky korpusové translatologie

Ve svém prvním zásadním článku „Corpus Linguistics and Translation Studies: Implications and Applications" Bakerová předpokládá, že budování různých typů korpusů a vývoj korpusové metodologie umožní translatologům odhalit „podstatu přeloženého textu jakožto zprostředkované komunikační události" prostřednictvím zkoumání „univerzálních rysů překladu" (Baker 1993: 243).

Translatoložka a komparatistka Maria Tymoczková pak shrnuje zrod korpusové translatologie následovně:

> „Corpus translation studies (CTS) has emerged at a critical time in the discipline of Translation Studies. Growing out of corpus linguistics and thus inherently having an allegiance to linguistic approaches to translation, CTS at the same time marks a turn away from prescriptive approaches to translation toward descriptive approaches." (Tymoczko 1998: 1)

Podle Tymoczkové korpusová translatologie pojímá překlad jako proces i jako výsledný produkt a zkoumá obojí:

> „CTS focuses on both the process of translation and the product of translation, and takes into account the smallest details of the text chosen by the individual translator, as well as the largest cultural patterns both internal and external to the text." (Tymoczko 1998: 2)

Také varuje před tím, aby se z korpusové translatologie nestala disciplína, která jen „objevuje objevené":

> „Researchers using CTS tools and methods must avoid the temptation to remain safe, exploiting corpora and powerful electronic capabilities merely to prove the obvious or give confirming quantification where none is really needed, in short, to engage in the type of exercise that after much expense of time and money ascertains what common sense knew anyway." (Tymoczko 1998: 7)

Laviosová (2002: 21) uvádí, že nejprve byl vliv korpusové lingvistiky na translatologii vnímán jen jako zdroj koherentnější a efektivnější metodologie, ale postupem

času se začaly formovat prvky a tendence, které z tohoto přístupu vytvořily nový, samostatný proud. Zárodky této nové disciplíny lze pak hledat právě v článcích Bakerové, které jsou podle Laviosové velmi inspirativní: „theoretically rich and brimming with ideas, hypotheses and suggestions“ (2002: 22).

O vývoji korpusové translatologie mnohé vypovídá i fakt, jak o této disciplíně referovaly a referují některé ze stěžejních přehledových publikací korpusové lingvistiky. Svého času jedna ze základních knih oboru *Corpus Linguistics* (McEnery & Wilson 1996) zahrnovala mimo jiné i přehled toho, jak se korpusy využívají v různých jazykovědných či příbuzných disciplínách – od pragmatiky či sociolingvistiky až po stylistiku nebo psychologii. O translatologii, příp. kontrastivní lingvistice zde však není ani zmínka. O deset let později, v obdobné referenční příručce do téhož spoluautora *Corpus-based Language Studies* (McEnery, Xiao & Tono 2006), zde již najdeme podkapitolu věnovanou Translation and contrastive studies. Prudký rozvoj této disciplíny je samozřejmě nejvíce patrný v množství vydaných monografií věnovaných přímo korpusové translatologii (např. Laviosa 2002, Olohan 2004, Anderman & Rogers 2008, Oakes & Ji 2012 a další).

V posledních letech našla korpusová translatologie uplatnění také ve výcviku nových překladatelů a v jazykové výuce (spojení korpusových metod a empirického výzkumu přináší inovativní výukové programy). Její poznatky jsou využívány i při rozvoji nástrojů pro strojový překlad (*Machine Translation*, MT) nebo počítačem podporovaný překlad (*Computer-Assisted Translation*, CAT). Celou tuto aplikovanou linii bychom také ve shodě s výše zmíněnou publikací (McEnery, Xiao & Tono 2006) mohli nazvat *praktickou* v protikladu k *teoretickému* směru v rámci disciplíny, který se zabývá samotným zkoumáním procesu překladu a porovnáváním různých lingvistických rysů na paralelních a srovnatelných korpusech.

2.3.3 Třetí kód, překladatelština a interference

Za východiska korpusové translatologie tedy můžeme považovat deskriptivní pohled na jazyk a překlad, orientaci na cílový text a cílovou kulturu a využití metodologie korpusové lingvistiky a empirických dat. Ústředním bodem zkoumání se pak stala myšlenka, že přeložené texty vykazují *určité společné rysy, jež je odlišují od textů nepřeložených*, napsaných v původním jazyce. Ještě před Monou Bakerovou, která pro rysy tohoto typu zavedla termín „univerzálie“, se v pracích translatologů objevovaly pod názvem „třetí kód“, „překladatelština“ nebo „interference“.

Třetí kód

Označení **třetí kód** použil ve své práci William Frawley, jenž vyšel z toho, že proces překladu je vlastně procesem překódování informace:

> „Since every translation is a recodification, the act of translation involves at least two codes. These I shall call the matrix code and the target code. The matrix code is the code of origin of the translation; it is the primary stimulus, the code that demands rerendering. The target code is the goal

> of the recodification, the code into which the matrix code is debatably rendered.“ (Frawley 1984: 252)

Spojením původního kódu a kódu cílového pak vzniká tzv. třetí kód, jenž tvoří jakousi podmnožinu obou kódů a má své vlastní zákonitosti:

> „That is, since the translation truly has a dual lineage, it emerges as a code in its own right, setting its own standards and structural presuppositions and entailments, though they are necessarily derivative of the matrix information and target parameters.“ (Frawley 1984: 257)

Frawley si tak povšimnul důležité věci: překlad považuje za sloučeninu obsahu původního sdělení a formy cílového jazyka: „The theory above has it that the matrix code provides input information.“ Výsledný text, třetí kód, tak vzniká jako konkrétní výběr z parametrů cílového kódu.

Zajímavým způsobem se Frawley staví i k otázce kvality překladu; věrnost předloze totiž automaticky nepovažuje za záruku dobrého překladu:

> „It should be quite evident that there can be no precise way of judging whether a translation is good or bad. Evaluative discussions on recodification are matters of preference solely. Consider, in this regard, the fact that the fidelity of a new linguistic text to its „original“ is often viewed as the criterion of goodness for interlingual translation. But is the „original“ text the matrix or the target code? Each contributes to the genesis of the translation.“ (Frawley 1984: 260)

Frawley tak navrhuje zcela upustit od konceptu *dobrého* a *špatného* překladu a hodnotit překlady spíše na škále mírný/konzervativní – radikální:

> „The fact is that a respectable theory of translation must abandon notions of good and bad (and fidelity) in recodification. And it must do so as readily as it abandoned identity and the ridiculous insistence on „preservation of meaning“. The closest that a theory of translation can come to an evaluative judgment is to label translations as moderate or radical and let the critics judge whether or not the moderate/radical translation is worth the effort to be considered.“ (Frawley 1984: 261)

Výraz třetí kód není v současné době užíván v translatologických pracích příliš často, ale objevuje se například v práci Linn Øveråsové (1998), která jej používá jako zastřešující termín pro popis jevů, které překlady odlišují od textů nepřekladových.

Překladatelština

Mnohem častěji se však v textech setkáváme s výrazem „**překladatelština**“ (*translationese*, vytvořeno podobně jako *journalese* či *legalese*). Tento výraz jako první použil pravděpodobně Martin Gellerstam (1986), ovšem v poněkud jiném významu, než který dnes převažuje. Gellerstam, jenž zkoumal vliv angličtiny ve

švédských překladech, popisuje překladatelštinu jako stopy či otisky (*fingerprints*), které zanechává původní jazyk v cílovém překladu. Jeho pojetí ještě nezahrnuje evaluativní hledisko. Dnes má však překladatelština v naprosté většině případů pejorativní nádech a označuje takový typ jevů v překladu, které nevznikají samotným procesem překladu, nýbrž nezvládnutím cílového jazyka nebo nepochopením originálu. Bakerová definuje překladatelštinu takto:

> „In some cases, when an unusual distribution of features is clearly a result of the translator's inexperience or lack of competence in the target language, this phenomenon is referred to as „translationese“.“ (Baker 1993: 249)

Najdeme však i neutrální interpretace tohoto výrazu, jako např. u Puurtinenové, která tak označuje prakticky jakýkoli specifický projev překladového jazyka:

> „The term *translationese* is used in a neutral sense, simply meaning translation-specific language, with no negative implications. [...] Translationese may be the result of source language interference, in some cases, however, features of translationese cannot be explained by interference.“ (Puurtinen 2003: 391)

Vzhledem k tomu, že v odborné literatuře lze narazit na obě interpretace, musíme vzít v potaz, že pokud dnes použijeme výraz překladatelština, nevyhnutelně se tak uchylujeme k možnému negativnímu hodnocení překladu či spíše překladatele – lze tedy říci, že označení překladatelština není bezpečným a zavedeným synonymem překladového jazyka ani shrnutím jeho předpokládaných univerzálních rysů.

Interference

Podobnou proměnu významu zaznamenal i výraz **interference**. Jak jsme viděli u Touryho, může označovat jak prostou skutečnost, že jazyk originálu má vliv na jazyk překladu (ať už negativní, či pozitivní), tak i jev, kdy do cílového jazyka/textu dochází k přenosu takových rysů původního jazyka, které v cílovém jazyce působí cizorodě, nepřirozeně či nevhodně. V tomto významu se interference většinou vztahuje k vlivu jednoho konkrétního jazyka na cílový jazyk, nikoli k překladovým textům jako celku.

Zdá se, že poslední dobou převažuje druhý význam slova interference, jak dokladují četné příspěvky z aktuálních translatologických konferencí (například ICLC7 – UCCTS3[7]). Interference je často kladena do protikladu k překladovým univerzáliím (De Sutter et al. 2013) nebo bývá nahrazována neutrálním označením „vliv zdrojového jazyka“ (*source language effect*, Granger 2013).

Shrneme-li tedy rozdíly mezi třetím kódem, překladatelštinou a interferencí, můžeme konstatovat, že označení třetí kód má spíše obecný a neevaluativní charakter a v současné době se s ním nesetkáme příliš často; nahradilo jej buď označení

[7]Podrobnější informace dostupné zde `http://www.iclc7-uccts3.ugent.be/`

překladový jazyk nebo se obecně mluví o vlastnostech překladového textu. Pojmy překladatelština a interference spojuje podle mnohých snaha hodnotit negativně konkrétní rysy přeloženého textu a mohly bychom je považovat takřka za synonymní; rozdíl mezi nimi však můžeme spatřovat v tom, na co kladou důraz. Interferencí se myslí negativní přenos rysů konkrétního zdrojového jazyka na výsledný text – důraz je zde kladen na zdrojový jazyk a text. Překladatelštinou lze na druhé straně označit prakticky všechny rysy, které v překladu působí rušivě a nepatřičně, tedy nejen ty, které vznikly přenosem ze zdrojového jazyka (interferencí), ale i ty, za něž může jednoduše neschopnost překladatele srozumitelně se vyjádřit v cílovém jazyce či porozumět originálu. Důraz je v tomto případě kladen spíše na osobu překladatele a jeho nekompetentnost.

Obecnou tendencí posledních let pak je postupné upouštění od takto evaluativních označení a preference výrazů neutrálních – ať už jde o „vlastnosti překladového jazyka“ (*properties*, Neumann 2014) nebo jeho „rysy“ (*features of translation*, Olohan 2004). Termínem, který je v tomto smyslu sice neutrální, ovšem jeho univerzální platnost bývá dnes už zpochybňována, jsou **překladové univerzálie** (*translation universals*). Těm je zapotřebí věnovat celý následující oddíl kapitoly.

2.3.4 Překladové univerzálie

Mona Bakerová ve svém článku z roku 1993 „Corpus Linguistics and Translation Studies: Implications and Applications“ shrnula situaci a vývoj translatologie až do devadesátých let a předeslala možnosti, kudy by se tato disciplína mohla dále ubírat, jaké výzkumné oblasti by měla zkoumat a jaké nástroje využívat. Největší ohlas však vzbudila představením tzv. *univerzálních rysů překladu*:

> „[...] features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistics systems.“ (Baker 1993: 243)

Hned v úvodní definici tak jasně vymezuje univerzální povahu těchto rysů v protikladu k jevům, které vznikají vlivem zdrojového jazyka (tedy interferencí). Poté nabízí příklady překladových univerzálií, jež jsou ovšem založeny na dílčích (nekorpusových) studiích a „běžném pozorování“:

> „Based on small-scale studies and casual observation, a number of scholars have noted features which seem, intuitively, to be linked to the nature of the translation process itself rather than to the confrontation of specific linguistic systems.“ (Baker 1993: 243)

Je pochopitelné, že Bakerová v roce 1993 nemohla disponovat velkým množstvím dat a korpusy, které jsou dnes téměř samozřejmostí; na její hypotézy je tak nutné nahlížet jako na prvotní impuls pro nový pohled na překladový jazyk. Předpokládaná univerzalita rysů překladového jazyka představuje beze sporu velmi lákavou myšlenku, ovšem vzbudila po právu značnou kritiku. Než se však podíváme na problémy teorie univerzálií, nejprve je třeba představit ty rysy, které Bakerová ve svém textu uvádí.

Překladové univerzálie podle Bakerové

Ačkoli se tradičně mluví o čtyřech univerzáliích podle Bakerové, sama jich v původním textu (1993: 243–245) uvádí víc a dokládá je studiemi, z nichž vycházela:

1. A marked rise in the level of explicitness compared to specific texts and to original texts in general. (Blum-Kulka 1986)

2. A tendency towards disambiguation and simplification. (Vanderauwera 1985)

3. A strong preference for conventional „grammaticality“ in interpreting [...] and a similar tendency towards general textual conventionality in a corpus of English translations of Dutch novels. (Vanderauwera 1985)

4. A tendency to avoid repetitions which occur in source texts. (Shlesinger 1991)

5. A general tendency to exaggerate features of the target language. (Toury 1980, Vanderauwera 1985)

6. A specific type of distribution of certain features in translated texts vis-à-vis source texts and original texts in the target language. (Shamaa 1978)

Z výčtu je zřejmé, že zdaleka ne všechny předpokládané „univerzální“ rysy jsou na stejné úrovni a některé stojí dokonce v protikladu. Například tendence k explicitaci (1), tedy snaha dovysvětlit text a doplnit do něj informace, vzbuzuje otázku, zda se nejedná o opak simplifikace (2), jež by se měla, obecně řečeno, projevovat zkracováním a zjednodušováním textu. Tendence vyhýbat se opakování při překladu (4) může souviset s tendencí k větší konvenčnosti překládaného textu (3), pokud v cílové kultuře panují normy, které opakování nepodporují. Obecná tendence přehánět rysy cílového jazyka (5) je natolik vágně formulovaná, že prakticky nelze potvrdit ani vyvrátit. Pravděpodobně však souvisí s později formulovanou „normalizací“, kdy překladatel nevědomky nadužívá jazykové rysy typické pro cílový jazyk, a text se pak jeví „normálnější“ (mohli bychom říci méně variabilní) než texty nepřekladové.

Ve své další studii (Baker 1996: 176–7) seznam univerzálií poněkud reviduje a dále pracuje se čtyřmi, které vešly do širšího povědomí:

1. **simplification** [...] the idea that translators subconsciously simplify the language or message or both

2. **explicitation** [...] the tendency to spell things out in translation, including, in its simplest form, the practice of adding background information

3. **normalisation or conservatism** [...] the tendency to conform to patterns and practices that are typical of the target language, even to the point of exagerating them

4. **levelling-out** [...] the tendency of translated text to gravitate around the centre of any continuum rather than to move towards the fringes

Nedá se ovšem říct, že by revidovaný seznam nevzbuzoval tytéž a další otázky – lze na jedné straně text zjednodušovat (tedy i zkracovat) při simplifikaci a na straně druhé zároveň obohacovat dovysvětlováním, jež s sebou nutně nese doplňování informací (tedy určité prodlužování textu)? Nebo se jedná o kompenzaci překladatele na odlišných rovinách, které nemusí nutně stát v protikladu? Jak objektivně vybrat rysy, které jsou typické pro cílový jazyk a projevují se pak přehnaným způsobem v překladu při normalizaci? Jak lze kvantitativně testovat hypotézu jevu zde ještě nazývaného *levelling-out*? Na některé z těchto otázek nelze odpovědět bez důkladné analýzy jak překladových textů, tak i jejich originálů, tedy na základě analýzy paralelního korpusu. Tato práce je však zaměřena na srovnání překladových textů s nepřekladovými, proto se věnuje jen těm univerzáliím a rysům, které je možné zkoumat pomocí jednojazyčného srovnatelného korpusu (viz 3.2). Takovým univerzáliím je pak věnována kapitola 4.

Kromě zjevných obsahových rozporů je problematické i samotné znění definic. Mnoho kritiků Bakerové považuje její formulace za velmi vágní a příliš obecné. Becher (2010: 8) se například domnívá, že hypotéza o explicitaci, kterou Bakerová převzala od Blum-Kulkové (1986), je ve své podobě neobhajitelná („unmotivated, unparsimonious and vaguely formulated“) a měla by být zcela opuštěna. U Houseové najdeme několik námitek vůči tomuto bakerovskému přístupu, které trefně sumarizují nejčastěji zmiňované **výhrady**. Daly by se shrnout následujícími slovy:

1. Pokud univerzálie v překladových textech existují, nemusí to však být univerzálie *překladové*, ale obecně jazykové, které se projevují v překladových textech.

2. Termíny typu explicitace/explicitnost/simplifikace/konvencionalizace a další jsou značně vágní a vyžadují velmi pečlivé vysvětlení a operacionalizaci.

3. Opomíjí se zde otázka rozdílnosti jednotlivých jazykových dvojic (*language-pair specificity*) a směrovost překladu (*directionality*) – univerzální rys se nemusí projevovat stejně u obou směrů překladu v rámci dané jazykové dvojice.

4. Ani jedna z hypotéz nebere v úvahu specifické rozdíly mezi žánry (*genre-specificity*), které mohou mít na tyto jevy vliv.

5. Také je třeba mít na paměti diachronní vývoj i v rámci kratšího časového úseku (např. 25 let), kdy se užití určitých jazykových jevů může lišit a poté zkreslit výsledky (např. modální částice v němčině v období od 1978–2002).

   (House 2008: 11–12)

Na původní seznam Bakerové bychom tedy měli nahlížet tak, že se jedná o pouhý náčrt rysů, se kterými se lze v překladech různého typu setkat; jejich univerzalita napříč jazyky i žánry zůstává otázkou, kterou se pokoušelo a pokouší zodpovědět mnoho translatologů i lingvistů. Odpovědí však zpravidla nebývá potvrzující ANO, které by ospravedlnilo koncept překladových univerzálií, ale častěji NE, nebo přesněji ANO, ALE (jev sice platí, ale jen pro tuto jazykovou kombinaci, žánr atd., není tedy univerzální).

Jak uvádí Mauranenová a Kujamäki v úvodu sborníku věnovaného problematice univerzálií (2004: 2), sama Bakerová na třetím kongresu EST s názvem *Claims, Changes and Challenges* v roce 2001 od svého termínu *univerzálie* upouští a uvažuje o tom, zda „the term was felicitous after all“. Ačkoli je dnes již téměř jisté, že v naprosté většině případů nejde o skutečně *univerzální* překladové rysy, ale spíše jevy závisející na žánru či zdrojovém jazyku, pravděpodobně z důvodu terminologické kontinuity se i v mnoha současných studiích stále pracuje s termínem univerzálie; se striktním odmítnutím tohoto termínu a jeho nahrazením neutrálním výrazem „rysy/vlastnosti“ (*properties*) se setkáme například v pracích německých translatologů sdružených v projektu CROCO.[8] V této práci je původní termín *univerzálie* využíván především při odkazování na tradičně vymezené jevy (Baker 2003), kdežto v případě nově definovaných rysů překladové češtiny je dávána přednost neutrálním výrazům: rysy, vlastnosti, jevy.

Bakerová si byla vědoma toho, že jí nastíněné hypotézy vyžadují další ověřování na dalších jazykových kombinacích a především na dalších datech. Co se týče metodologie, v obecných rysech navrhla, aby se translatologové, kteří disponují korpusem textů přeložených např. z různých jazyků do angličtiny, pokoušeli nalézt jevy a vzorce, které:

- „occur across the corpus, irrespective of whether the source texts are French, Hebrew or Chinese,
- do not occur, or do not occur to the same degree/with the same frequency, in original English texts.“

  (Baker 1993: 245)

Oba tyto metodologické pokyny se vztahují na univerzálie z hlediska opozice *přeložené* texty v. texty *nepřeložené*, nikoli však původní ve smyslu zdrojové (*source texts*). V případě hledání univerzálií je tak nezbytné rozlišovat mezi rysy, které odlišují přeložené texty od originálu, a těmi, které je vymezují vůči nepřeloženým textům (bez ohledu na zdrojový text). Praktické rozlišení, které je uplatňováno i v této práci, přinesl Andrew Chesterman (2004a: 8, 2004b: 39).

Chestermanovo rozdělení univerzálií

Chesterman se ve svém článku kromě samotného rozdělení (viz dále) zabývá různými typy hypotéz o překladových univerzáliích. Ve shodě s ním můžeme říct, že prakticky všechny hypotézy, které byly doposud o univerzáliích vyřčeny, jsou hypotézami *deskriptivními*; jejich cílem je tedy popis daného jevu (tedy že k němu vůbec dochází). Pakliže se nashromáždí dostatek dat, která hypotézu potvrdí, je možné v dalším kroku formulovat hypotézy *explanatorní*, které se snaží osvětlit příčinu zkoumaného jevu (tedy proč k němu dochází).

[8] `http://fr46.uni-saarland.de/croco/publication_en.html`

Chesterman pracuje s praktickou definicí univerzálií:

> „In simple terms, we can define a translation universal as a feature that is found (or at least claimed) to characterize all translations: i.e. a feature that distinguishes them from texts that are not translations."
>
> (Chesterman 2004a: 3)

Vůči jejich univerzálnosti se však v zásadě vymezuje hned následujícím výrokem:

> „More strictly: to qualify as a universal, a feature must remain constant when other parameters vary. In other words, a universal feature is one that is found in translations regardless of language pairs, different text-types, different kinds of translators, different historical periods, and so on."
>
> (Chesterman 2004a: 3)

Shromáždit potřebné množství dat pro ověření takto absolutně pojaté univerzality překladových rysů je velmi náročné a v praxi zřejmě neproveditelné. Stejně tak nemohou fungovat ani univerzální *preskriptivní* výroky o tom, jaké by překlady *měly* nebo *neměly* být, neboť tato tvrzení nutně předpokládají, že všechny překlady jsou stejné, a tudíž na ně lze vztáhnou tatáž hodnotící kritéria. To samozřejmě není pravda; výroky tohoto typu jsou tak nutně založeny na přílišných generalizacích.

Totéž platí i pro jistý typ kritiky překladu, jež se snaží charakterizovat všechny překlady jako celek (Chestermana 2004: 4), např. *deformující tendence literárního překladu* podle Antoinea Bermana (Munday 2001: 149–151), které popisují jevy, k nimž při překladu dochází. Ovšem nejedná se o hypotézy deskriptivní, jak by se mohlo zdát, ale de facto preskriptivní: už samotným názvem „deformující" dává Berman najevo, že tyto jevy v překladu žádoucí nejsou a že překlady, jež je vykazují, jsou tím pádem méněcennými texty. V českém prostředí zde můžeme vzpomenout na spisovatele Milana Kunderu, který o překladech (především svých vlastních děl) smýšlí podobně a některá jeho pozdější díla tak zůstávají pro českého čtenáře z tohoto důvodu nedostupná.

Tendence popisovat a vymezovat překlad pomocí negativních výroků však není vůbec ojedinělá, naopak. Jak je vidět na předešlých stranách této kapitoly, v dějinách translatologie se překlad vždy analyzoval především ve vztahu ke zdrojovému textu, jako jeho odvozenina (ať už jde o polaritu volný–věrný překlad nebo o zkoumání různých posunů). Každá odchylka od originálu se zpravidla hodnotila na škále vhodná–nevhodná, žádoucí–nežádoucí. Je zcela pochopitelné, že při porovnávání dvou textů jsou lépe patrné právě odlišnosti, nikoli shody, ovšem záleží na postoji badatele, jak tyto odlišnosti interpretuje. Korpusová translatologie se však pokusila odhlédnout od hodnocení a vykročit směrem k objektivnímu popisu překladu a překladového jazyka, bez nutnosti vztahovat se k němu skrze originál.

Právě s postavením překladu a s jeho vztahem ke zdrojovým a nepřekladovým textům souvisí Chestermanovo klíčové **rozdělení univerzálií** na *S-universals* a *T-universals*. Vychází přitom z rozlišení mezi třemi typy textů:

1. překlady v jazyce A
2. jejich zdrojové texty v jazyce B (originály ve smyslu původní texty, jež byly podkladem pro překlad)
3. nepřeklady v jazyce A (originály ve smyslu původně psaná, nepřekladová literatura)

V závislosti na tom, jaké dva soubory chceme zkoumat, Chesterman (2004: 6) rozlišuje **dva typy vztahů**. V obou případech je ke zkoumání překladu zapotřebí *referenční korpus*: buď typu 2., nebo 3., ovšem vztahy mezi soubory mají odlišný charakter. Pokud chceme srovnávat překlad a originál (1. a 2.), zajímá nás pravděpodobně vztah ekvivalence (*equivalence*) v širokém slova smyslu: nakolik si texty odpovídají po obsahové i stylistické stránce. Pokud však porovnáváme překlady s nepřeklady (1. a 3.), o ekvivalenci nelze hovořit; kritériem srovnání je zde přirozenost či přijatelnost překladového textu. Chesterman tento vztah označuje obtížně přeložitelným výrazem jako *target text family fit* nebo krátce „textual fit", což bychom mohli charakterizovat jako spřízněnost či podobnost překladového textu se srovnatelnými texty nepřekladovými.

Na základě tohoto rozdělení pak rozeznává univerzálie dvojího typu. **S-univerzálie** odrážejí vztah mezi překlady a jejich zdrojovými texty (vztah ekvivalence) a týkají se způsobu, jakým překladatelé nakládají se zdrojovým textem (S = *source*). Oproti tomu **T-univerzálie** lze hledat ve srovnání překladových textů s nepřekladovými (vztah podobnosti). T-univerzálie jsou výsledkem toho, jak překladatelé zacházejí s cílovým jazykem (T = *target*). K oběma skupinám uvádí Chesterman (2004: 8) příklady potenciálních projevů. V případě S-univerzálií může jít například o simplifikaci (zjednodušování z hlediska originálu), o explicitaci a s ní související prodlužování překladového textu, snahu o neopakování týchž výrazů, normalizaci dialektu nebo o užívání konvenčnějších kolokací oproti originálu. U T-univerzálií zmiňuje také simplifikaci (ovšem z hlediska srovnání s neřpekladovými texty), užití netypických lexikálních clusterů (Mauranen 2000) nebo naopak nedostatečné zastoupení takových jevů, jež jsou pro cílový jazyk jinak specifické (*unique items hypothesis*, Tirkkonen-Condit 2002, 2004).

Mezi oběma těmito typy univerzálií (či překladových rysů) je tedy třeba pečlivě rozlišovat, protože oba vyžadují pro výzkum různé korpusy. Uvážíme-li, jaký důraz kladou deskriptivní obory, včetně korpusové translatologie, na reprezentativní datovou základnu, je výběr a především složení korpusu pro výzkum skutečně klíčovým krokem, proto je materiálové základně a s ní související metodologii věnována celá kapitola 3.

Jak interpretovat výzkumy o univerzáliích

Jak vyplývá z předchozího přehledu, původní překladové univerzálie podle Bakerové jsou definovány poměrně vágně a široce a jejich bližší vymezení může být značně problematické. Přístup badatelů se liší – téže univerzálii mohou být připisovány odlišné lingvistické projevy. O zmapování této poněkud chaotické situace se pokusil translatolog Federico Zanettin (2013), který na základě dvaceti vybraných studií

z korpusové translatologie, jež se věnují univerzáliím, sestavil pomocný interpretační rámec, skládající se ze čtyř rovin, od nejabstraktnější až po nejkonkrétnější.

Vymezení, které nabízí, je užitečné a přehledné a umožňuje lepší operacionalizaci hypotéz a jejich testování v praxi na reálných datech, proto z něj vycházím i v této práci. Zanettin rozlišuje **čtyři roviny abstrakce**:

1. rovinu **teorie**, která v tomto případě zahrnuje výchozí hypotézu, že všechny překladové texty vlivem procesu překladu sdílejí určité vlastnosti, které je odlišují od podobných nepřekladových textů;

2. rovinu **deskriptivních rysů** podporujících danou teorii, zde konkrétní překladové univerzálie: původní simplifikace, explicitace, normalizace, *levelling-out* podle Bakerové; novější *unique items hypothesis* (Tirkonnen-Condit 2002, 2004), *shining-through* (Teich 2003), *gravitational pull* (Halverson 2003) a další;

3. rovinu **jazykových indikátorů** (*linguistic indicators*), jimiž se realizuje ten který deskriptivní rys na různých jazykových rovinách, např. lexikální hustota (*lexical density*) nebo netypické kolokace;

4. rovinu **formálních operátorů** (*formal operators*), s jejíž pomocí se zmíněné abstraktní lingvistické rysy zkoumají v textech (konkrétní metody výpočtu a testy).

   (Zanettin 2013: 21)

První rovina s výchozí hypotézou, která předpokládá, že překladové a nepřekladové texty se liší, je nejobecnější, a proto velmi obtížně ověřitelná. Je proto nezbytné rozložit tuto výchozí hypotézu na několik dílčích hypotéz, které popisují domnělé rysy překladového jazyka (překladové univerzálie). Aby bylo možné tyto dílčí hypotézy testovat, je třeba každou z nich operacionalizovat a určit, jak se projevuje (rovina jazykových indikátorů), a poté uplatnit vhodné metody výpočtu (rovina formálních operátorů).

U výchozí hypotézy a dílčích hypotéz lze vycházet z dosavadních výzkumů, ačkoli i tam je třeba přesně specifikovat vybrané překladové rysy (viz konkrétní části kapitoly 4), neboť jednotliví badatelé si mnohdy odporují nebo označují obdobný jev různě. Nejdůležitější a zároveň nejobtížnější krok ve výzkumu překladových rysů však představuje samotná identifikace relevantních jazykových indikátorů a s tím související volba vhodných prostředků k jejich zkoumání. Je zjevné, že operacionalizací abstraktních ukazatelů, jakými může být například lexikální kreativita nebo naopak repetitivnost, vždy nutně dochází k redukci či generalizaci; s tím musí badatel u kvantitativního výzkumu na datech tohoto rozsahu počítat. O to důležitější však je pečlivě uvážit a zdůvodnit výběr výzkumných metod a těžit ze smysluplné kombinace kvantitativního a kvalitativního přístupu (viz 3.3.1).

Kapitola 3

Data a metodologie

V posledních letech se začaly ozývat kritické hlasy, že se bakerovský přístup v translatologii neopírá o dostatečně propracovanou metodologii (např. De Sutter, Goethals, Leuschner & Vandepitte 2012). Výtky se týkají nejen definic konkrétních univerzálií a jejich hypotéz (o nichž jednotlivě a podrobně pojednává kapitola 4), ale také obecně teoretických a metodologických otázek, jež zpravidla souvisejí s charakterem zkoumaného souboru dat. Badatelům bývá vytýkán nedostatečný popis využitého korpusu a mnohdy i jeho design. Situaci neulehčuje ani fakt, že v pojmenování korpusů používaných v korpusové translatologii nepanuje vždy shoda. Z toho důvodu je nejprve vhodné uvést na tomto místě stručný přehled typů a názvů korpusů[1], s nimiž se lze v translatologickém výzkumu nejčastěji setkat, a tím rovněž představit klasifikaci korpusů, která je použita v této práci. Těžiště této kapitoly pak spočívá v popisu korpusu Jerome, který byl sestaven speciálně pro účely zkoumání překladové češtiny a tvoří hlavní materiálovou základnu této práce. Závěrečná podkapitola shrnuje základní metodologické požadavky pro korpusovětranslatologický výzkum a uvádí výchozí výzkumnou hypotézu práce.

3.1 Typy korpusů v translatologickém výzkumu

Pojednat na tomto místě celou typologii korpusů je prakticky nemožné a ani to není cílem tohoto oddílu. Korpusy můžeme rozdělit podle mnoha různých kritérií: z hlediska jejich funkce ve výzkumu (na referenční, nereferenční či oportunistické nebo monitorovací), podle obsahu (obecné či specializované), časového zařazení textů (synchronní a diachronní), typu textu (vyvážené nebo žánrové) či jazyka (mluvené či psané) atd. Podstatným kritériem pro využití korpusů v translatologii a obecně i v kontrastivní lingvistice je však především **počet zahrnutých jazyků** v korpusu a s tím související další charakteristiky.

Mezi nejvyužívanější korpusy v kontrastivní lingvistice patří především korpusy vícejazyčné (*multilingual*); translatologie však nachází uplatnění i pro specifický typ korpusů jednojazyčných (*monolingual*). V širokém slova smyslu se za vícejazyčné korpusy označují ty korpusy, které obsahují více než jeden jazyk. V užším chápání

[1] Text následující podkapitoly 3.1 vychází z článku Chlumská (2014).

pak zahrnují tři a více jazyků, protože pro dvojjazyčné korpusy existuje samostatný název (*bilingual*). Zde budeme využívat obecnou dichotomii vícejazyčný – jednojazyčný (McEnery, Xiao & Tono 2006: 47).

Vícejazyčné (a koneckonců i některé jednojazyčné) korpusy můžeme dále rozdělit podle toho, jaké texty obsahují (originály = zdrojové texty pro překlad, typ 2 podle Chestermana; překlady; nepřekladové texty = původně psané, typ 3 podle Chestermana). A právě zde dochází k terminologickým nejasnostem. Nejenže pro tentýž typ korpusu existuje několik konkurenčních – a často protikladných – názvů, ale i samotné označování zahrnutých textů (původní v. nepřekladové) se v různých pracích liší. Mezi nejvíce problematické termíny patří paralelní (*parallel*), srovnatelný (*comparable*) a překladový (*translation/translational*) korpus. Běžně se pod těmito a dalšími názvy můžeme setkat s vícejazyčnými korpusy tří typů (McEnery, Xiao & Tono 2006: 47):

1. s korpusem zdrojových textů a jejich překladů (do jednoho či více cizích jazyků),
2. s korpusem zahrnujícím texty vybrané podle týchž kritérií (žánru, zaměření, délky apod.) v různých jazycích (příp. v různých varietách jednoho jazyka),
3. s kombinací obojího.

První typ zde v souladu s aktuálními tendencemi budeme označovat jako *paralelní* korpus, druhý typ jako korpus *srovnatelný* a třetí jako korpus *reciproční*.

3.1.1 Paralelní korpus

Definice a charakteristika paralelního korpusu

Paralelní korpus je tedy korpus, který obsahuje původní, zdrojové texty v jazyce A a jejich překlady v jazyce B, příp. ve více jazycích. Paralelní korpus může být jednosměrný (*uni-directional*), tj. může obsahovat pouze překlady z jazyka A do jazyka B, nebo obousměrný (*bi-directional*), tedy překlady z A do B i z B do A.

Ačkoli pro paralelní korpus bychom nalezli různé definice, zdá se, že v současné době převažuje zde uváděný význam. V tomto smyslu použila označení paralelní korpus už Bakerová (1995: 230) ve své kategorizaci korpusů v translatologii a v dnešní době jej přebírají další translatologové (Laviosa 2002: 36) i lingvisté (např. McEnery, Xiao & Tono 2006: 47; Hunston 2002: 15). V tomto významu se termín používá i v české korpusové tradici, např. v Ústavu Českého národního korpusu FF UK (paralelní korpus InterCorp[2]) nebo v Ústavu formální a aplikované lingvistiky MFF UK (paralelní treebank PCEDT[3]).

Odlišnou definici prosazoval Stig Johansson (1998: 4), autor dodnes hojně používaného anglicko-norského paralelního korpusu[4], který chápal termín paralelní jako obecný, zastřešující pro všechny zmíněné typy korpusů. Korpus originálů

[2] http://www.korpus.cz/intercorp

[3] http://ufal.mff.cuni.cz/pcedt2.0/

[4] https://www.hf.uio.no/ilos/english/services/omc/enpc/

a překladů (tedy v našem pojetí paralelní) označoval za překladový (*translation corpus*). V tomto významu se však dnes již tento termín zpravidla nepoužívá.

Vytvářet paralelní korpus je velmi náročné, a to nejen z technického hlediska (např. větné zarovnání textů z typologicky odlišných jazyků). Zjevným problémem může být, stejně jako u jednojazyčných korpusů, reprezentativnost. U paralelních korpusů však kromě běžných otázek (po zastoupení žánrů apod.) vyvstávají i další. Paralelní korpusy zpravidla nemívají obecný charakter (ve smyslu vyvážených obecných korpusů, jež se snaží co nejvěrněji zachytit všechny nejčastěji zastoupené žánry a textové typy v daném jazyce), ale z důvodu dostupnosti textů se specializují na jeden či několik málo textových typů/žánrů, např. na právní dokumenty EU, titulky k filmům (např. v korpusu OPUS[5]) nebo beletrii. Také je třeba zmínit, že paralelní korpusy jsou zpravidla vázány na psaný jazyk – paralelní korpus mluveného jazyka (tedy již ne korpus originálů a překladů, ale původních a tlumočených promluv) by byl žánrově ještě omezenější a na výstavbu náročnější. Jako příklad korpusu, který stojí na pomezí psaného a mluveného jazyka, můžeme uvést často využívaný korpus Europarl[6], obsahující přepisy projevů poslanců Evropského parlamentu.

Využití paralelního korpusu

Přes všechny možné nedostatky nebo obtížně řešitelné problémy jsou paralelní korpusy v translatologii i kontrastivní lingvistice zcela nenahraditelným zdrojem dat. V rámci translatologie se dokonce dá říct, že právě paralelní korpusy zprostředkovaly onen přechod od preskripce k deskripci (Baker 1995: 231). Odhalují totiž skutečné problémy, se kterými se překladatelé setkávají, i jejich řešení, a tak představují nedocenitelný zdroj informací pro začínající překladatele nebo studenty překladatelství. Také umožňují zkoumat normy překladu, jak se uplatňovaly v různých kulturních či historických kontextech. Nezastupitelnou roli mají i v odvětví strojového překladu, kde představují hlavní zdroj dat. Podle McEneryho, Xiaa & Tona (2006: 49) však samy o sobě nejsou příliš vhodné pro výzkum rozdílů mezi jazyky, jelikož nelze zanedbat možný vliv překladového jazyka na výsledný text. Pro srovnání jazykových jevů ve více jazycích je tak vhodné doplnit zdroj dat i o vícejazyčný srovnatelný korpus, jenž je vlivu překladu ušetřen (viz 3.1.2).

Obecně řečeno jsou paralelní korpusy ideální pro výzkum toho, jak je myšlenka v jednom jazyce převedena do jazyka druhého – klíčovým slovem by zde byla především ekvivalence. V případě takovýchto studií, kdy je hlavním cílem zjistit, jak se určitý jazykový jev projevuje v druhém jazyce, je však důležitý **směr překladu** (*directionality*). Paralelní korpusy, u jejichž textů není směr překladu uveden (např. není znám zdrojový jazyk, z něhož se překládalo), pro tento typ studií vhodné nejsou. Mohou však zcela jistě posloužit zájemcům o překlad určitého slova či fráze.

V rámci korpusové translatologie se paralelní korpus uplatňuje při výzkumu překladových univerzálií nebo překladového jazyka jako takového – ať už s cílem odhalit interferenci z cizího jazyka nebo poukázat na překladatelštinu (*translationese*). Zde je nutno upozornit na to, že klasický paralelní korpus je vhodný pouze

[5] http://stp.lingfil.uu.se/~joerg/published/ranlp-V.pdf
[6] http://www.statmt.org/europarl/

pro zkoumání S-univerzálií (Chesterman 2004a: 8). Pro výzkum T-univerzálií je zapotřebí jednojazyčný srovnatelný korpus (viz 3.1.2).

3.1.2 Srovnatelný korpus

Definice a charakteristika srovnatelného korpusu

Srovnatelný korpus se tedy skládá z částí (subkorpusů), které byly sestaveny podle stejných kritérií výběru textů/vzorků, a jsou tak obdobně vyvážené a reprezentativní. Ačkoli v korpusové lingvistice se pod pojmem *comparable corpus* rozumí takřka vždy korpus vícejazyčný (tedy složený z obdobně sestavených subkorpusů v alespoň dvou různých jazycích), korpusová translatologie mnohem častěji pracuje se srovnatelným korpusem jednojazyčným. Ani v jednom případě ale nejde o korpus originálů a jejich překladů; pro ten převažuje v úzu označení *paralelní*.

Zatímco **vícejazyčný srovnatelný korpus** zahrnuje originální, původně psané texty ve více jazycích (nikoli tedy zdrojové texty a jejich překlady), **jednojazyčný srovnatelný korpus** lze rozdělit na dvě části: nepřekladovou a překladovou. Obsahuje tedy dva subkorpusy v témže jazyce, opět sestavené podle téhož klíče se srovnatelnou velikostí, reprezentativností a vyvážeností, jeden s texty původně psanými, nepřekladovými (*non-translated*) a druhý s texty překladovými (*translated*). Subkorpus, příp. korpus přeložených textů se dnes někdy označuje za korpus překladový (*translation/al*).

Stejně jako u paralelního korpusu i pro srovnatelný korpus najdeme různé definice (kromě výše zmíněného rozlišení mezi vícejazyčným a jednojazyčným). Aijmerová – Altenberg (1996) a Grangerová (1996: 38) pro označení srovnatelného korpusu používali výraz *paralelní*, který např. Johansson (1998: 4) považoval za obecné označení, pod které se vešel dnešní paralelní korpus i oba typy korpusu srovnatelného. Bakerová (1995: 232) zase pojmem *comparable corpus* odkazuje pouze k jednojazyčnému, translatologickému korpusu, kdežto pro vícejazyčný srovnatelný používá termín *multilingual*. Tento způsob pojmenování však dnes přijímán není (Fernandes 2006).

V dnešní době už význam termínu srovnatelný korpus tolik nekolísá (kromě aspektu vícejazyčnosti/jednojazyčnosti), ovšem jisté rozdíly přece jen najdeme. Zatímco Hunstonová (2002: 15) mezi srovnatelné korpusy zahrnuje i korpusy obsahující různé variety téhož jazyka (ne ve vztahu k překladovosti), např. International Corpus of English[7], který zahrnuje milion slov několika variet angličtiny, McEnery, Xiao & Tono (2006: 48) zastávají opačný názor a tento typ korpusu v rámci jednoho jazyka za srovnatelný nepovažují. Argumentují tím, že všechny korpusy jakožto zdroj pro lingvistický výzkum jsou vždy vhodné pro komparativní výzkum, ať už jsou vícejazyčné nebo jednojazyčné (např. v BNC lze zkoumat mluvený vs. psaný). Pro korpusy typu International Corpus of English tak raději volí termín komparativní (*comparative*). Vzhledem k tomu, že tento typ korpusu by jinak spadal do kategorie jednojazyčný srovnatelný korpus, kde je hlavním zástupcem korpus překladových a nepřekladových textů, má zavedení dalšího termínu patrně svůj význam. Nutno

[7] http://ice-corpora.net/ice/

však říct, že jak korpus komparativní, tak korpus jednojazyčný srovnatelný (translatologický) mají odlišné cílové uživatele, dokonce možná i disciplíny, takže by z kontextu mělo být i bez použití nového termínu patrné, o který druh jednojazyčného srovnatelného korpusu se jedná.

U srovnatelného korpusu musíme otázku **reprezentativnosti** chápat opět poněkud jinak než u korpusu obecného nebo paralelního. Zatímco pro paralelní korpus je zásadní spíše výběr díla a překladu a otázky s tím spojené, srovnatelný korpus je zcela závislý na uplatnění týchž kritérií výběru v obou či více subkorpusech. Vybrané texty, ať už úplné nebo vzorky, by měly patřit k témuž textovému typu, žánru či časovému období. Jejich srovnatelnost tak musí být chápana jako souhrn co možná nejvíce charakteristik, velikostí počínaje, žánrovým zařazením konče. V případě, že srovnatelný korpus není sestaven pečlivě, je zde riziko, že veškerá tvrzení z něj odvozená ztratí svou platnost.

Využití srovnatelného korpusu

Jak již bylo řečeno výše, paralelní a srovnatelné korpusy mají nejen odlišné složení, ale především využití. Vícejazyčný srovnatelný korpus je ideálním zdrojem dat pro kontrastivní výzkum, neboť nehrozí vliv překladového jazyka. Svoje uplatnění ale nachází i v aplikované translatologii, především pak ve výuce překladatelů. Malé a vysoce specializované vícejazyčné srovnatelné korpusy totiž mohou začínajícím překladatelům pomoci seznámit se s charakteristickými prvky žánru či odvětví a osvojit si terminologii, která může v mnoha případech působit na překladatele i v jeho rodné řeči jako cizí jazyk (Friedbichler & Friedbichler 1997, podle McEnery & Xiao 2012: 94).

Jednojazyčný srovnatelný korpus (translatologický, nikoli komparativní) je pak typickým specializovaným korpusem v korpusové translatologii. Slouží k objevování typických rysů překladového jazyka v porovnání s nepřekladovým a tvoří základ výzkumu T-univerzálií (např. tendence k simplifikaci, Chlumská & Richterová 2014). Pro výzkum překladové a nepřekladové češtiny byl nedávno v Ústavu Českého národního korpusu vytvořen již zmiňovaný korpus Jerome[8] (Chlumská 2013), který splňuje kritéria jednojazyčného srovnatelného korpusu. Jeho struktuře se podrobně věnuje další oddíl této kapitoly.

3.1.3 Reciproční korpus

Posledním typem korpusu, který v sobě svým způsobem kombinuje jak paralelní, tak srovnatelný korpus, je korpus „reciproční“ (*reciprocal*) (Zanettin 2011: 21). Ten můžeme charakterizovat jako paralelní korpus, v němž jsou rovnoměrně zastoupeny oba směry překladu, jedná se tedy o zvláštní typ obousměrného paralelního korpusu (*bi-directional parallel corpus*). Reciproční korpusy bývají zpravidla jen dvojjazyčné (např. již zmiňovaný Johanssonův English Norwegian Parallel Corpus[9]), neboť shromáždit stejný počet překladů z a do jednoho jazyka (zvlášť malého) je

[8] http://korpus.cz/jerome.php

[9] https://www.hf.uio.no/ilos/english/services/omc/enpc/

nesnadný úkol. Reciproční korpus tak obsahuje originály jazyka A, překlady do jazyka B, originály v jazyce B a překlady do jazyka A o stejném počtu a pokud možno i srovnatelného charakteru. Umožňuje tak výzkum jak paralelní (oběma směry), tak srovnatelný (originály A a B, překlady A a B, příp. i originály A a překlady A).

Mnozí však namítají (např. Zanettin 2011: 21), že tato srovnatelnost je pouze zdánlivá, neboť nesplňuje základní požadavek srovnatelného korpusu, totiž uplatnění stejných kritérií pro výběr textů, jejich žánrové zařazení apod. Jediným kritériem takto srovnatelného korpusu (např. originály A a B) je totiž jen skutečnost, že jde o zdrojové texty. Řešením této výhrady by však mohl být opačný výchozí postup – pečlivé sestavení srovnatelného korpusu originálů v obou jazycích, u nichž víme, že existují překlady do daného jazyka, a následné doplnění těchto překladů do korpusu.

Zde však zejména u malých jazyků narážíme na problém. Jak upozorňuje Bernardiniová a Zanettin (2004: 57), překlady z malého jazyka a do něj se radikálně liší. Zatímco z velkých jazyků do malých bývá překládáno téměř cokoli, od klasických děl až po současné populární čtivo (detektivky, romány pro ženy apod.), v opačném směru převažují překlady starších kanonických děl tzv. vysoké literatury (*high-brow literature*). Srovnání takto rozdílných literárních děl je takřka nemožné. V této situaci nezbývá badateli nic jiného než omezit svůj reciproční korpus na klasická díla a rezignovat tak na synchronní výzkum, nebo shromáždit to málo srovnatelných moderních textů a jejich překladů a spokojit se s korpusem mnohem menším.

Pokud je možné jej v dané jazykové kombinaci vytvořit, je reciproční korpus bezesporu zajímavým řešením pro translatology, kteří zkoumají překladové univerzálie, umožňuje totiž pátrat jak po S-univerzáliích, tak T-univerzáliích současně a ověřovat možné interpretace i vzhledem k vlivu zdrojových dat.

3.2 Korpus Jerome

Korpus Jerome[10] je jednojazyčný srovnatelný korpus, který je určen ke zkoumání překladové češtiny v porovnání s češtinou nepřekladovou. Před zveřejněním korpusu byl výzkum překladové češtiny obtížněji uskutečnitelný; korpusy synchronního psaného jazyka (řada SYN), kterými čeština disponuje, sice překladové texty zahrnují (viz 3.2.1), ale nebylo pro uživatele snadné si vytvořit vlastní vyvážený subkorpus. Ačkoli byl korpus Jerome zamýšlen především jako materiálová základna pro výzkum představený v této práci, je zdarma přístupný všem registrovaným uživatelům Českého národního korpusu a může sloužit k výzkumu překladové češtiny všem zájemcům z řad translatologů i bohemistů.

Výběr dat je (nejen) v korpusovém výzkumu vždy zásadním krokem, proto se následující podkapitoly korpusu Jerome podrobně věnují. Nejprve budou vysvětlena **kritéria** výběru textů do korpusu (viz 3.2.1), v další části (viz 3.2.2) pak následuje **popis korpusu** z hlediska velikosti, počtu děl / autorů / zdrojových jazyků atd.

[10]Korpus vznikl v rámci vnitřního grantu FF UK 2013 VG027 řešitelek Chlumská – Richterová a byl zveřejněn všem zájemcům začátkem roku 2014 na adrese www.korpus.cz v rámci velké infrastruktury Český národní korpus.

3.2.1 Kritéria výběru textů

Jak již bylo zmíněno výše, při vytváření jednojazyčného srovnatelného korpusu je nezbytně nutné si nejprve určit hlavní kritéria, která budou dodržena u obou částí korpusu, překladové i nepřekladové:

> „The two corpora are set up according to similar design criteria, e.g. according to text genre, topic, time span, distribution of male and female authors, readership, average number of words in each text."
>
> (Laviosa 2002: 36)

Jakkoli se požadavek formulovaný v definici Sary Laviosové jeví jako oprávněný a pochopitelný, v praxi zpravidla není možné dodržet všechna relevantní kritéria, zvláště pokud chceme vytvořit velký korpus v řádu desítek milionů slov. Je nutné stanovit si priority a těm pak podřídit výběr textů do korpusu.

Prioritou při sestavování korpusu Jerome byla především jeho výsledná **velikost**, která by umožnila alespoň částečnou generalizaci výzkumných zjištění o překladové češtině. Srovnatelnost obou částí korpusu byla zachována v následujících kategoriích: počet tokenů u překladů i nepřekladů, typ textu (u vyváženého subkorpusu i žánr, viz 3.2.3), časové rozpětí vydání textu a heterogenita autorů/překladatelů. Požadavek vyvážit korpus z hlediska pohlaví autora či překladatele se ukázal být v přímém rozporu s prioritou velikosti korpusu – pokud bychom skutečně trvali na stejném počtu autorů a autorek (resp. překladatelů a překladatelek), výsledný korpus by dosahoval pouze velikosti v řádu stovek tisíc tokenů.

Jako zdroj dat pro korpus Jerome posloužila **databáze textů ČNK**, konkrétně pak texty zahrnuté do korpusu SYN[11]. V korpusech psané češtiny SYN (kromě publicistických) tvoří překlady přibližně třetinu všech textů, ale poměr se liší v závislosti na typu textu[12] v korpusech SYN, uvádí tabulka 3.1).

překlady	beletrie	odborná literatura	publicistika
SYN2010	66 %	20,3 %	0 %
SYN2005	58,2 %	30 %	0 %

Tabulka 3.1: Počet překladů v korpusech SYN2005 a SYN2010

Počty zahrnutých překladových děl (v průměru) přibližně odráží skutečnou situaci překladové literatury u nás (viz tabulka 3.2). Avšak texty se do těchto korpusů synchronní psané češtiny zahrnují podle jiných požadavků, než které je třeba zohlednit u translatologického korpusu, proto bylo nutné vybrat vhodné texty **ručně** a doplnit ke každému také translatologicky relevantní anotaci (viz dále). Korpus byl poté na Ústavu Českého národního korpusu po technické stránce připraven ke zveřejnění.

[11] `http://wiki.korpus.cz/doku.php/cnk:syn`

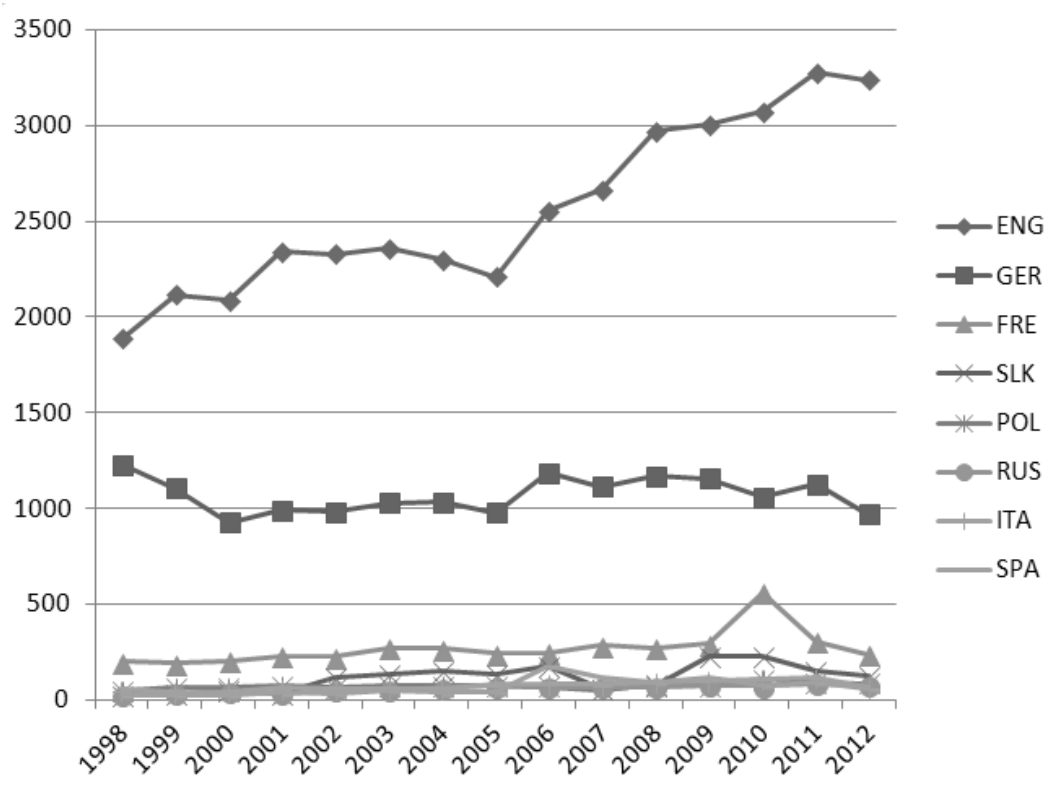
[12] U publicistiky je však třeba brát v potaz, že část textů mohou být ve skutečnosti překlady nebo zčásti převzaté články, ačkoli to u nich nebývá explicitně uvedeno. Vzhledem k tomu, že publicistika není předmětem tohoto výzkumu, ponechávám toto téma stranou.

Zastoupení jazyků

Korpus pro výzkum překladových rysů by měl v ideálním případě obsahovat v překladové části obdobný počet textů (resp. tokenů) z co možná nejvíce typologicky odlišných jazyků, aby se vyloučila možná interference z převažujícího zdrojového jazyka. Tento požadavek předpokládá, že tvůrce korpusu má k dispozici obdobné množství textů ve vybraných jazycích. To však – zvláště u malých jazyků – ve skutečnosti neplatí. Zpravidla je zde jeden zdrojový jazyk, jehož překlady výrazně převažují nad ostatními – v posledních letech je tím jazykem nejen v českém prostředí **angličtina**.

rok	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
celkem	16451	15749	15350	17019	18029	18520	17598	17054	18985	17247
překlady	4602	4604	4423	5240	5384	5546	5777	6044	6514	5871
ENG	2362	2301	2211	2556	2665	2969	3005	3071	3276	3238

Tabulka 3.2: Počet vydaných neperiodických publikací v letech 2003-2012



Obrázek 3.3: Nejpřekládanější jazyky podle statistik NKP (v počtu děl)

Podíváme-li se podrobněji na statistiky překladové literatury (neperiodických publikací – tedy beletrie a odborné literatury)[13] za desetileté období, zjistíme, že

[13] `http://www.nkp.cz/sluzby/sluzby-pro/sluzby-pro-vydavatele/vykazy`

překlady z angličtiny tvoří více než polovinu celkového počtu překladů (viz tabulka 3.2). Za angličtinou se dlouhodobě drží němčina a francouzština a poté slovanské jazyky, ruština, slovenština, polština (viz graf 3.3).

Dá se tedy říci, že český čtenář přichází do kontaktu převážně s překlady z angličtiny a ty se pak významně podílejí na jeho představách o překladovém jazyku. Z tohoto předpokladu tak vychází celá **koncepce korpusu Jerome**. Jeho cílem je zobrazit překladovou češtinu tak, jak se s ní pravděpodobně setkává běžný uživatel jazyka, tedy i s převahou textů přeložených z jediného jazyka. I v korpusu je proto ponechán jako převažující zdrojový jazyk angličtina (přesné údaje o počtu jazyků a děl viz 3.2.2). Tím se podařilo zachovat i hlavní prioritu korpusu – jeho velikost. Ovšem aby bylo možné v dalším výzkumu rozeznat překladové rysy/univerzálie od možné interference angličtiny, byl v rámci korpusu Jerome sestaven také jazykově vyvážený subkorpus (viz 3.2.3).

Autor a překladatel

> „The translation corpus should be representative in terms of the range of original authors and of translators.“
>
> (Baker 1995: 234)

Důležitým faktorem při sestavování korpusu je jeho **heterogenita**. Zvláště u menších korpusů může převaha jednoho autora významně zkreslit výsledky zkoumání, neboť nelze rozlišit vliv idiolektu od rysů překladového jazyka. Korpus Jerome je sice co do velikosti rozsáhlým korpusem, ale i zde byla snaha zachovat co možná největší pestrost autorů a děl.

Při výběru textů tak bylo stanoveno **pravidlo tří** – žádný autor není v korpusu zastoupen více než třikrát. Pokud je autor součástí kolektivu autorů, počítá se tento kolektiv jako další autor a může být opět zastoupen až třikrát. Počet tři byl stanoven arbitrárně jako kompromis mezi snahou o heterogenitu a úsilím o zařazení co možná nejvíce textů do korpusu. U kategorie překladatel bylo toto pravidlo ještě doplněno o omezení, že tentýž překladatel se sice může v korpusu objevit až třikrát, ovšem pokaždé s překladem jiného autora.

Jak uvádí Olohanová (2004: 48) nebo Hareide & Hofland (2012: 87), **pohlaví autora a překladatele** („gender“) také může hrát důležitou roli proměnné ve výzkumu. Z toho důvodu byly do korpusu Jerome tyto údaje doplněny, ovšem nebyly zohledněny jako kritérium výběru, neboť to by výslednou velikost korpusu spolu s ostatními kritérii snížilo na minimum. Dostupnost těchto údajů však znamená, že si uživatel může sám vytvořit vlastní subkorpus např. pouze překladatelek překládajících mužské autory nebo naopak, příp. vyloučit překlady pořízené kolektivem autorů nebo se na ně naopak zaměřit.

Zde je na místě zmínit, že kromě autora a překladatele se na výsledné podobě textu může významnou měrou podílet i **redaktor** či korektor. Vzhledem k tomu, že jejich identitu u mnoha textů neznáme a informace o ní není součástí bibliografického značkování, nebylo možné vliv redakční práce při sestavování korpusu zohlednit. Můžeme však předpokládat, že zmíněná heterogenita autorů, překladatelů (a také

nakladatelství) zaručuje, že na textech pracovali různí redaktoři a nemělo by tedy docházet k přílišným zkreslením.

Doba vydání

Dalším podstatným kritériem při tvorbě srovnatelného korpusu je rok vydání díla. Cílem korpusu Jerome je zobrazit **současný jazyk**, ten lze ovšem definovat různým způsobem. Tradičně se současný jazyk chápe jako období tří generací (Cvrček et al. 2010: 34). Toto období je však z hlediska vývoje moderní češtiny příliš dlouhým časovým úsekem – důležitým předělem je zde rok 1989 a začátek devadesátých let, kdy došlo k proměně jazyka. Do korpusu Jerome proto byly zařazeny texty, které byly vydány mezi lety **1992–2009** (rokem 2009 končí nejnovější tituly zveřejněné v korpusu SYN2010).

Datem vydání se zde však myslí datum, kdy vyšla ta která konkrétní kniha zařazená do korpusu, nikoli datum prvního vydání díla. Samotná skutečnost, že je kniha vydána znovu, svědčí o tom, že je po ní poptávka, že je čtena a tudíž se podílí na recepci textů současného jazyka. Datum prvního vydání však zcela jistě relevantní je, proto byla informace o něm doplněna do anotace (viz dále). Uživatel si tak u menších sond bude moci ověřit, zda charakteristika jazyka díla není ovlivněna dobou jeho vzniku.

3.2.2 Charakteristika korpusu a jeho složení

Začneme-li obecnou charakteristikou, korpus Jerome je stejně jako všechny korpusy synchronní psané češtiny vytvořené v rámci ČNK lemmatizovaný a morfologicky značkovaný. **Lemmatizace**[14], tedy přiřazení základového (slovníkového) tvaru (lemmatu) každé pozici[15] v korpusu, je u flektivních jazyků, jako je čeština, obzvlášť užitečná; umožňuje vyhledat najednou celé tvarové paradigma slov.

Při **morfologickém značkování** je poté každému slovu v korpusu dodána značka (tag) s hodnotami příslušných morfologických kategorií, vč. slovního druhu, podle kterých lze také vyhledávat a zobrazit si tak například frekvenční seznam všech spojek či předložek v korpusu.

Anotace textů

Kromě anotace jednotlivých pozic v korpusu bývá korpus zpravidla anotován prostřednictvím tzv. strukturních atributů, které se vztahují k celé strukturní jednotce, nejčastěji k celému textu (opusu). Textová anotace psaných textů (jinými slovy metainformace) obvykle zahrnuje informace o autorovi, názvu díla, roku a místu vydání, nakladatelství, příp. zdrojovém jazyce a překladateli. Korpus Jerome disponuje doplněnou anotací, ve které nechybí uvedení prvního vydání, informace o tom, zda se jedná o překlad, a údaje o pohlaví autora a překladatele.

[14]Podrobnější informace o lemmatu viz `http://wiki.korpus.cz/doku.php/pojmy:lemma`

[15]Pozicí zde rozumíme veškeré řetězce znaků oddělované v korpusu z obou stran mezerami, tj. řetězce alfabetických znaků (neboli slova), číslice, interpunkci, příp. kombinaci uvedeného.

Následující kompletní **seznam strukturních atributů**[16] v korpusu Jerome zahrnuje v relevantních případech i příklad *hodnot*, kterých mohou nabývat[17].

autor	jméno autora ve formě *Příjmení, Jméno*
nazev	úplný název díla/opusu
nakladatel	nakladatelství nebo organizace, která dílo vydala
mistovyd	místo vydání
rokvyd	rok vydání toho konkrétního díla, které je zahrnuto v korpusu, nemusí jít o první vydání
prvnivyd	rok prvního vydání díla
isbnissn	identifikátor ISBN, příp. ISSN
preklad	překladatel/ka díla, nejedná-li se o původně české dílo
srclang	zdrojový jazyk v třípísmenné podobě, např. *ENG*, *GER*, *FRE*, *CZE* atd.
txtype_group	makroskupina textových typů: *beletrie*, *odborná* literatura, *publicistika*
txtype	blíže určený typ textu, např. *NOV* (romány), *ENC* (encyklopedie) atd.
genre	žánr (určený orientačně na základě pojednávaného tématu), např. *ART* (výtvarné umění), *LIN* (lingvistika), *MAT* (matematika) atd.
med	médium (způsob vydání textu), např. *B* (kniha), *J* (časopis) atd.
syn	obsahuje údaj o tom, z jakého korpusu řady SYN byl text vybrán (*2000*, *2005*, *2006pub*, *2010*)
status	informace o tom, zda je text překlad (*překlady*), nebo nepřeklad (*nepřeklady*)
autor_pohlavi	pohlaví autora, může nabývat hodnot *M* (muž), *Z* (žena), *KOL* (kolektiv min. 2 známých autorů bez ohledu na pohlaví), *Y* (kolektiv autorů, jejichž počet ani jména nemáme k dispozici)
preklad_pohlavi	pohlaví překladatele (*M*, *Z*, *KOL*)
sub_balance	informace o tom, zda je daný text zahrnut do jazykově vyváženého subkorpusu, pokud ano, nabývá atribut hodnot *beletrie* a *odborná*, pokud ne, zůstává pole prázdné

Všechny strukturní atributy lze využít ke specifikaci vyhledávání, příp. k vytvoření subkorpusu podle požadavků uživatele. Je tak například možné vyhledávat

[16] Úplná podoba strukturních atributů vypadá takto: opus.autor, opus.nazev atd.

[17] Seznamy a vysvětlení zkratek u atributů, jako je txtype, genre či srclang, jsou zveřejněny zde `http://wiki.korpus.cz/doku.php/seznamy:index`.

pouze v subkorpusu beletristických textů, textů přeložených z němčiny nebo naopak pouze v jazykově vyváženém subkorpusu. Lze si také pro specializovaný výzkum vytvořit například subkorpus mužských autorů přeložených překladatelkami a podobně.

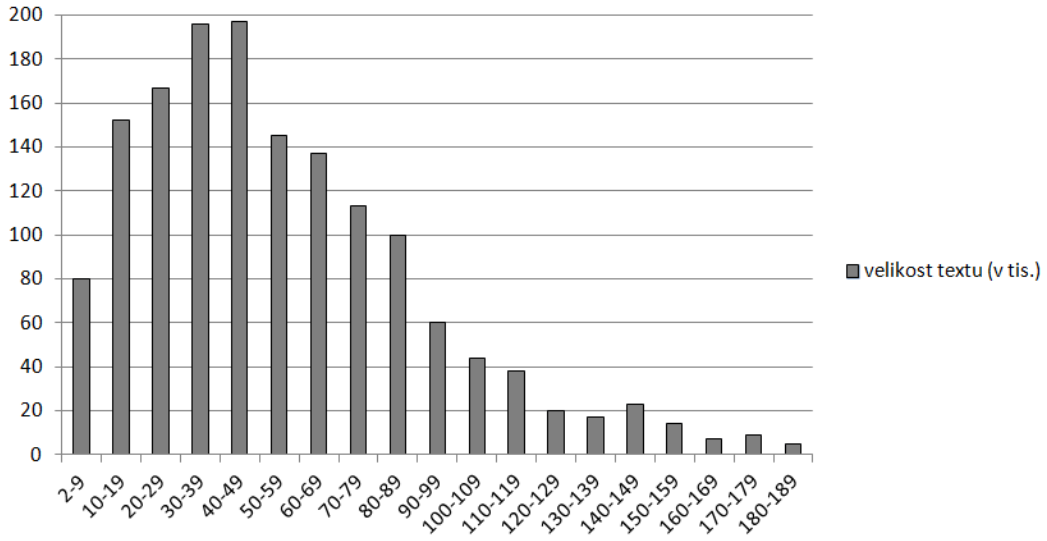
Velikost

Jak bylo zdůvodněno výše, velikost korpusu Jerome byla určujícím kritériem při jeho sestavování. Nakonec se při dodržení ostatních relevantních kritérií podařilo dosáhnout velikosti **85 milionů pozic**. Velikost korpusů bývá tradičně udávána ve slovech, tedy řetězcích alfabetických znaků (bez číslic a interpunkce). Vzhledem k tomu, že i interpunkce a její použití může být relevantním faktorem při zkoumání překladového jazyka (viz 4.1.1), uvádím údaje o velikosti v pozicích (neboli tokenech), tedy všech jednotkách oddělených v textu mezerami (vč. interpunkce a číslic). Tabulka 3.4 přehledně shrnuje velikost srovnatelných částí korpusu.

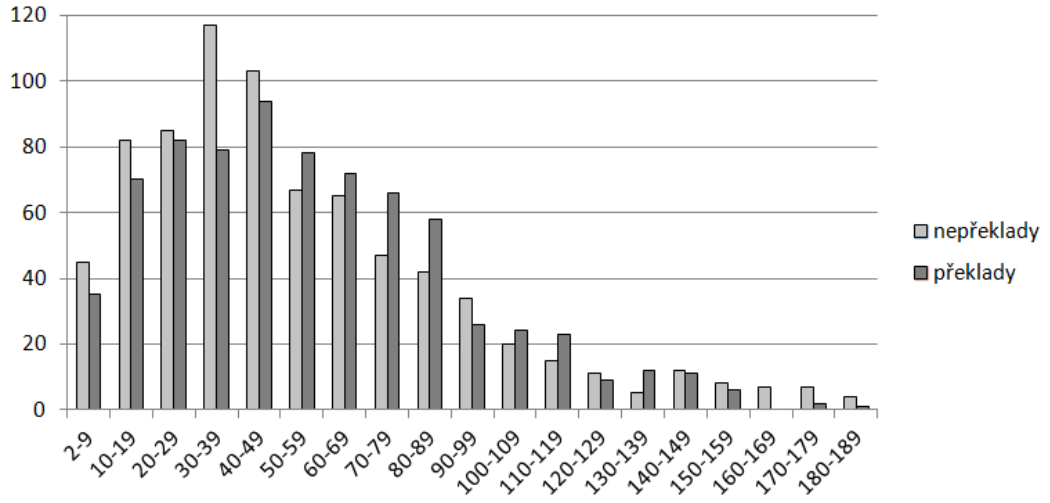
JEROME	nepřeklady	překlady	celkem
beletrie	26 551 540	26 617 523	53 169 063
odborná	15 949 930	15 946 319	31 896 249
celkem	42 501 470	42 563 842	**85 065 312**

Tabulka 3.4: Počet pozic v korpusu Jerome

Co se týče **velikosti jednotlivých textů**, cílem bylo, aby škála zahrnovala nejrůznější typy, od krátkých textů až po rozsáhlejší. Do korpusu byly proto vybrány texty v rozmezí délky 2 000 – 190 000 pozic. Následující graf 3.5 shrnuje počet děl v korpusu Jerome podle velikosti (v kategoriích po 10 tis. pozic).



Obrázek 3.5: Počet děl v korpusu Jerome podle velikosti



Obrázek 3.6: Počet překladů a nepřekladů v korpusu Jerome podle velikosti

Graf 3.6 zobrazuje srovnání délky textů v překladové a nepřekladové části korpusu. Bohužel nebylo možné docílit stejného počtu textů v dílčích kategoriích, ale v celkovém souhrnu jsou obě části korpusu (překladová i nepřekladová) srovnatelné. Jak je patrné z grafu, v překladové i nepřekladové části panuje u délky textů obdobný trend. Velikost textu však může hrát určující roli u některých statistických testů, proto je vždy třeba rozdíl mezi texty zohlednit (podrobněji o této problematice pojednává kapitola 4.2).

Heterogenita autorů a překladatelů

Po uplatnění pravidla tří (viz 3.2.1) a vyfiltrování relevantních textů obsahuje korpus Jerome texty 1 244 různých autorů nebo autorských kolektivů. V překladové části korpusu pak najdeme texty 607 překladatelů nebo překladatelských kolektivů. Do korpusu byla také ručně dodána anotace informující o pohlaví autora nebo překladatele, ovšem tento faktor nebyl při výběru textů zohledněn s ohledem na primární cíl při sestavování korpusu – jeho velikost. Korpus tedy není vyvážen podle pohlaví autora či překladatele, ovšem je možné texty podle tohoto atributu (autor_sex a preklad_sex) filtrovat. Atributy mohou nabývat následujících hodnot:

M	muž
Z	žena
KOL	kolektiv dvou a více autorů či překladatelů bez ohledu na pohlaví, u nichž známe jména
Y	kolektiv autorů, jejichž počet ani jména neznáme

Původním záměrem při tvorbě korpusu sice bylo, aby se v něm nevyskytovaly texty, u nichž neznáme jména autorů, ale v případě odborné literatury (především u té vydávané časopisecky) je velmi obtížné dohledat autora textu. Opět bylo nutné

volit, prioritu dostala pestrost žánrů v odborné literatuře a neznámý autorský kolektiv byl v menší míře ponechán. Shrnutí kategorie pohlaví ukazují následující tabulky 3.7 a 3.8.

AUTOR	beletrie N	beletrie P	odborná N	odborná P	celkem
M	286	313	167	183	949
Z	76	119	55	67	317
KOL	33	12	74	41	160
Y	0	0	87	13	100

Tabulka 3.7: Počet autorů podle pohlaví v celém korpusu Jerome

PŘEKLADATEL	beletrie P	odborná P	celkem
M	168	133	301
Z	273	138	411
KOL	33	33	66
Y	0	0	0

Tabulka 3.8: Počet překladatelů podle pohlaví v celém korpusu Jerome

Jak vyplývá z tabulek, v korpusu Jerome tak najdeme výrazně více textů napsaných muži (949 oproti 317), kdežto u překladů je situace opačná – více překladových textů v korpusu je dílem překladatelky (411) oproti překladateli (301), přičemž rozdíl je nejvíce patrný v překladech beletrie (273 oproti 168). U překladů odborné literatury jsou rovnoměrně zastoupeny překladatelky i překladatelé.

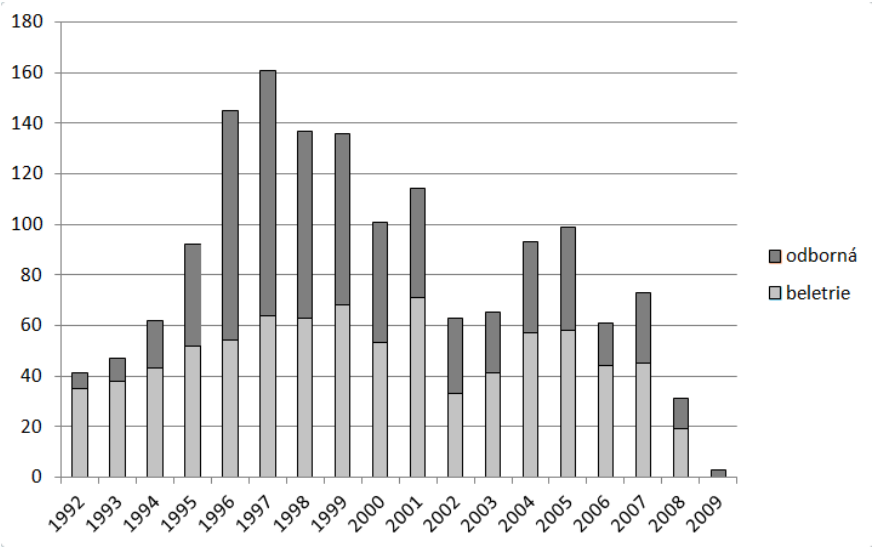
Doba vydání textů

Graf 3.9 ukazuje počet textů vydaných v jednotlivých letech. Jak je patrné, převažují texty z konce devadesátých let, cílem však nebylo vytvořit monitorovací korpus se stejným počtem textů za každý rok a sledovat vývoj rysů překladového jazyka rok za rokem, proto uvedené složení vyhovuje výzkumným požadavkům.

Textové typy a žánry

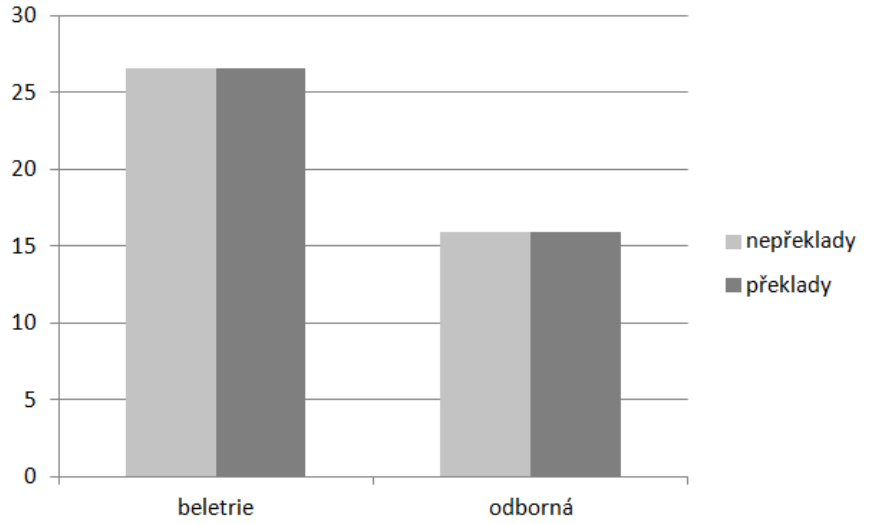
Při tvorbě srovnatelného korpusu je nutné volit texty obdobného textového typu nebo žánru[18]. Texty zařazované do databáze ČNK lze rozdělit podle obou kategorií. Ambice zahrnout do obou částí korpusu srovnatelný počet textů stejného

[18]Použití termínů *textový typ* a *žánr* se u různých lingvistů a v různých tradicích liší. V této práci jsou používány ve shodě s pojetím ČNK, kde textový typ zahrnuje tři hlavní velké kategorie: beletrii, publicistiku a odbornou literaturu (vč. dalšího dělení na romány, povídky, encyklopedie atd.), kdežto žánr odkazuje k tématu textu, oboru či disciplíně (např. matematika, biologie, právo atd.).



Obrázek 3.9: Roky vydání textů (v počtu děl)

žánru v odborné literatuře se ukázala být nereálná, dodržet se ji však podařilo u menšího vyváženého subkorpusu (viz 3.2.3). Požadavek srovnatelnosti byl u korpusu Jerome naplněn na nejvyšší úrovni textového typu, tedy u skupin **beletrie** a **odborná literatura**, které obsahují takřka totožný počet tokenů, viz graf 3.10. Odborných textů, které splnily kritéria složení korpusu, je méně než beletristických, ovšem překladová a nepřekladová část je v rámci textového typu srovnatelná, což byl klíčový požadavek.

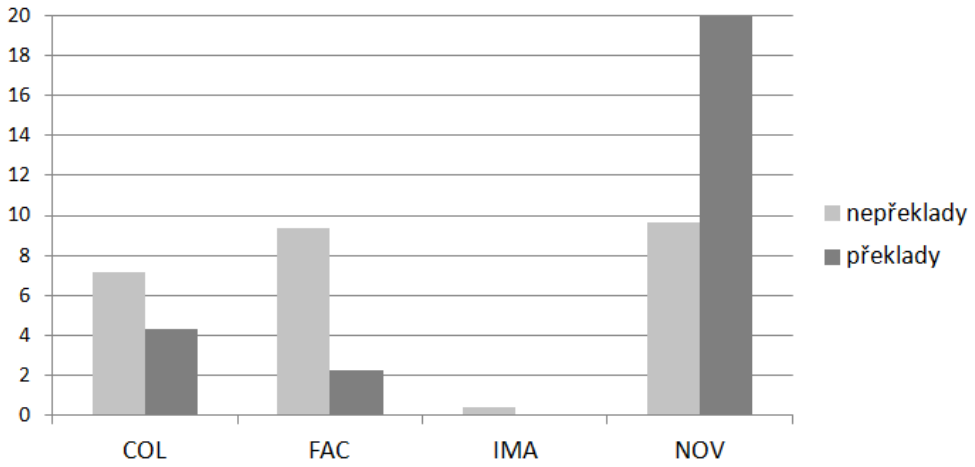


Obrázek 3.10: Textové typy v korpusu Jerome (v mil. pozic)

V rámci **beletrie** rozlišujeme v ČNK následující textové typy, jež jsou zahrnuty do korpusu Jerome:

COL soubor povídek, jednotlivá povídka
FAC literatura faktu (např. memoáry)
IMA jiné imaginativní texty
NOV román

Jak ukazuje graf 3.11, v překladové části převažují romány (NOV), zatímco v nepřekladové je poměr textových typů COL, FAC a NOV vyrovnanější. Texty z kategorie IMA jsou v obou souborech zastoupeny pouze okrajově. Problematický je nepoměr ve výskytu **textového typu FAC** (125 textů v nepřekladové beletrii a 40 v překladové); ačkoli je zde literatura faktu řazena do beletrie, fakticky stojí spíše na pomezí mezi populárně-naučnou a beletristickou literaturou, což se samozřejmě odráží v obsahové i formální charakteristice textů (viz 4.1). Srovnání překladové a nepřekladové části korpusu Jerome tak může probíhat pouze na nejvyšší rovině textového typu, v tomto případě na rovině beletrie zahrnující všechny zmíněné typy, přičemž při interpretaci dat je nezbytné brát zmíněný rozdíl na nižších úrovních v potaz.

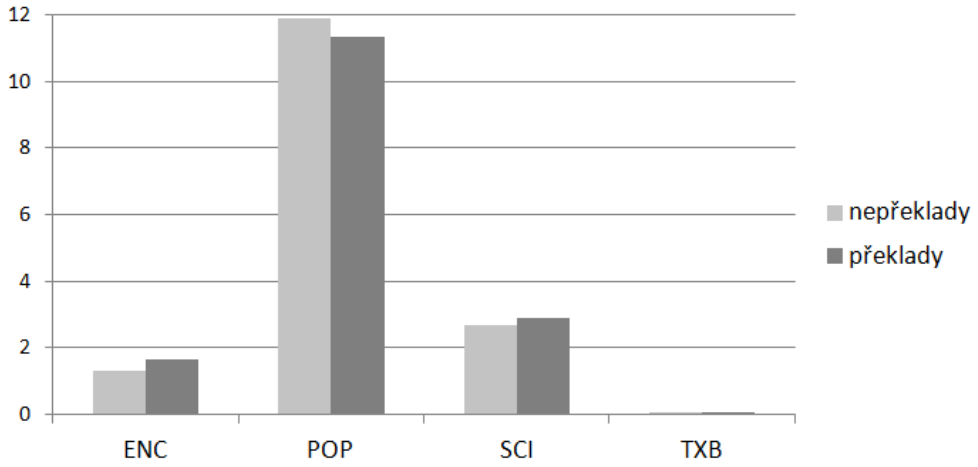


Obrázek 3.11: Textové typy v rámci beletrie (v mil. pozic)

V **odborné literatuře** můžeme dále vydělit následující textové typy:

ENC abecedně, systematicky a jinak uspořádaná díla
POP populárněnaučná literatura, též profesní a zájmové časopisy
SCI vědeckonaučná literatura
TXB učebnice

Z grafu 3.12 vyplývá, že v odborné literatuře je počet textů v překladové části a nepřekladové části z hlediska textových typů lépe srovnatelný, ovšem s ohledem na situaci v beletrii bude ve výzkumu zohledněna pouze nejvyšší rovina, tedy odborná literatura.



Obrázek 3.12: Textové typy v rámci odborné literatury (v mil. pozic)

Počet zdrojových jazyků

Ve shodě se všeobecným trendem posledních let, kdy se v českém prostředí vydávají především překlady z angličtiny, převažuje tento jazyk výrazně i v korpusu Jerome. Složení korpusu bylo motivováno snahou přiblížit se skutečné situaci překladové literatury u nás (viz 3.2.1), nikoli primárně sestavit korpus pro výzkum překladových univerzálií (kde by měly být všechny jazyky zastoupeny rovnoměrně), viz 3.2.3.

Následující tabulky 3.13 a 3.14 uvádí v úplnosti počty pozic a textů přeložených z uvedených jazyků. Podle atributu **srclang** lze rovněž texty filtrovat a ověřit tak možný vliv příslušného zdrojového jazyka.

JAZYK	počet pozic	počet textů	*JAZYK*	počet pozic	počet textů
ENG	18 274 340	283	*FIN*	182 722	3
FRE	2 211 599	45	*POR*	128 687	2
GER	2 161 026	48	*ICE*	125 594	1
RUS	729 066	13	*SLK*	109 237	2
POL	564 995	6	*HEB*	103 399	2
SPA	333 009	10	*HUN*	98 970	2
DAN	321 388	4	*NOR*	74 565	2
SWE	313 474	5	*GRN*	74 179	1
ITA	309 627	6	*SLV*	40 211	1
JAP	237 073	4	*SER*	22 867	1
DUT	201 495	3	-	-	-

Tabulka 3.13: Počet zdrojových jazyků v překladové beletrii

JAZYK	počet pozic	počet textů	*JAZYK*	počet pozic	počet textů
ENG	8 748 715	154	*LAT*	87 904	2
GER	3 999 797	90	*ROM*	86 695	1
FRE	1 338 413	23	*SPA*	74 106	1
POL	640 664	11	*SWE*	68 204	1
RUS	269 247	4	*HUN*	57 215	1
ITA	231 986	9	*GRA*	41 819	1
SLK	163 887	2	*SER*	29 824	2
MIX	88 215	1	*ARA*	19 628	1

Tabulka 3.14: Počet zdrojových jazyků v překladové odborné literatuře

Otázka kvality překladu u zařazených textů

Problém kvality překladů zahrnutých v korpusu úzce souvisí s otázkou **reprezentativnosti** korpusů obecně. Při sestavování jednojazyčných korpusů, jež si kladou za cíl reflektovat skutečnou situaci jazyka v určité jeho fázi či podobě (např. synchronní korpusy psaného jazyka), patří mezi klíčová kritéria pro zařazení konkrétního textu do korpusu například doba vzniku textu, příslušnost k určitému žádoucímu textovému typu či žánru nebo autor (ve shodě se snahou o co největší heterogenitu textů v korpusu) a v neposlední řadě také formální charakteristiky textu (délka, poměr textu a obrázků, tabulek a dalších netextových prvků, které se do korpusu nezařazují).

Jak uvádí Cvrček a Kováříková (2011: 121), „korpus principiálně není budován selektivně, rozhodnutí o zařazení textu není subjektivní (možná subjektivnost je potlačena množstvím textů)“. Z toho vyplývá, že při výběru textů do korpusu nehraje roli jejich subjektivně hodnocená kvalita (ať už ji chápeme z hlediska obsahu/tématu nebo formy/jazyka). Ponechme nyní stranou korpusy specializované (např. pro výuku jazyka na školách nebo výcvik překladatelů), jejichž charakter je do jisté míry preskriptivní: jejich složení má odrážet určitý žádoucí standard a kritéria pro výběr textů jsou tudíž odlišná. U korpusů zaměřených obecně, na jazyk, jak ve skutečnosti vypadá, k žádnému filtrování textů podle kvality nedochází.

U překladových textů vstupuje do hry navíc i překladatel a jeho volby a řešení, jež mohou vést k více či méně zdařilému překladu. Otázka, jak bychom definovali dobrý, zdařilý či přiměřený překlad, nás vrací zpět k dějinám teorie překladu a především ke kritice překladu. Není pochyb o tom, že translatologové disponují prostředky a metodami, jak hodnotit kvalitu překladu, jakkoli se jejich kritéria mohou lišit (viz například koncept ekvivalence podle Nidy, teorie skoposu a další, viz 2.1). V naprosté většině případů však **kritika překladu** vyžaduje pečlivou kvalitativní analýzu překladu v porovnání s originálem (zdrojovým textem). Vzhledem k objemu dat v korpusu Jerome a k zaměření na kvantitativní výzkum je však taková analýza všech děl nereálná, a to nejen kvůli časové náročnosti, ale také kvůli tomu, že jednojazyčný srovnatelný korpus nezahrnuje zdrojové texty. Navíc pokud bychom chtěli filtrovat překladové texty podle kvality, museli bychom totéž provést i s texty nepřekladovými (viz výše).

Zde je nutné dodat, že otázka kvality překladových děl v korpusu představuje výzvu nejen v případě tohoto konkrétního korpusu, ale v korpusové translatologii obecně.[19] Zabývají-li se tvůrci korpusů vůbec otázkou kvality překladu, tradičně rozlišují pouze mezi publikovanými překlady (*published translations corpora*) a překlady studentskými (*translation learner corpora*), fundovanost překladatele zpravidla nijak neověřují.[20] V dnešní době, kdy se texty do korpusu běžně získávají i z internetu, se navíc může objevit i mnoho knižně nepublikovaných neprofesionálních překladů, strojových překladů apod., které s sebou přinášejí další problémy. Korpus Jerome však žádné takové texty nezahrnuje, proto tuto otázku ponechávám stranou.

Fakt, že při sestavování korpusu Jerome nebyla kvalita překladu (ani textu jako takového u nepřekladové části) zohledněna, nemá v žádném případě sugerovat, že mezi překlady nejsou rozdíly, ba naopak. Tím, že do korpusu jsou zahrnuta nejrůznější díla z databáze ČNK, se otevírá prostor pro to, aby byly zastoupeny jak velmi kvalitní, průměrné, tak i ty méně kvalitní překlady, tedy **široké spektrum textů**, s nímž se běžný čtenář překladů reálně setká. Základním předpokladem zde však zůstává, že samotná skutečnost, že překlady byly publikovány knižně pod hlavičkou fungujícího nakladatelství a musely tedy projít alespoň minimální redakční kontrolou, zaručuje alespoň určitou úroveň jejich přijatelnosti pro čtenáře.

Z pohledu statistiky pak můžeme tentýž postoj shrnout odkazem na tzv. normální (Gaussovo) rozdělení pravděpodobnosti, kdy očekáváme, že v populaci bude nejvíce zastoupena střední hodnota jevu (tj. průměrné, všeobecně přijatelné překlady) a jevy okrajové (tedy výjimečně dobré překlady a výjimečně špatné překlady) budou tvořit pouze malou část. Velikost korpusu a heterogenita děl (zaručená pravidlem tří) pak představují další záruku, že případné nekvalitní překlady nebudou v tomto objemu dat zásadním způsobem zkreslovat výsledek.

3.2.3 Vyvážený subkorpus

Jak bylo zdůrazněno výše, hlavními přednostmi korpusu Jerome jsou jeho velikost a heterogenita textů, autorů i překladatelů. Vzhledem k těmto prioritám však nebylo možné korpus vyvážit z hlediska dalších kritérií, jako je srovnatelný počet textů ze všech zahrnutých zdrojových jazyků (u překladů) nebo srovnatelný počet žánrů v odborné literatuře. Dodržení těchto kritérií však může ve výzkumu překladových rysů hrát klíčovou roli, neboť eliminuje vliv různých proměnných na charakteristiku překladového jazyka (např. interferenci z převažujícího zdrojového jazyka nebo vliv převažujícího žánru):

> „The isolation of translation universals and norms may be demanding in terms of corpus resources, since several translation and reference subcorpora are needed in order to disentangle source language, genre-related and diachronic variables.“ Zanettin (2013: 30)

[19] Za aktuální informace a cenné poznámky k hodnocení kvality překladu v korpusech vděčím translatologu Federicu Zanettinovi (osobní komunikace).

[20] Ojedinělým příkladem budiž Johansson (2004), který pro účely své studie provedl cílený výběr kvalitních překladatelů, ale v tomto případě se nejedná o sestavování rozsáhlého korpusu, nýbrž o vytvoření poměrně malého vzorku za specifickým účelem.

Z tohoto důvodu byl v rámci korpusu Jerome (ručním výběrem z jeho textů) vytvořen vyvážený srovnatelný subkorpus, který splňuje výše uvedené požadavky. Jeho velikost je nutně řádově menší, viz tabulka 3.15. Do subkorpusu byly vybrány **texty o délce** 20 000 – 160 000 tokenů, přičemž pro překladovou část byla stanovena kvóta v rozmezí 110 000 – 160 000 tokenů pro každý zastoupený jazyk (kde to bylo možné, bylo vybráno více kratších textů, a nikoli jeden dlouhý). Subkorpus slouží především k přesnější interpretaci výsledků, které byly zjištěny na korpusu Jerome, ale může být využit i k samostatnému výzkumu překladových rysů (univerzálií), aniž by badatel riskoval vliv interference převažující angličtiny.

SUBKORPUS	nepřeklady	překlady	celkem
beletrie	1 768 079	1 765 433	3 533 512
odborná	779 288	774 610	1 553 898
celkem	2 547 367	2 540 043	**5 087 410**

Tabulka 3.15: Počet pozic ve vyváženém subkorpusu

Subkorpus obsahuje opět texty nepřekladové i překladové s takřka stejným počtem pozic a tentokráte i se stejným počtem textů (beletrie 33, odborná literatura 16 v každé části). V beletristické části jsou zahrnuty texty přeložené ze čtrnácti jazyků, v odborné pouze ze šesti, neboť v dalších jazycích nebylo k dispozici tolik textů, jejichž objem by byl min. 110 000 tokenů. Zahrnuty však byly v obou případech jazyky z typologicky odlišných skupin. Při srovnávání obou textových typů je nutné tuto skutečnost zohlednit. Tabulky 3.16 a 3.17 shrnují **srovnatelné počty jazyků** v beletrii i odborné literatuře obsažené v subkorpusu.

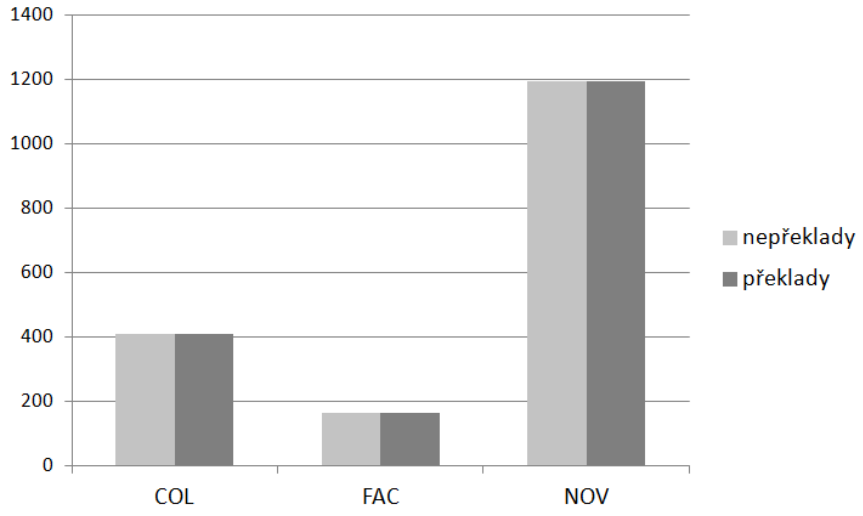
JAZYK	počet pozic	počet textů	*JAZYK*	počet pozic	počet textů
DAN	130 296	2	*ITA*	132 926	3
DUT	128 176	2	*JAP*	124 090	2
ENG	140 094	3	*POL*	124 129	2
FIN	112 765	1	*POR*	128 687	2
FRE	124 914	3	*RUS*	127 251	3
GER	128 435	3	*SPA*	119 510	3
ICE	125 594	1	*SWE*	118 566	3

Tabulka 3.16: Počet zdrojových jazyků v překladové beletrii subkorpusu

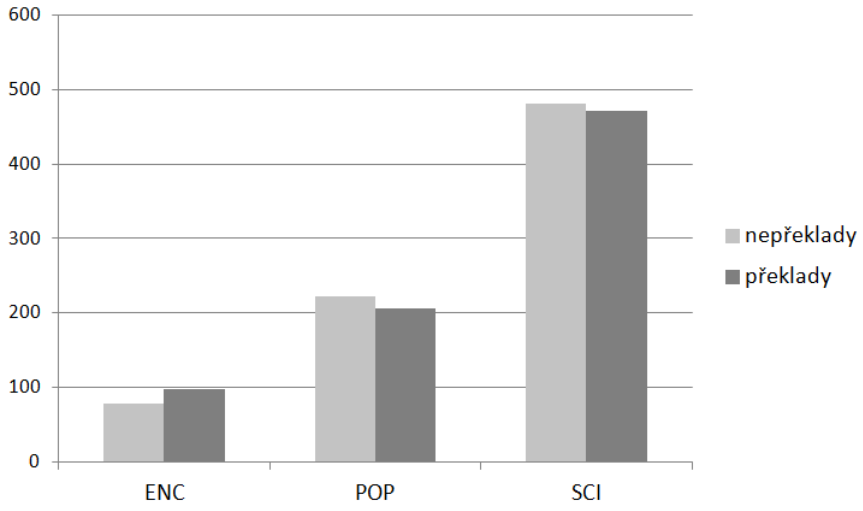
JAZYK	počet pozic	počet textů	*JAZYK*	počet pozic	počet textů
ENG	113 149	3	*ITA*	123 496	3
FRE	126 298	3	*POL*	155 261	3
GER	134 504	3	*RUS*	121 902	1

Tabulka 3.17: Počet zdrojových jazyků v překladové odborné literatuře subkorpusu

Podíváme-li se na **zastoupení textových typů a žánrů** v subkorpusu, obě části jsou srovnatelné i na nižší úrovni, než je beletrie a odborná literatura. V beletrii jsou rovnoměrně zahrnuty povídky (COL), romány (NOV) a literatura faktu (FAC), v odborné literatuře pak populárněnaučné texty (POP), vědeckonaučné texty (SCI) a abecedně řazená díla / encyklopedie (ENC), jak ukazují grafy 3.18 a 3.19.



Obrázek 3.18: Textové typy zahrnuté v subkorpusu – beletrie (v tis. pozic)



Obrázek 3.19: Textové typy zahrnuté v subkorpusu – odborná literatura (v tis. pozic)

Aby nedošlo ke zkreslení vlivem jednoho žánru, v odborné literatuře obsahují obě části subkorpusu ve stejném počtu následující žánry, vybrané pragmaticky z dostupných žánrů s ohledem na zdrojový jazyk a velikost (v závorce počet textů):

AGR zemědělství, lesnictví (1)
ARS jiný text z oblasti věd o umění (1)
BIO obecná biologie (1)
BOT botanika (1)
ENV ekologie (1)
ETH etnografie (1)
EXC okultní vědy, magie (1)
HOU domácí hospodářství (1)
IND průmysl, technika (1)
PHI filozofie (1)
POL politologie (1)
PSY psychologie (1)
REL náboženství, teologie (2)
SOC sociologie (1)
THE divadlo, balet (1)

Shrneme-li tedy hlavní motivaci pro vytvoření takto vyváženého subkorpusu, můžeme říci, že na rozdíl od celého korpusu Jerome, který slouží pro výzkum charakteristických rysů překladové češtiny, jak se s ní setkává průměrný čtenář, vyvážený subkorpus by měl svým složením poskytnout datovou základnu pro výzkum a deskripci překladových univerzálií vyskytujících se v textech přeložených z různých jazyků, tedy bez nežádoucího vlivu interference, a umožnit tak lepší interpretaci výsledků.

3.3 Metodologické zásady

Jak již bylo řečeno v kapitole 2.3, korpusová translatologie se od svého vzniku v polovině devadesátých let 20. století zabývá především zkoumáním a hledáním překladových univerzálií a obecně typických rysů překladového jazyka. Za přibližně dvacet let výzkumu vzniklo mnoho studií na mnoha jazycích, které se pomocí nejnovějších poznatků a nástrojů snažily potvrdit či korigovat výchozí hypotézy Mony Bakerové. Na počátku nového tisíciletí se však začaly množit **kritické hlasy**, že tento bakerovský přístup není podložen dostatečně propracovanou metodologií. Výtky by se daly shrnout do dvou hlavních stanovisek:

1. Původní hypotézy jsou formulovány velmi vágně a je těžké vysledovat, z čeho vycházejí a na čem jsou založeny, což samozřejmě velmi snižuje relevanci výsledků a ztěžuje jejich interpretaci.

2. Mnoho korpusových studií v translatologii zcela zanedbalo vliv dalších faktorů, jako je zdrojový jazyk nebo žánr.

   (De Sutter et al. 2012: 138)

Translatologové z Ghentské university v čele s Gertem De Sutterem (De Sutter et al. 2012: 137) proto formulovali **několik metodologických požadavků**, jež by měly studie z korpusové translatologie splňovat:

1. to provide a meticulous overview of the corpus materials used and of the exact procedures for selecting, annotating and sifting the data;

2. to comment on any specific problems encountered during data selection and annotation, including explicit and motivated statements as to the solutions being adopted;

3. to include elaborate testing for statistical significance as a complement of, not in opposition to, thorough qualitative analysis.

Ve shodě s těmito zásadami, jež by v obecné rovině měly platit pro veškeré korpusové lingvistické studie, byl v kapitole 3.2 detailně představen korpus Jerome, včetně podrobného zdůvodnění výběru textů do korpusu. S požadavkem přesnější formulace a odůvodněné operacionalizace hypotéz o překladových univerzáliích/rysech se vypořádávají příslušné části kapitoly 4 věnované jednotlivým zkoumaným rysům. Výhodám a nevýhodám kvantitativního a kvalitativního výzkumu se pak věnuje následující část této kapitoly.

3.3.1 Kvantitativní a kvalitativní přístup

Na první pohled by se mohlo zdát, že se kvantita a kvalita, zvláště při výběru dat, navzájem vylučují – buď si badatel zvolí jedno, nebo druhé:

> „For practical reasons corpus compilers may have to choose between focusing on quantity or on quality, often one at the expense of the other.“
>
> (Zanettin 2013: 31)

Jak bylo patrné i z popisu korpusu Jerome, snaha vytvořit a využít velký reprezentativní korpus tak s sebou nutně nese ústupky na nižších rovinách (vyváženost podle žánrů, zdrojového jazyka, pohlaví apod.). Ovšem tyto nedostatky lze kompenzovat jak vytvořením vyváženého subkorpusu, tak i doplňujícími případovými studiemi na omezených, pečlivě vybraných datech. V korpusové translatologii je tak více než kde jinde uplatňován přístup, který se snaží oba způsoby zkoumání, kvantitativní a kvalitativní, kombinovat:

> „Quantitative and qualitative approaches are radically intertwined in corpus-based translation studies, and they are not mutually exclusive. On the one hand larger corpora which typically display little annotation can be enriched with further layers of annotation. [...] On the other hand, small-scale qualitative studies based on intensive annotation are needed to confirm the findings from large-scale quantitative studies.“
>
> (Zanettin 2013: 31)

Jak tato kombinace vypadá v praxi? Je pochopitelné, že vzhledem k čím dál naléhavějšímu požadavku spolehlivé a velké datové základny není dost dobře možné, aby byl takový výzkum v translatologii prováděn kvalitativně – tedy bez využití statistických metod, ať už základních (frekvence, distribuce), nebo sofistikovanějších (faktorová analýza, testování statistické signifikance apod.)[21]. **Kvantitativní přístup** se v tomto případě jeví jako ideální, neboť umožňuje zpracovávat rozsáhlá data a vyvozovat z nich relativně obecné závěry o příslušném souboru textů. Výhody kvantitativního výzkumu (vyplývající ze spolehlivosti a autentičnosti velkého objemu dat, eliminace subjektivního faktoru při výzkumu a možnosti generalizace) se projevují především na rovině deskripce (k čemu v jazyce dochází a v jaké míře) – na rovině interpretace (proč k tomu v tom kterém případě dochází) mohou mít takto obecná zjištění omezenou platnost a mohou vyžadovat doplnění a vysvětlení pomocí cílených a kvalitativně zaměřených sond.

Kromě dílčích kvalitativních studií se **kvalitativní pohled** v korpusové translatologii může odrážet v rozšířené anotaci, jak uvádí Zanettin, nejen na úrovni textů, ale na úrovni větných segmentů či slov: kromě standardně využívaného morfologického značkování a lemmatizace je možné využít i syntaktickou anotaci nebo značkování ryze evaluativní, jež je nutné provádět manuálně (např. značení chyb v překladu u korpusů studentů translatologie nebo kategorizace explicitačních prostředků u korpusu CROCO[22]).

Je však třeba mít na zřeteli, že kvalitativní přístup bývá ve skutečnosti uplatňován i u kvantitativního výzkumu, a to všude tam, kde je třeba pragmaticky nastavit hranici relevance výsledků (tzv. *cut-off point*) nebo manuálně třídit výsledky, např. u konkordančních řádků nebo u jiných kvantitativně zjištěných dat (např. v případě kategorizace a třídění kolokací, frekvenčních seznamů či n-gramů).

Kvantitativní a kvalitativní přístup tak nestojí nutně v protikladu, nýbrž se mohou vzájemně doplňovat. Těžiště této práce spočívá v kvantitativních analýzách na rozsáhlých datech, které mají za cíl zmapovat současnou překladovou češtinu v co největší šíři, její obecné vlastnosti a charakteristiky. Tam, kde analýzy ukazují potenciálně zajímavá data, je pak na nižší rovině doplňují dílčí, úzce zaměřené sondy, jejichž cílem je přinést konkrétní příklady a možná vysvětlení zkoumaných jevů.

3.3.2 Výchozí hypotéza

Zatímco při zkoumání S-univerzálií na paralelním korpusu se badatel může soustředit jak na shody překladu s originálem, tak na rozdíly mezi nimi, v případě T-univerzálií (na datech z *jednojazyčného* srovnatelného korpusu) jsou v centru pozornosti odlišnosti. Podobnost překladových textů s nepřekladovými v rámci jednoho jazyka totiž

[21]Zde vycházím z definice kvantitativní analýzy jakožto výzkumu založeného na statistických metodách (McEnery & Hardie 2012: 249).

[22]Tento obousměrný paralelní (reciproční) anglicko-německý korpus byl vytvořen přímo za účelem výzkumu překladových rysů a byl manuálně označkován z hlediska prostředků vyjadřujících explicitaci. Vzhledem k časové náročnosti takové anotace čítá korpus pouhý jeden milion slov. Více informací na stránkách projektu: `http://fr46.uni-saarland.de/croco/deliverable_en.html`.

může značit nejen přiměřenost překladu, ale i pouhou přirozenost využitého jazyka, nic však už nevypovídá o společných rysech překladů coby samostatné skupiny textů. Z toho důvodu je výzkum překladové češtiny zaměřen právě na to, čím se tato odlišuje od textů nepřekladových.

Zde je však nutné připomenout základní premisu, totiž že žádné dva texty nejsou zcela totožné a všechny se od sebe nějakým způsobem odlišují, samotná **odlišnost překladového a nepřekladového textu** tedy ještě není důkazem existence specificky překladových rysů. Abychom mohli interpretovat zjištěné odlišnosti jako vliv procesu překladu, je třeba mít jednak k dispozici velké množství textů z obou srovnávaných souborů, jednak identifikovat dostatečné množství rysů, jejichž větší či menší zastoupení v daných textech může pomoci seskupit tyto texty do skupin, jež jsou si podobné. Teprve v případě, že se tyto skupiny budou překrývat se skupinou překladů nebo nepřekladů, pak můžeme usuzovat, že daná kombinace rysů je skutečně typická pro texty, které ne/prošly procesem překladu (viz kapitola 4).

Budeme-li tedy předpokládat, že překladové texty vykazují jiné vlastnosti než nepřekladové, nulová hypotéza[23] na nejvyšší rovině abstrakce bude znít takto:

H_0: Překladová a nepřekladová čeština se neliší.

Tato hypotéza pak bude konkrétním způsobem testována na vybraných rysech, a pakliže se nepotvrdí (testy a sondy odhalí rozdíly, které budou společné pouze překladovým textům), bude platit alternativní hypotéza:

H_1: Překladová čeština se od nepřekladové odlišuje.

Tato výchozí hypotéza (v podobě nulové a alternativní hypotézy), jež je vzhledem ke své obecnosti v této podobě netestovatelná, tvoří **základ dílčích hypotéz**, které se vztahují k vybraným deskriptivním rysům překladových textů (viz Zanettinovo rozdělení na s. 37). Těm se podrobně věnuje následující kapitola 4.

[23] Nulová hypotéza je tvrzení, které obvykle deklaruje „žádný rozdíl“ mezi zkoumanými soubory dat (Hendl 2009: 182).

Kapitola 4

Rysy překladové češtiny

Tato kapitola představuje těžiště výzkumu: přináší analýzy vybraných rysů překladového jazyka na základě korpusu Jerome. Zkoumá-li badatel rozdíly mezi jazykem v překladech a jazykem nepřekladových textů, musí mít stále na zřeteli, že předmětem jeho výzkumu je stále **tentýž jazykový systém**, v tomto případě čeština. Na rozdíl od kontrastivních translatologických studií (např. při výzkumu S-univerzálií, viz s. 37) tak mezi možné vlivy nevstupuje ani typologická odlišnost zkoumaných jazyků (případně pouze v podobě interference, viz 2.3.3) ani jazykově specifické jevy; rozdíly tedy nemusí být tak zjevné. Nelze předpokládat, že čeština v překladech bude na první pohled jiná než jazyk původních česky psaných děl – výrazné odlišnosti můžeme očekávat spíše na nižších rovinách (např. nikoli ve slovní zásobě jako celku, ale v preferenci konkrétních kolokací nebo typů výrazů). Abychom to však mohli tvrdit s jistotou, je třeba pečlivě srovnat i kategorie na nejvyšších úrovních obecnosti.

Všechny části této kapitoly jsou proto strukturovány tak, že postupují od nejvyšší roviny abstrakce až po nejnižší ve shodě se Zanettinovým interpretačním rámcem (viz s. 37). Každá část (s výjimkou 4.1) je věnována konkrétnímu rysu překladového jazyka, přičemž **výběr zkoumaných rysů** se opírá o dosavadní translatologické studie o univerzáliích provedených na jiných jazycích, které jsou v příslušné části vždy odůvodněny a okomentovány. Jednotlivé rysy založené na T-univerzáliích (viz s. 37) zahrnují simplifikaci (4.2), levelling-out/konvergenci (4.3) a (ne)typické slovní kombinace (4.4) vycházející z výzkumu n-gramů (častých sekvencí o délce n-slov).

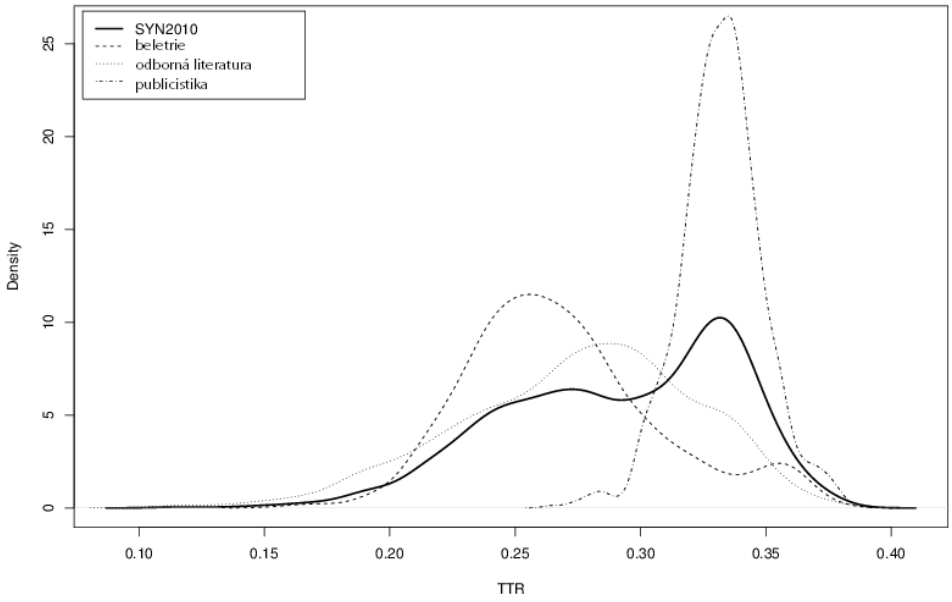
Úvodní část 4.1 má obecnější charakter; není zaměřena na jeden konkrétní překladový rys, ale snaží se popsat překladovou češtinu z hlediska frekvence a distribuce slovních druhů a jejich kombinací (tzv. POS-gramů[1]) a zároveň zohlednit rozdílnost textových typů (beletrie, odborná literatura). Tento přehled pak také slouží jako možný zdroj dat pro další analýzu v rámci jednotlivých překladových rysů. Cílem této kapitoly je také poukázat na limity často používaných statistických testů.

[1] POS = *part of speech*

4.1 Obecné frekvenční charakteristiky

Frekvence neboli četnost výskytu jevu je v korpusové lingvistice jedním z klíčových konceptů. Slouží nejen jako prvotní informace o užívání zkoumaného jevu, jež vypovídá o jeho zařazení mezi centrum nebo periferii jazyka (tak, jak je zachycen v reprezentativním korpusu), ale může být také ukazatelem jazykové změny v diachronní i synchronní perspektivě a cenným údajem při srovnávání více jevů. Frekvenční distribuce jevů v korpusu představuje zpravidla odrazový můstek pro další analýzu a může upozornit na rozdíly mezi zkoumanými soubory, v tomto případě mezi překladovými a nepřekladovými texty.

V analýzách je však třeba brát v potaz nejen předpokládanou odlišnost mezi překlady a nepřeklady, ale také **rozdíl mezi textovými typy** v korpusu Jerome – beletrií a odbornou literaturou, které vstupují do veškerých srovnávacích analýz jako další proměnné. Fakt, že je nezbytné zohlednit při výzkumu žánrovou[2] odlišnost, je v lingvistice v posledních letech velmi zdůrazňován (viz například Biber & Conrad 2009). O tom, že tyto rozdíly platí i v češtině, svědčí například graf 4.1, který ukazuje zcela rozdílné hodnoty poměru mezi typy a tokeny (TTR) ve třech textových typech korpusu SYN2010 (Cvrček & Chlumská v tisku). Veškerá srovnání v rámci korpusu Jerome jsou proto prováděna odděleně na každém textovém typu zvlášť.



Obrázek 4.1: Srovnání TTR v různých žánrech

[2]Výraz žánr je zde použit v širokém slova smyslu jako ekvivalent anglického *genre*; v terminologii české korpusové lingvistiky však odkazuje především na rozdíl mezi textovými typy.

4.1.1 Frekvenční distribuce slovních druhů

Slovní druhy, klasifikující slova do skupin podle jejich vlastností a významů, představují v jazyce vyšší rovinu obecnosti – ačkoli bývají tradičně popisovány v rámci morfologie, jejich vymezení souvisí především s lexikonem a má důsledky i pro syntax. S kategorií slovních druhů pracují i některé výzkumy o překladových univerzáliích, především ty, které zkoumají bohatost slovní zásoby pomocí statistických měr, jako je např. lexikální hustota (*lexical density*) porovnávající výskyt autosémantik a synsémantik v textu (viz 4.2).

Analyzovat distribuci slovních druhů ve velkém korpusu je možné jen tehdy, je-li označkován. Morfologické značkování v korpusech ČNK (a tedy i v korpusu Jerome), zahrnující i zařazení slova k příslušnému slovnímu druhu, je výsledkem automatického procesu, který zahrnuje morfologickou analýzu a desambiguaci. Úspěšnost morfologického značkování u korpusů ČNK se pohybuje přibližně okolo 95 % (nerozpoznané slovní druhy jsou v následujících tabulkách označeny jako „neurčeno“). Kromě slovního druhu lze podle morfologické značky identifikovat i interpunkci, kterou také zahrnuji do frekvenčního přehledu v tabulkách 4.2 a 4.3, protože může být jedním z ukazatelů vlivu překladu v textech (viz Rodríguez-Castro 2011).

Abychom mohli s jistotou říci, že případné rozdíly mezi soubory nejsou pouhým dílem náhody, tedy že jsou s ohledem na velikost vzorku statisticky významné, je třeba provést některý z testů **statistické signifikance** (*statistical significance*), jako je například test chí-kvadrát (viz např. Volín 2007: 124 nebo Cantos Goméz 2013: 75) nebo Mann-Whitneyův U test (viz Volín 2007: 151). Oba tyto testy jsou neparametrické a používají se pro nezávislé proměnné. V této práci jsou využity oba dva, proto zde uvádím jejich stručnou charakteristiku. Test chí-kvadrát slouží k testování rozdílů mezi četnostmi výskytu v určitých kategoriích a vychází z rozdílů mezi očekávanými a skutečně pozorovanými hodnotami. Mann-Whitneyův U test rovněž umožňuje porovnání dvou skupin případů a jeho výsledek naznačuje, zda jsou rozdíly mezi nimi náhodné, nebo statisticky významné, jinými slovy, zda testované veličiny patří pomyslně do jednoho souboru nebo do dvou odlišných.

Výsledky srovnání (viz tabulky 4.2 a 4.3) ukazují, že v rámci obou textových typů, beletrie i odborné literatury, se distribuce slovních druhů v překladech i nepřekladech liší, avšak nikterak výrazně; rozdíly se pohybují maximálně v řádu jednotek procent. Provedený test chí-kvadrát potvrdil, že všechny slovní druhy s výjimkou adverbií v beletrii vykazují v distribuci v obou testovaných souborech (překladu i nepřekladu) **statisticky významné rozdíly** (na hladině významnosti $p < 0{,}001$).

Co však můžeme z těchto výsledků vyvodit? Samotný fakt, že rozdíly jsou statisticky signifikantní, pouze značí, že máme dostatek dat na to, abychom mohli tvrdit, že z čistě matematického hlediska rozdíly mezi soubory existují; nevypovídá však už nic o tom, jak významná tato zjištění jsou z hlediska výzkumného či lingvistického. Navzdory tomu mnoho kvantitativně zaměřených studií nejen v korpusové translatologii končívá právě u výpočtu statistické signifikance, na jehož základě postuluje závěry o vědecké významnosti výsledků. Tento postup čelí především v posledních letech oprávněné kritice, některá odborná periodika (např. *Basic and Applied So-*

cial Psychology) dokonce odmítají přijmout příspěvek, jenž vychází pouze z těchto statistik.

Jak lze tedy posoudit, do jaké míry jsou výsledky z vědeckého hlediska významné? Test statistické signifikance může a měl by být pouze prvním krokem, který potvrdí, zda má smysl data dále testovat. Pokud jsou zjištění skutečně statisticky významná, je vhodné otestovat i tzv. **věcnou významnost**, jinými slovy jak relevantní je daný výsledek pro náš výzkum. Mezi míry, které testují tuto sílu účinku (*effect size*), patří například Cohenovo d, Glassovo delta nebo Haysovo omega (viz např. Soukup 2013).

V tomto případě byla použita obdobná míra DIN neboli *difference index* (Cvrček & Fidler, v tisku). DIN je modelována podle míry Dice a vychází z relativních frekvencí jevu ve zkoumaném souboru (A) a referenčním souboru (B). Počítá se následujícím způsobem:

$$DIN = 100 \times \frac{relFQ(A) - relFQ(B)}{relFQ(A) + relFQ(B)} \tag{4.1}$$

Zkoumaným souborem jsou v tomto případě překladové texty, referenčním nepřekladové texty (počítáno pro beletrii i odbornou literaturu zvlášť). DIN může dosahovat hodnot od -100 do 100, přičemž obecně platí, že:

- hodnota -100 znamená, že daný jev se ve zkoumaném souboru vůbec nevyskytuje, je pouze v referenčním korpusu,
- hodnota 0 znamená, že daný jev má zhruba stejnou relativní frekvenci ve zkoumaném souboru i v referenčním souboru (není tedy prominentní ani pro jeden ze souborů),
- hodnota 100 značí, že jev se vyskytuje pouze ve zkoumaném souboru (může se tedy jednat o velmi typický a prominentní jev),
- hodnoty v rozmezí absolutních hodnot 75–100 je možné považovat za obzvlášť zajímavé, např. v případě vyhodnocování klíčových slov v textu (tzv. *keywords*).

V praxi však dosahuje výrazně okrajových hodnot zpravidla jen malá část zkoumaných jevů; většinu lze se srovnatelnou frekvencí nalézt v obou zkoumaných souborech. Výsledky srovnání je třeba vždy hodnotit v kontextu – můžeme očekávat, že DIN bude dosahovat jiných (pravděpodobně vyšších) hodnot u jednotlivých lexémů a jiných hodnot (méně výrazných) u větších skupin, jako jsou například slovní druhy. Rozdílné výsledky však můžeme čekat i v rámci kategorie slovních druhů jako takové, kde je při interpretaci nezbytné zohlednit, že některé slovní druhy jsou malé a relativně uzavřené (jako např. zájmena), kdežto jiné představují rozsáhlou a diverzifikovanou třídu (např. substantiva nebo slovesa). V prvním případě, u synsémantické skupiny zájmen, můžeme očekávat, že slov, která se vyskytují jen v jednom ze souborů, bude velmi málo, kdežto v případě substantivních autosémantik bude výskyt slov do značné míry ovlivněn tématem textu a překryv tak může být menší (viz 4.1).

Podíváme-li se znovu na výsledky v 4.2 a 4.3 z hlediska míry DIN, zjistíme, že rozdíly, ač statisticky signifikantní, nedosahují příliš vysokých hodnot. Naprostá většina slovních druhů v beletrii i odborné literatuře se pohybuje v rozmezí -10 až +10, což znamená, že jejich výskyt je v obou souborech relativně srovnatelný a nejde tedy o zcela odlišné tendence. To potvrzuje také Pearsonův korelační koeficient, který je pro překladové i nepřekladové texty v obou textových typech vyšší než 0,98 (značí tedy velmi silnou korelaci mezi oběma soubory).

Dílčí rozdíly v případě zájmen (DIN = 10,67) v překladové odborné literatuře, číslovek v nepřekladové beletrii (DIN = -18,86) a citoslovcí (DIN = -14,56) v nepřekladové odborné literatuře můžeme do značné míry přičítat konkrétnímu složení korpusu. V případě číslovek, které vykazují nejvýraznější rozdíl, jde o nesporný vliv literatury faktu (FAC). V nepřekladové beletrii korpusu Jerome byl zaznamenán pětkrát větší výskyt nejfrekventovanějšího typu, což jsou číslovky psané číslicí (6 568,51 ipm oproti 1 337,92 ipm v překladech), přičemž bližší analýza potvrdila, že téměř 50 % všech těchto výskytů spadá do textového typu FAC, který má v nepřekladové beletrii větší zastoupení.

slovní druh	nepřeklady		překlady		DIN
	tokeny	%	tokeny	%	
substantiva	5 502 041	20,72	4 832 060	18,15	-6,61
adjektiva	1 988 453	7,49	1 700 914	6,39	-7,92
pronomina	2 981 630	11,23	3 364 521	12,64	5,91
numeralia	446 766	1,68	305 719	1,15	-18,86
verba	4 201 372	15,82	4 684 899	17,60	5,32
adverbia	1 803 657	6,79	1 806 615	6,79	-0,04
prepozice	2 176 670	8,20	2 006 794	7,54	-4,18
konjunkce	1 889 785	7,12	1 885 087	7,08	-0,25
partikule	364 230	1,37	340 423	1,28	-3,50
interjekce	27 059	0,10	31 220	0,12	7,02
neurčeno	301 556	1,14	276 834	1,04	-4,40
interpunkce	4 868 321	18,34	5 382 437	18,15	4,89
celkem	26 551 540	100,00	26 617 523	100,00	0,00

Tabulka 4.2: Srovnání frekvenční distribuce slovních druhů – beletrie

Oproti tomu rozdíl ve výskytu citoslovcí v odborné literatuře nevyplývá ze složení korpusu jako takového, nýbrž z chybné lemmatizace a tagování u několika málo slov (*PR*, *PI*, *HR*). Tyto zkratky pak v celkovém součtu svou frekvencí zkreslují celkový výsledek citoslovcí v nepřekladové odborné literatuře. S těmito dílčími chybami však musí badatel, který využívá automaticky anotovaný korpus, počítat.

Konečně, srovnáme-li zájmena v překladové a nepřekladové odborné (DIN = 10,67), z frekvenčního seznamu vyčnívá především rozdíl v případě ukazovacího zájmena *ten*, které má v překladech výskyt 10 302,38 ipm oproti 8 379,16 v nepřekladové odborné literatuře (rozdíl je statisticky signifikantní na hladině významnosti $p < 0{,}001$). Tento výsledek může souviset s kategorií určenosti a s interferencí z angličtiny a dalších jazyků, jež používají členy (viz s. 78).

Fakt, že na nejvyšší úrovni slovních druhů nedochází mezi překlady a nepřeklady k zásadním rozdílům, však nemusí znamenat, že se překladové a nepřekladové texty

slovní druh	nepřeklady tokeny	%	překlady tokeny	%	DIN
substantiva	4 439 389	27,83	4 045 297	25,37	-4,63
adjektiva	1 962 341	12,30	1 719 111	10,78	-6,60
pronomina	1 114 150	6,99	1 379 899	8,65	10,67
numeralia	442 884	2,78	401 701	2,52	-4,86
verba	1 860 967	11,67	2 187 532	13,72	8,08
adverbia	799 276	5,01	852 786	5,35	3,25
prepozice	1 424 911	8,93	1 348 440	8,46	-2,75
konjunkce	1 018 342	6,38	1 136 385	7,13	5,50
partikule	176 158	1,10	170 193	1,07	-1,71
interjekce	3 098	0,02	2 310	0,01	-14,56
neurčeno	233 009	1,46	198 145	1,24	-8,07
interpunkce	2 475 405	15,52	2 504 220	15,70	0,59
celkem	15 949 930	100,00	15 946 319	100,00	0,00

Tabulka 4.3: Srovnání frekvenční distribuce slovních druhů – odborná literatura

neliší na nižší rovině obecnosti, v rámci jednotlivých slovních druhů. Pro další analýzu byly vybrány dva velké slovní druhy (substantiva a verba), které mohou coby autosémantika poukázat na určité lexikální tendence ve zkoumaných souborech. Bližší pohled je věnován také zájmenům, u nichž se může projevovat interference, a interpunkci, jejíž výskyt může souviset s hypotézou o simplifikaci (viz 4.2).

Cílem analýzy bylo zjistit, zda a jak se frekvenční seznamy těchto slovních druhů liší a především kolik a případně jakých slov se vyskytuje typicky jen v jednom ze souborů (překladech či nepřekladech). K vyhodnocení relevance rozdílu byla opět využita míra DIN. Rozsah slov pro analýzu nebyl předem nijak omezen (např. náhodným vzorkem), aby byl obraz co možná úplný. Jediné omezení se týká logického požadavku, aby dané slovo dosahovalo v celém korpusu Jerome minimálně takové frekvence, která odpovídá alespoň jednomu výskytu v každém ze čtyř sledovaných souborů (nepřekladové beletrii a odborné literatuře a překladové beletrii a odborné literatuře), a mohlo být tedy v rámci nich porovnáno. Po zohlednění velikosti souborů činí tato hranice minimální absolutní frekvence 5,4 výskytů, zaokrouhleno 6 výskytů (v přepočtu na relativní frekvenci 0,07 ipm).

Substantiva

Prvním krokem bylo získat frekvenční seznam substantiv s minimální absolutní frekvencí 6 ze všech čtyř zkoumaných souborů[3] spolu s údajem o pořadí ve frekvenčním seznamu (*rank*) a relativní frekvencí (*ipm*). Tyto čtyři seznamy byly poté sloučeny do dvou (zvlášť pro beletrii a odbornou literaturu) a deduplikovány, takže výsledné seznamy obsahují všechna lemmata vyskytující se v daném textovém typu, vč. informace, jaké pořadí a frekvenci mají v překladech a původních českých dílech.

[3]Z celkového počtu 88 705 substantivních lemmat v beletrii a 85 152 v odborné literatuře splnilo tuto hranici 69 453 lemmat a 67 652 lemmat.

Ze seznamů byla rovněž odstraněna lemmata s velkým písmenem (viz tabulka 4.4), která indikují propria, neboť můžeme předpokládat, že z tohoto hlediska se od sebe texty obsahově a tematicky liší vždy, avšak o charakteru překladových textů to nic nevypovídá. Pro další analýzu byla použita pouze apelativa.

SUBSTANTIVA	beletrie lemmata	%	odborná lemmata	%
apelativa	41 205	59,33	41 357	61,13
propria	28 248	40,67	26 295	38,87
celkem	69 453	100,00	67 652	100,00

Tabulka 4.4: Substantiva s min. frekvencí 6 v textových typech – apelativa a propria

Po vypočítání indexu DIN pak byly seznamy apelativ rozděleny podle toho, zda se slova vyskytují převážně či výhradně v jednom nebo ve druhém souboru nebo zda jsou oběma souborům společná bez rozdílu (viz tabulka 4.5). Z výsledků vyplývá, že většina substantivních lemmat je podle očekávání v obou textových typech společná oběma zkoumaným souborům – překladům i nepřekladům. Podíváme-li se ovšem na **slova s vysokou hodnotou DIN** (kladnou či zápornou), zjistíme, že množství substantiv, která jsou typická pro nepřekladové texty (tedy s hodnotou -75 až -100), je vyšší než počet substantiv převažujících v překladech (75 až 100): celkově 23,7 % substantiv oproti 6,16 % v beletrii a 22,11 % oproti 10,96 % v odborné literatuře (na základě provedeného testu chí-kvadrát[4] jsou rozdíly statisticky signifikantní na hladině významnosti $p < 0{,}001$, pro beletrii $\chi^2 = 4\ 436{,}71$, pro odbornou literaturu $\chi^2 = 1\ 221{,}56$).

APELATIVA	DIN	beletrie lemmata	%	odborná lemmata	%
A – pouze v překladech	100	2 099	5,09	3 406	8,25
B – výrazně v překladech	75–99	439	1,07	1 122	2,71
C – společná	-74 až 74	28 900	70,14	27682	66,93
D – výrazně v nepřekladech	-75 až -99	2 086	5,06	1 743	4,21
E – pouze v nepřekladech	-100	7 681	18,64	7 404	17,90
celkem	-	41 205	100,00	41 357	100,00

Tabulka 4.5: Srovnání relativní frekvence substantiv podle DIN

Tento nepoměr by mohl naznačovat, že škála používaných substantiv v překladech je menší než u nepřekladových textů, což může být dalším ukazatelem nižší bohatosti slovní zásoby v textech (viz simplifikace v části 4.2). Tento přehled se však týká pouze substantiv coby jednotlivých slov, ve skutečnosti může k dalším odlišnostem či naopak kompenzaci výše uvedených rozdílů docházet na úrovni slovních či slovnědruhových kombinací, tedy n-gramů či POS-gramů (viz další část 4.1.2).

[4]Tento test byl použit pro výpočet statistické signifikance v celé této kapitole 4, není-li uvedeno jinak.

Chceme-li slova, která jsou typická pro jednu či druhou skupinu, z kvantitativního hlediska blíže popsat, první informací může být opět **frekvence v korpusu**, tzn. zda se slova řadí mezi frekvenční špičku nebo spíše na periferii slovní zásoby. Pro každé substantivum v beletrii a odborné literatuře byla proto zjištěna jeho relativní frekvence v daném textovém typu v korpusu Jerome.

Tabulka 4.6 potvrzuje očekávání, že slova vyskytující se výhradně v jednom ze souborů (skupiny A a E) jsou velmi málo frekventovaná (například ipm 0,15 v beletrii odpovídá v tomto souboru 8 výskytům slova). Průměrná hodnota může být navíc zkreslena několika málo více frekventovanými slovy v souboru, proto v tabulce uvádím i modus, tedy hodnotu, která se v daném souboru vyskytuje nejčastěji; ta je pro všechny skupiny ještě přibližně o řád nižší než průměr a u překladových a nepřekladových textů se neliší. Na základě těchto pozorovaných veličin můžeme konstatovat, že mezi hodnotou DIN a frekvencí panuje nepřímá úměra – čím specifičtější je slovo pro určitý soubor, tím méně frekventované a více tematicky a žánrově vázané je.

APELATIVA	beletrie		odborná	
	průměr	modus	průměr	modus
A – pouze v překladech	0,15	0,04	0,33	0,06
B – výrazně v překladech	3,79	0,34	5,50	0,50
D – výrazně v nepřekladech	2,89	0,30	6,19	0,56
E – pouze v nepřekladech	0,24	0,04	0,40	0,06

Tabulka 4.6: Relativní frekvence substantiv v jednotlivých skupinách

Vedle většiny řídce se vyskytujících slov, o nichž nelze na základě několika málo výskytů mnoho říci, najdeme v analyzovaných skupinách i substantiva frekventovanější. Samotná frekvence však nemusí být dostačujícím kritériem, jak používané dané slovo ve skutečnosti je. Příkladem mohou být dvě slova z korpusu Jerome (*nedorozumění* a *křižník*), jež mají totožnou frekvenci 898 výskytů, ovšem jejich distribuce v korpusu se výrazně liší. Slovo *křižník* najdeme v 80 textech téměř výhradně z odborné literatury, kdežto na *nedorozumění* narazíme ve 458 různých textech z obou textových typů, beletrie i odborné. Dá se tedy očekávat, že se slovem *nedorozumění* se uživatel jazyka setká častěji, byť je v korpusu frekventované stejně.

Jednou z měr, která zohledňuje právě distribuci slova v korpusu z hlediska jeho výskytu v různých textech, je ARF neboli **průměrná redukovaná frekvence** (Salický & Hlaváčová 2003). Tato míra ukazuje, jak rovnoměrně je slovo v korpusu rozloženo, bere tedy v úvahu jeho disperzi. Čím je rozložení rovnoměrnější, tím více se hodnota ARF blíží frekvenci slova a naopak; pro výrazy, jejichž výskyty jsou v korpusu soustředěny do jediného shluku, např. v rámci jediného textu, se hodnota ARF blíží jedné bez ohledu na frekvenci. Pro výpočet ARF jednotlivých substantiv byl použit referenční korpus SYN2010, který zde reprezentuje vzorek současného psaného jazyka (bez rozlišení textových typů).

Následující výběr zahrnuje **dvacet nejfrekventovanějších substantiv** z beletrie ze dvou skupin B a D, jež mají zároveň co nejvyšší referenční hodnotu ARF; skupiny A a E neobsahují takřka žádná více frekventovaná slova, proto zde byly

ponechány stranou. Ze seznamu byly ručně vytříděny chyby v lemmatizaci (např. vlastní jména lemmatizovaná s malým písmenem) a zkratková slova (*stol.*, *sv.*, *km* apod.)

- B (výrazně v překladech): *tóra, seržant, lord, sir, monsieur, senátor, drahoušek, sendvič, láma, šerif, mademoiselle, grál, sáhib, samuraj, pastor, brandy, zlatíčko, konstábl, astma, kimono*

- D (výrazně v nepřekladech): *fakulta, rozhledna, socialismus, komunismus, čeština, bronz, freska, nadace, cikán, rozhled, hajný, zájezd, fabrika, myslivec, primář, sloh, beseda, protektorát, vstupné, polemika*

Z přehledu jasně vyplývá, že frekvence těchto substantiv je v naprosté většině případů ovlivněna tématem a zaměřením konkrétního textu a o specifických rysech překladového jazyka příliš mnoho nevypovídá. Seznam slov naznačuje, že v nepřekladových textech v korpusu Jerome jsou tematizovány například epochy spjaté s českou historií (*socialismus*, *komunismus*, *protektorát*). Dále zde najdeme slova, která jsou typičtější pro publicistiku (*vstupné*, *nadace*, *beseda*) nebo odbornou literaturu (*polemika*, *fakulta*); v beletristické části korpusu se nacházejí v tzv. literatuře faktu (FAC), která stojí na pomezí mezi beletrií a odbornou literaturou a zastupují ji v korpusu především memoáry. Vzhledem k tomu, že korpus mohl být vyvážen jen na nejvyšší úrovni beletrie a odborné literatury, v nepřekladové části najdeme textů FAC výrazně víc (viz graf 3.11), a proto se tato slova dostala do popředí. U překladů nepřekvapí, že převažují cizojazyčná oslovení a názvy vycházející z reálií cizích zemí (vzhledem ke skladbě korpusu jsou to především slova z anglofonní oblasti).

Kromě těchto substantiv se však v překladech výrazně častěji používají výrazy *drahoušek* (606 výskytů oproti 56 v nepřekladových textech, $\chi^2 = 456{,}66$) a *zlatíčko* (300 výskytů oproti 40, $\chi^2 = 197{,}94$; rozdíly jsou u obou slov statisticky signifikantní na hladině významnosti $p < 0{,}001$), které zcela jistě odrážejí odlišný úzus oslovení, ovšem vyjadřují také oblibu překladatelů v těchto konkrétních překladových protějšcích. Pro srovnání, výskyt obdobného výrazu *miláček* je v překladových i nepřekladových textech vyrovnanější (999 výskytů v překladech vůči 712, $\chi^2 = 48{,}52$, opět statisticky signifikantní na $p < 0{,}001$). Výraz *drahoušek* se v překladové beletrii vyskytuje v naprosté většině v textech přeložených z angličtiny (530 výskytů ve 118 textech), dále pak z francouzštiny (35 výskytů v 9 textech), ruštiny (11 výskytů ve 3 textech) nebo dánštiny (10 výskytů ve 2 textech). U výrazu *zlatíčko* je situace obdobná: výrazně převažuje v textech z angličtiny (248 výskytů v 83 textech), po několika málo výskytech najdeme i v původně německy psaných (10) a francouzsky psaných (9). Počet různých textů, v nichž se zkoumané ekvivalenty vyskytují, potvrzuje, že nejde o idiolekt jediného překladatele. Kontrolní pohled do paralelního korpusu InterCorp[5] potvrzuje, že výraz *zlatíčko* je v angličtině nejčastěji překladem slov *honey*, *precious*, *sweetheart* a *dear*, podobně jako *drahoušek*, kterým se překládá *dear*, *darling*, *sweetheart* a *honey*.

[5]Anglicko-česká část, verze 7 z 19. 12. 2014.

Chceme-li ověřit **vliv literatury faktu** (textový typ FAC), můžeme porovnat seznamy substantiv z ad hoc utvořených beletristických subkorpusů překladové a nepřekladové češtiny (pouze textové typy NOV, COL a IMA). Po replikování postupu, kterým jsme získali výše uvedený seznam dvaceti nejfrekventovanějších substantiv s hodnotami DIN (75 až 99 a -75 až -99), získáme obdobný soupis substantiv bez vlivu FAC. V seznamu podle očekávání chybí žánrově specifická slova (*fakulta*, *nadace*, *polemika*), ovšem převažující tendence zůstávají totožné. V překladové beletrii zůstávají na čelných místech výrazy označující cizokrajné reálie (*tóra, seržant, lord, sir, kaplan, monsieur, libra, sendvič, láma, šerif, mademoiselle, vévodkyně, sáhib, samuraj, grál*) a v první desítce zůstává i *drahoušek*. V nepřekladové beletrii pak na předních místech figurují výrazy *soudruh, stráň, sumec, dědek, cikán, hajný, myslivec, lágr, panoš, basket, primář, hlavolam, náves, zájezd, jatky* a další, které jsou úzce spjaty s tématem textů v subkorpusu.

Na základě této analýzy můžeme konstatovat, že největší rozdíly mezi užitím konkrétních substantivních lexémů v překladech a nepřekladových textech pramení v naprosté většině případů z tématu textů a/nebo zahrnutého žánru/textového typu, nikoli z vlivu překladového jazyka: nejvíce odlišná substantiva (na základě míry DIN) jsou velmi málo frekventovaná a úzce spjatá s konkrétním tématem či zaměřením textu. Kromě dílčí preference u dvou výrazů (*drahoušek, zlatíčko*) se v rámci zkoumaných souborů nevydělily, např. na základě sémantiky, žádné skupiny substantiv s výrazně odlišným užitím (s hodnotou DIN nad +/-75), jež by bylo dáno vlivem překladu. Doplňující pohled na substantivní lemmata s nižší absolutní hodnotou DIN (35–74) a tedy s méně vyhraněným užitím potvrdil, že tato slova patří sice do vyššího frekvenčního pásma, ovšem stejně odrážejí především téma textu (*výraz, paže, policie, vražda, dolar* v překladech, *les, pivo, tatínek, hospoda, básník* v nepřekladových textech).

Kombinacím substantiv s ostatními slovními druhy je věnována část 4.1.2.

Verba

Stejnou metodou jako v případě substantiv byla v korpusu Jerome vyhledána také všechna slovesná lemmata (s minimální frekvencí 6). Z přehledu v tabulce 4.7 vyplývá, že v beletrii je situace stejná jako u substantiv: sloves, jež se typicky vyskytují v nepřekladových textech, je celkem 15,09 % oproti pouhým 3,99 % v překladech (statisticky signifikantní na hladině významnosti $p < 0{,}001$, $\chi^2 = 771{,}61$). V odborné literatuře je však poměr opačný a rozdíly, ač také signifikantní ($\chi^2 = 27{,}13$, $p < 0{,}001$), nejsou tak výrazné: 13,29 % v překladech oproti 10,57 % v nepřekladových textech. Na rozdíl od substantiv obsahují skupiny B a D zanedbatelný počet sloves; naopak společných sloves mají překlady a nepřeklady přibližně o 10 % více než substantiv.

Chceme-li zjistit, nakolik se v textech u sloves projevuje **repetitivnost**, můžeme se podívat na poměr typů (v tomto případě unikátních lemmat) a tokenů (všech výskytů lemmat ve všech tvarech). Tento poměr (*type-token ratio* neboli TTR) je pro snadnost svého výpočtu často používanou mírou, jež je ovšem velmi závislá na délce textu/korpusu (podrobněji o TTR viz 4.2). V tomto případě jsou vždy oba

VERBA	DIN	beletrie lemmata	%	odborná lemmata	%
A – pouze v překladech	100	429	3,97	1 045	13,03
B – výrazně v překladech	75–99	2	0,02	21	0,26
C – společná	-74 až 74	8 751	80,92	6 105	76,13
D – výrazně v nepřekladech	-75 až -99	12	0,11	17	0,21
E – pouze v nepřekladech	-100	1 620	14,98	831	10,36
celkem	-	10 814	100,00	8 019	100,00

Tabulka 4.7: Srovnání relativní frekvence sloves podle DIN

srovnávané soubory (překladová a nepřekladová beletrie; překladová a nepřekladová odborná literatura) srovnatelně velké, TTR nám proto může posloužit jako orientační ukazatel. Většinou se TTR počítá jako poměr typů k tokenům, ovšem pro lepší vizualizaci byl v tabulce zvolen opačný poměr, tedy kolik tokenů připadá na jeden typ.

VERBA	beletrie typy	tokeny	TTR	odborná typy	tokeny	TTR
nepřeklady	10 383	4 201 372	404,64	6 953	1 860 967	267,65
překlady	9 182	4 684 899	510,23	7 171	2 187 532	305,05

Tabulka 4.8: Typy a tokeny u slovesných lemmat

Tabulka 4.8 ukazuje, že v beletristických překladech dochází k častějšímu opakování týchž sloves. Na jedno slovesné lemma zde připadá v překladu přibližně 510 tokenů (výskytů), kdežto v případě nepřekladů je to jen 404 ($\chi^2 = 12{,}29$, $p < 0{,}001$). V odborné literatuře u sloves k výrazným rozdílům mezi překladovými a nepřekladovými texty nedochází (výsledek není statisticky signifikantní, $\chi^2 = 2{,}44$, $p = 0{,}12$). Větší repetitivnost sloves v překladech může opět odrážet nižší bohatost slovní zásoby, která je jedním z argumentů v hypotéze o simplifikaci (viz 4.2).

Z porovnání seznamů slovesných lemmat na základě nastavených parametrů míry DIN vyplynulo, že skupiny B a D (tedy sloves převažujících v jednom nebo druhém souboru) obsahují velmi málo slov (viz 4.7); buď se slovesa vyskytují pouze v jednom ze souborů (pak se jedná o velmi málo frekventovaná a silně tematicky vázaná slova jako v případě substantiv), nebo jsou oběma souborům společná (dosahují hodnot DIN v rozmezí -74 do +74). Tato společná skupina je však velmi široce definovaná; zaměříme-li se na drobnější rozdíly, můžeme i v ní najít vhodné kandidáty s odlišným výskytem v překladech. Snížíme-li hranici relevance podle hodnoty DIN na hodnoty 35–74 a setřídíme-li slova podle frekvence a jejich referenční ARF, získáme soupis sloves, která jsou výrazněji zastoupena v jednom ze souborů a přitom patří mezi frekventované a rovnoměrně distribuované typy. Aby v rámci beletrie nedošlo ke zkreslení vlivem literatury faktu (FAC), byly texty tohoto typu z této analýzy vyloučeny. Následující seznam zahrnuje **dvacet nejfrekventovanějších slovesných lemmat** vybraných na základě zmíněných kritérií:

- (převážně v překladech): *prohlásit, promluvit, přikývnout, poznamenat, zavrtět, zírat, zamířit, strávit, zadívat, zmínit, ujistit, zamumlat, odmlčet, přimět, vzhlédnout, zamračit, zaváhat, potřást, zavraždit, naléhat*

- (převážně v nepřekladech): *pravit, divit, řvát, počít, vypravovat, závidět, leknout, optat, vznikat, žrát, lítat, sežrat, vozit, volit, kecat, loučit, sypat, hučet, nalézat, usínat*

Na obou seznamech nalezneme slova, která můžeme přičítat tématu textů (např. sloveso *zavraždit* v žánru krimi, *pravit* u Karla Čapka nebo značný výskyt slovesa *volit* v knize Ludvíka Vaculíka *Poslední slovo*), ale lze zde také vypozorovat určité lexikální tendence v obou souborech. V nepřekladových textech častěji narazíme na slovesa s příznakem expresivity (*řvát, žrát, sežrat, kecat*), a to nikoli v několika málo dílech, nýbrž rovnoměrně v rámci beletristické části korpusu (s výjimkou vyloučených textů FAC). Rozdíl v užití těchto sloves v překladech může opět poukazovat na tendenci překladatelů volit neutrálnější, nepříznakové lexikální prostředky. V seznamu také najdeme několik sloves nedokonavých *vznikat, nalézat, usínat*, jejichž dokonavé protějšky se vyskytují v překladech i nepřekladech srovnatelně, kdežto tato nedokonavá varianta je v překladech řídká.

Na seznamu z překladových textů se objevují slovesa, která se v češtině využívají jako uvozovací (*prohlásit, poznamenat, zmínit, ujistit, zamumlat, odmlčet*), což svědčí o větším výskytu přímé řeči v překladové beletrii v korpusu Jerome (to potvrzuje i podrobnější pohled na interpunkci a vyšší výskyt uvozovek, viz 4.15). Slovesem, které poněkud vybočuje z řady, je *přimět*, neboť se často vyskytuje v kombinaci s infinitivem (ve 24 % v překladech a 15 % v nepřekladech), který je v překladových textech také frekventovanější (viz dále). Vyšší výskyt slovesa *přimět* může také odrážet složení překladové části korpusu – na základě zdrojového jazyka textů, v nichž se toto sloveso vyskytuje, můžeme usuzovat, že jde o vliv angličtiny a jejích konstrukcí typu *make/get/force sb do sth.*

Soubor sloves můžeme blíže charakterizovat i pohledem na nižší, gramatickou rovinu. Podíváme-li se do frekvenčního seznamu sloves, jaké konkrétní **slovesné tvary** se v překladových a nepřekladových textech nejvíce liší, zjistíme, že celkově překlady obsahují přibližně o 20 % více infinitivních tvarů ($\chi^2 = 9\ 741{,}56$, $p < 0{,}001$; DIN = 9,04), viz tabulka 4.9.

INFINITIVY	beletrie ipm	odborná ipm	celkem ipm
nepřeklady	16 670,14	16 072,36	32 742,50
překlady	19 031,11	20 222,53	39 253,64

Tabulka 4.9: Srovnání počtu infinitivních tvarů

Významnější rozdíly však najdeme v užití jednotlivých sloves v infinitivním tvaru. Slovesa s variantami **infinitivních koncovek *-ci*** a ***-ct*** (např. *říci*/*říct*, *moci*/*moct*) mají odlišnou distribuci: v překladech (bez rozlišení textového typu)

mírně převažuje varianta *-ct* (55,8 %), jež je považována za novější a stylově neutrální, kdežto v nepřekladových textech je častější varianta druhá: *-ci* se vyskytuje v 60,8 %, *-ct* ve zbývajících 39,2 % případů. Podíváme-li se na textové typy zvlášť, potvrdí se nejen obecná odlišnost beletrie a odborné literatury, ale i výraznější rozdíl v užití obou variant, viz 4.10 a 4.11. Zatímco v beletrii se projevuje tendence překladatelů k výběru novější varianty (v 69,6 % případů), v odborné literatuře převažuje v obou zkoumaných souborech varianta *-ci*, kterou můžeme považovat za formálnější. Možné zdůvodnění, zda se v odborné literatuře nejedná o jiná slovesa (např. modální *moci*), můžeme při pohledu na konkrétní realizace těchto infinitivních variant v odborné literatuře vyloučit – výrazně nejfrekventovanějším slovesem je v beletrii i odborné literatuře v obou souborech sloveso *-říci/-říct*.

BELETRIE	nepřeklady		překlady		test	
	ipm	%	ipm	%	χ^2	p-hodnota
-ci	232,27	52,6	186,83	30,4	133,86	$p < 0{,}001$
-ct	209,67	47,4	426,75	69,6	1 956,43	$p < 0{,}001$
celkem	414,94	100,00	613,58	100,00	1 008,07	$p < 0{,}001$

Tabulka 4.10: Srovnání infinitivní koncovky -ci/-ct v beletrii (vč. FAC)

ODBORNÁ	nepřeklady		překlady		test	
	ipm	%	ipm	%	χ^2	p-hodnota
-ci	225,27	83,2	296,62	84,3	155,35	$p < 0{,}001$
-ct	45,52	16,8	55,44	15,7	15,35	$p < 0{,}001$
celkem	270,79	100,00	352,06	100,00	168,92	$p < 0{,}001$

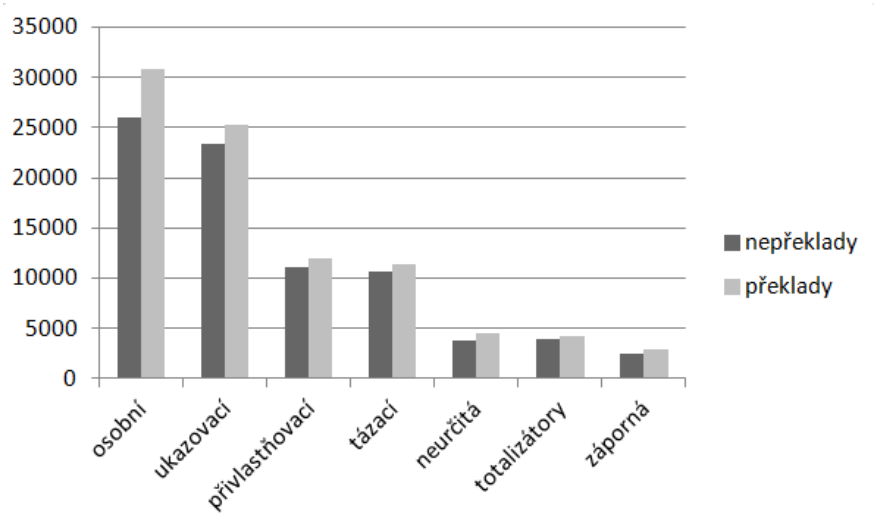
Tabulka 4.11: Srovnání infinitivní koncovky -ci/-ct v odborné literatuře

U slovesa *říci/říct* se rovněž projevují nejvýraznější rozdíly: v překladech se objevuje výrazně ve variantě *říct* (311,9 ipm, tedy 78,5 %), kdežto v nepřekladových textech dosahuje její výskyt 54,4 % (153 ipm). Preference novější varianty by mohla naznačovat tendenci překladatelů volit aktuálnější, bezpříznakový jazykový prostředek, což může souviset s tendencí k přehnané normalizaci u překladových textů. Možný vliv roku vydání díla zde můžeme vyloučit, neboť v obou souborech (u překladů i nepřekladových textů), v nichž se obě varianty vyskytují, jsou rovnoměrně zastoupeny texty starší i novější.

Snahou překladatelů vybírat bezpříznakové varianty bychom mohli zdůvodnit i skutečnost, že v překladové beletrii se hovorová varianta *bejt* vyskytuje takřka dvakrát méně často než u nepřekladových textů (30,13 ipm ve 119 textech oproti 57,28 ipm ve 146 textech; $\chi^2 = 222{,}24$, $p < 0{,}001$), a to navzdory již zmíněnému rozdílu ve složení obou beletristických částí korpusu (viz graf 3.11 a možný vliv literatury faktu).

Pronomina

Zájmena tvoří oproti substantivům a slovesům nepříliš početnou a prakticky uzavřenou třídu slov, kterou lze rozdělit do tradičně vymezovaných skupin. Graf 4.12 ukazuje, jak se liší distribuce jednotlivých druhů zájmen **v beletristické části** korpusu Jerome. Z výsledků srovnání vyplývá, že nepatrně vyšší výskyt zájmen v překladech (viz tabulka 4.2) se projevuje především u zájmen osobních a ukazovacích. Z hlediska hodnoty DIN vykazují největší rozdíly zájmena záporná (DIN = 8,76) a osobní (DIN = 8,47) a dále pak neurčitá (DIN = 7,62) a přivlastňovací (DIN = 3,97).

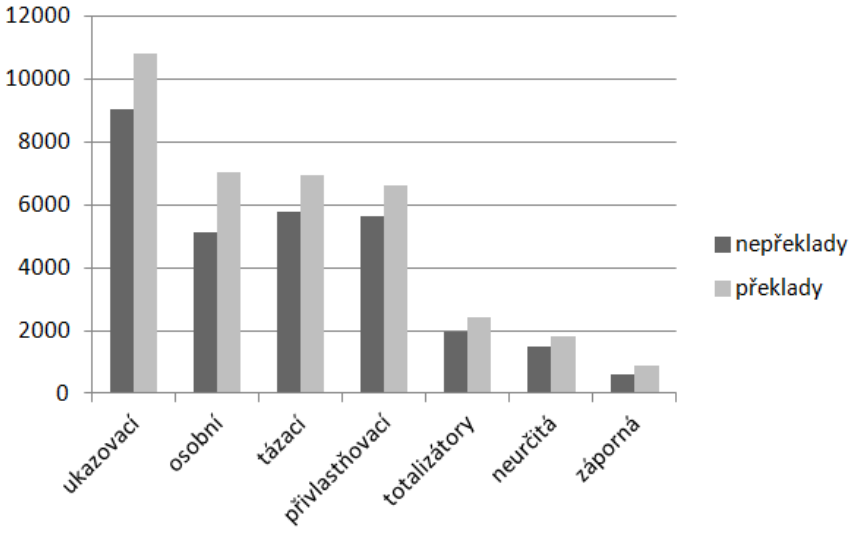


Obrázek 4.12: Srovnání relativní frekvence druhů zájmen (v tis. slov) – beletrie

Podíváme-li se na konkrétní zájmena, na výrazně odlišnou frekvenci užití (hodnoty DIN od -50 do -100 a od 50 do 100) narazíme především u zájem neurčitých, která v rámci tohoto slovního druhu spolu s ukazovacími zájmeny představují skupinu s největší slovotvornou potencionalitou; škála různých neurčitých výrazů se tak může lišit text od textu, výsledky ovšem naznačují, že bohatší spektrum najdeme v nepřekladových textech. Následující zájmena se v korpusu Jerome vyskytují především v nepřekladové beletrii: *čertvíco*, *všelijaký*, *všelicos*, *ledasco*, *kterýsi*, *ledaco*.

Distribuce zájmen **v odborné literatuře** se nejvíce liší u zájmen ukazovacích, osobních a tázacích. Na první pohled je také patrné, že zájmena v odborné literatuře netvoří tak prominentní slovní druh jako v beletrii (zvláště pak zájmena osobní). Z hlediska míry DIN se nejvíce odlišují zájmena záporná (18,15) a osobní (15,60).

V souvislosti se zájmeny ukazovacími se v českých lingvistických kruzích vedou již mnoho desetiletí diskuze o tom, zda čeština má a vyjadřuje **kategorii určenosti** podobně jako jiné jazyky (např. angličtina či němčina). Již Zubatý v roce 1917 v časopise Naše řeč vznesl otázku, zda ukazovací zájmeno *ten* neplní v češtině funkci určitého členu. O tři roky později se však sám z puristického hlediska kriticky vymezuje vůči srovnávání českého *ten* a německým *der*. O mnoho let později se k tomuto



Obrázek 4.13: Srovnání relativní frekvence druhů zájmen (v tis. slov) – odborná literatura

tématu vrací Mathesius (1947) v závěru své stati o zájmenech *ten*, *ta*, *to* v hovorové češtině. Srovnání s anglickým určitým členem *the* je podle něj možné tehdy, stojí-li zájmeno *ten* před „výrazy komparativními, superlativními a před diferencujícími větami relativními“, ovšem o skutečném členu v češtině podle něj hovořit nelze, jelikož tento nepatří do jazykového systému, nýbrž do významové výstavby věty (Mathesius 1947: 188). Zájmeno *ten* v mluvené češtině považuje za protějšek anglického určitého členu i Kodýtek v *Mluvnici současné češtiny* (Cvrček et al. 2010).

Použití ukazovacího zájmene ve funkci členu je v češtině i ve výše zmíněných případech (např. před adjektivy v superlativu) fakultativní (Chlumská & Kováříková 2009: 148). Vyjdeme-li z těchto informací, pak bychom mohli zformulovat alternativní hypotézu, že v překladových textech z jazyka, který kategorií určenosti ve svém systému disponuje, bude vlivem interference výskyt zájmena *ten* v těchto případech (v superlativních konstrukcích) vyšší než v textech původně česky psaných. Nulová hypotéza zní, že výskyt zmíněné konstrukce bude v obou zkoumaných souborech stejný. Na základě výsledků provedených v beletristické části korpusu Jerome můžeme nulovou hypotézu vyvrátit: v překladech dosahuje konstrukce *ten* + ADJ v superlativu 211,93 ipm, což tvoří 19 % výskytů superlativní konstrukce jako takové. V nepřekladové beletrii má frekvenci 156,34 ipm (přibližně 11 %). V odborné literatuře činí výskyt 171,62 ipm v překladech (9 %) vůči 107,53 ipm (6 %). Rozdíly jsou v obou textových typech statisticky signifikantní. Překlady z angličtiny tvoří podle očekávání největší část textů (203,94 ipm v 283 textech v beletrii a 183,14 ipm ve 138 textech v odborné literatuře), tuto strukturu však najdeme hojně zastoupenou i v překladech ze švédštiny, ruštiny či francouzštiny (na rozdíl od angličtiny se však jedná se o jednotlivá díla, nelze tedy vyloučit vliv idiolektu překladatele).

Interpunkce

Výzkumu interpunkce nebyla v korpusové translatologii doposud věnována příliš velká pozornost. Jedním z důvodů je jistě fakt, že volba interpunkce do značné míry závisí na autorovi (příp. redaktorovi) textu a vypovídá tak především o autorském stylu. Kreativní využití interpunkce je typické především pro literární texty. Většina translatologických studií, které o interpunkci vznikly, vychází z kontrastivního pohledu (srovnání zdrojového textu a cílového překladu, tedy z pohledu S-univerzálií). Malmkjærová (1997) a Mayová (1997) se ve svých kvalitativních studiích zaměřily na to, jak „věrná“ je interpunkce v překladech ve vztahu k originálu, tedy zda-li se překladatelé staví spíše do role redaktora a text upravují, aby odpovídal konvencím v cílovém jazyce, nebo respektují daný autorský styl. Mayová i Malmkjærová se na základě svého výzkumu přiklánějí k první možnosti.

Z novějších studií, které se interpunkci věnují především z pohledu cílového textu a porovnávají překladové a nepřekladové texty, stojí za zmínku výzkum Mónicy Rodríguez-Castrové (2011), která porovnávala větnou interpunkci na malém srovnatelném korpusu publicistických textů původně španělsky psaných a přeložených z angličtiny. Výsledky jejího zkoumání naznačují, že překladové texty obsahují méně čárek a více teček, což Rodríguez-Castrová přisuzuje simplifikační strategii překladatelů a tendenci psát kratší a srozumitelnější věty (Rodríguez-Castro 2011: 52).

Podíváme-li se nejprve na přehled (viz tabulka 4.14), jak velkou část zkoumaných souborů v korpusu interpunkce[6] zabírá, zjistíme, že výskyt je srovnatelný, pouze v překladové beletrii můžeme pozorovat o něco vyšší frekvenci.

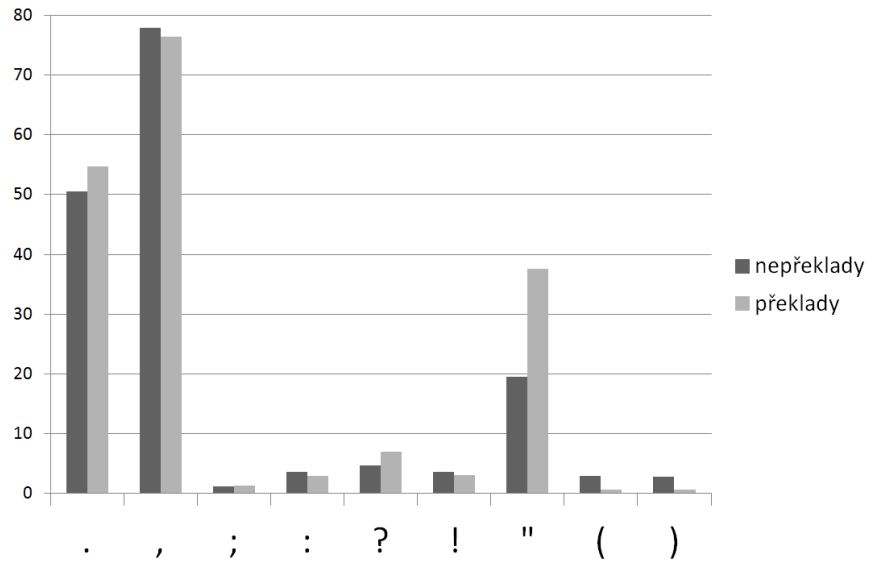
INTERPUNKCE	beletrie		odborná	
	ipm	%	ipm	%
nepřeklady	183 353,62	18,3	155 198,49	15,5
překlady	202 214,05	20,2	157 040,63	15,7

Tabulka 4.14: Poměr interpunkce u překladů a nepřekladů

Pro bližší analýzu byla vybrána pouze ta interpunkční znaménka, která nejčastěji strukturují text: tečka, čárka, středník, dvojtečka, otazník, vykřičník, uvozovky a kulaté závorky. Vzhledem k tomu, že interpunkce má v korpusu Jerome vždy vlastní pozici (je oddělena z obou stran mezerami), je například u tečky při hledání nezbytné specifikovat, že má jít o tečku koncovou, nikoli za řadovou číslovkou nebo zkratkou. K tomu byl využit strukturní atribut označující konec věty (</s>).

Jak ukazuje graf 4.15, obdobnou tendenci, jakou popisuje Rodríguez-Castrová, pozorujeme i v překladové části korpusu Jerome. V překladech beletrie se objevuje přibližně o 10 % více teček a o 50 % více otazníků, zatímco čárky indikující delší souvětí nebo vsuvky se vyskytují více v nepřekladové části (viz tabulka 4.17). Rozdíly ve výskytu závorek a uvozovek plynou ze složení korpusu. Závorky můžeme spíše čekat v textech na pomezí beletrie a odborné literatury (opět vliv textového

[6]Interpunkce v tomto pojetí (v rámci morfologického značkování označena jako Z) zahrnuje veškerá interpunkční znaménka členící text, ale také matematické symboly. Ty však tvoří poměrně zanedbatelnou část.

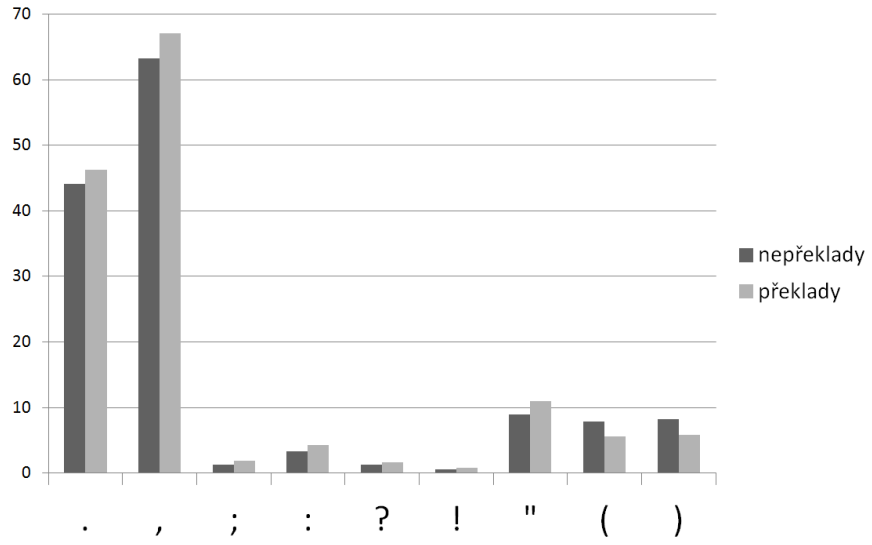


Obrázek 4.15: Srovnání vybrané interpunkce (v tis.) – beletrie

typu FAC v nepřekladové literatuře), kdežto uvozovky značí přímou řeč (převaha románů NOV v překladech). V odborné literatuře je situace odlišná, podle grafu 4.16 najdeme v překladech nejen více teček, ale všech zkoumaných interpunkčních znamének, s výjimkou závorek. Překladový odborný text se tak z hlediska interpunkce zdá být nepatrně více členěn.

Vypočítáme-li pro zmíněná interpunkční znaménka a jejich naměřenou frekvenci hodnotu DIN, dostaneme pro beletrii hodnoty v rozmezí -64,75 (pro závorky v nepřekladových textech) a 31,79 (pro uvozovky v překladech). Všechny rozdíly jsou opět statisticky signifikantní ($p < 0{,}001$), přičemž pro tečku vykazuje míra DIN v beletrii hodnotu 3,96 a v odborné literatuře 2,34, pro otazník 20,12 v beletrii a 13,8 v odborné. Opačné tendence vykazuje vykřičník, který je častější v nepřekladové beletrii a překladové odborné literatuře. Středník má vyšší frekvenci v překladech v obou textových typech. Z pohledu na zdrojový jazyk dokumentů vyplývá, že jde především o vliv angličtiny (1 174,8 ipm ve 229 textech), vyšší počet středníků však najdeme i v textech přeložených z portugalštiny (9 969,9 ipm ve dvou textech), finštiny (8 072,4 ipm ve 3 textech) či španělštiny (3 249,2 ipm v 10 textech). Vysoká relativní frekvence je ale dána především vlivem jednoho či dvou děl (nelze tedy vyloučit idiolekt autora či spíše překladatele), kdežto v případě angličtiny jde o konstantní jev.

Platí-li, že výskyt většího počtu interpunkčních znamének, jež značí konec věty, v překladech poukazuje na tendenci překladatelů k jednodušším větám, měl by tento rozdíl být patrný i v samotné délce vět. Srovnání délky vět proto bylo provedeno v rámci výzkumu simplifikace (výsledky viz kapitola 4.2).



Obrázek 4.16: Srovnání vybrané interpunkce (v tis.) – odborná literatura

INTERPUNKCE	beletrie nepřeklady	překlady	DIN	odborná nepřeklady	překlady	DIN
tečka	50 523,02	54 687,73	3,96	44 134,05	46 245,72	2,34
čárka	77 859,51	76 387,05	-0,95	63 263,10	67 119,00	2,96
středník	1 175,22	1 310,87	5,46	1 291,10	1 863,75	18,15
dvojtečka	3 583,97	2 869,50	-11,07	3 257,19	4 285,31	13,63
otazník	4 652,08	6 996,29	20,12	1 266,71	1 672,42	13,80
vykřičník	3 538,51	3 003,96	-8,17	497,18	750,89	20,33
uvozovky	19 467,27	37 615,07	31,79	8 948,57	10 909,04	9,87
kulatá závorka levá	2 880,89	616,44	-64,75	7 820,79	5 597,03	-16,57
kulatá závorka pravá	2 841,42	652,99	-62,63	8 227,37	5 843,61	-16,94

Tabulka 4.17: Srovnání výskytu interpunkčních znamének (v ipm)

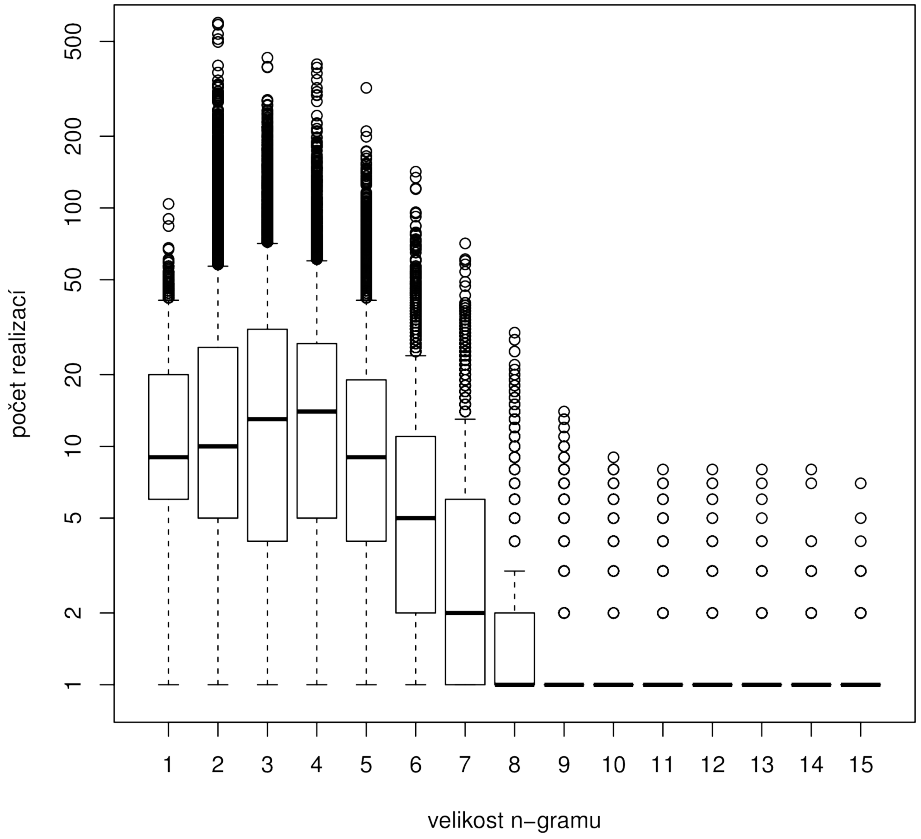
4.1.2 Časté kombinace slovních druhů (POS-gramy)

Typické sekvence slovních druhů, tzv. **POS-gramy**, a jejich nejčastější zástupci mohou leccos napovědět o povaze textu, např. o jeho dynamice (s ohledem na převahu verbálních či nominálních kombinací) či lexikální hutnosti (kumulace autosémantik nebo naopak preference kombinací se synsémantiky). Výzkum POS-gramů bývá často kombinován s výzkumem n-gramů (tedy konkrétních lexikálních realizací) a uplatňuje se i v korpusové translatologii (např. Corpas Pastor, Mitkov, Afzal & Pekar 2008) nebo ve výzkumu dětské literatury (Thompson & Sealey 2007: 12).

Srovnání nejčastějších POS-gramů

Pro výzkum POS-gramů v korpusu Jerome byla zvolena délka čtyř po sobě jdoucích jednotek, která byla určena na základě výzkumu desambiguace kontextu jako struktura, která má v korpusu průměrně největší počet různých realizací (viz graf 4.18

podle Cvrček & Václavík, v tisku). Lze proto předpokládat, že struktura 4-gramu je dostatečně dlouhá na to, aby se v ní mohly projevit kombinatorické tendence, a zároveň ne příliš dlouhá, aby alespoň její nejčastější konkrétní realizace měly dostatečný výskyt pro obecné vyhodnocení a porovnání.



Obrázek 4.18: Vztah mezi velikostí n-gramu a počtem jeho realizací (SYN2010)

Pro každý textový typ zvlášť byly v korpusu Jerome vyhledány nejčastější po sobě jdoucí sekvence slovních druhů o délce čtyř pozic, přičemž byla v tomto případě předem z analýzy vyloučena interpunkce, neboť vzhledem ke své frekvenci (viz tabulka 4.14) bude nutně tvořit nejčastější POS-gramy a potenciálně zajímavější kombinace ostatních slovních druhů by tak mohly zůstat opomenuty.

Tabulka 4.19 obsahuje souhrn POS-gramů, které se v překladové a nepřekladové beletrii (s vynecháním literatury FAC) nejvíce odlišují. Nejvýraznější rozdíl (DIN = -15,00) se týká sekvence N-N-R-N[7] (substantivum – substantivum – prepozice – substantivum), která převažuje v překladech. V odborné literatuře (viz tabulka 4.20) vykazuje největší rozdíl konstrukce N-N-A-N (substantivum – substantivum – adjektivum – substantivum), která je preferována v nepřekladové odborné literatuře (DIN = -25,18).

[7]Zkratky vycházejí z morfologického tagu a jejich seznam je k dispozici na adrese: http://wiki.korpus.cz/doku.php/seznamy:tagy.

POS-gramy	nepřeklady		překlady		
	rank	ipm	rank	ipm	DIN
J-V-P-V	45.	968,72	30.	1 197,64	10,57
P-R-P-V	14.	1 577,46	11.	1 904,80	9,40
J-V-P-N	54.	917,44	35.	1 095,45	8,84
J-P-N-V	47.	955,29	32.	1 135,03	8,60
V-P-R-P	20.	1 401,91	14.	1 665,14	8,58
J-V-P-R	34.	1 116,30	24.	1 310,92	8,02
N-R-N-N	29.	1 198,63	45.	984,03	-9,83
N-R-N-V	33.	1 123,68	50.	910,19	-10,50
A-N-J-N	36.	1 082,34	57.	875,54	-10,56
R-N-A-N	13.	1 608,97	25.	1 237,35	-13,06
A-N-A-N	30.	1 196,89	52.	908,38	-13,70
N-N-R-N	35.	1 087,22	72.	803,52	-15,00

Tabulka 4.19: Srovnání POS-gramů v beletrii (NOV, COL, IMA)

POS-gramy	nepřeklady		překlady		
	rank	ipm	rank	ipm	DIN
V-P-A-N	71.	1 112,29	31.	1 481,72	14,24
V-R-P-N	61.	1 210,60	28.	1 519,35	11,31
P-V-R-N	25.	1 940,07	10.	2 378,48	10,15
P-V-A-N	41.	1 520,32	20.	1 806,87	8,61
V-N-R-N	37.	1 620,13	18.	1 874,98	7,29
R-P-A-N	56.	1 270,29	36.	1 428,67	5,87
N-A-A-N	16.	2 346,78	27.	1 585,38	-19,36
N-R-N-A	21.	2 093,11	43.	1 373,67	-20,75
A-N-N-N	31.	1 709,10	69.	1 104,33	-21,50
N-A-N-N	24.	1 968,22	55.	1 250,01	-22,32
N-A-N-R	18.	2 236,56	42.	1 376,74	-23,80
N-N-A-N	19.	2 188,41	47.	1 307,95	-25,18

Tabulka 4.20: Srovnání POS-gramů v odborné literatuře

Z obou tabulek je na první pohled patrné, že v překladech se více prosazují verbální konstrukce, kdežto v nepřekladových textech jsou to konstrukce nominální se substantivy a adjektivy. Potvrzuje to trend, který vyplývá z tabulek 4.2 a 4.3: překlady vykazují vyšší výskyt sloves, kdežto původně česky psané texty obsahují přibližně o 2 % více substantiv. Zajímavý je také fakt, že 4-gramy s nejrozdílnějším výskytem se v obou textových typech vůbec nepřekrývají – v tabulkách 4.19 a 4.20 nenajdeme jedinou společnou konstrukci. To jen potvrzuje odlišnost beletrie a odborné literatury a poukazuje na to, že je nezbytné oba textové typy analyzovat zvlášť. Následující část kapitoly 4.1 je věnována stručnému přehledu konkrétních realizací nejvýraznějších 4-gramů, jehož cílem je popsat možné odlišnosti a identifikovat případné opakující se lexikální vzorce.

POS-gramy typické pro překlady

Pro bližší analýzu byly vybrány následující 4-gramy vyskytující se s největším rozdílem v **překladové beletrii**: J-V-P-V (DIN = 10,57), P-R-P-V (DIN = 9,40) a J-V-P-N (DIN = 8,84). U každé konstrukce uvádím i normalizovaný počet (v ipm) jejích různých realizací v obou souborech, aby bylo zjevné, zda pozorovaný rozdíl v užití vyplývá z několika málo velmi frekventovaných realizací, nebo z celkové větší bohatosti realizací (pro srovnání viz tabulka 4.19).

- J-V-P-V (konjunkce – verbum – pronomen – verbum)
  - počet realizací v překladu: 846,88 ipm, v nepřekladových textech: 765,32 ipm
  - převažující realizací této konstrukce v překladech je *jako by to byl/a/o*, která má takřka dvojnásobnou frekvenci oproti nepřekladovým textům (20,04 ipm oproti 11,80 ipm; DIN = 25,88); časté jsou i ostatní kombinace s *jako by*
  - dále mezi frekventované zástupce této konstrukce patří *když jsem se* + V a *že by se/to* + V, jež však mají srovnatelný výskyt v obou zkoumaných souborech
- P-R-P-V (pronomen – prepozice – pronomen – verbum)
  - počet realizací v překladu: 1 266,42 ipm, v nepřekladových textech: 1 238,81 ipm
  - tuto konstrukci tvoří v naprosté většině n-gramy začínající zvratným zájmenem *se*
  - v překladech výrazně dominuje realizace *se na* + P + *podívat*, která má dvojnásobnou frekvenci oproti nepřekladovým textům (69,39 ipm oproti 33,49 ipm; DIN = 34,90); tento jev může být ovlivněn skladbou textů a jejich tématem
  - specifickým n-gramem je *se na* + P + *zadívat*, který jistě souvisí s předchozím jevem, ovšem preference výrazu *zadívat* je typická pro překlad (12,44 ipm ve 143 textech oproti 3,55 ipm ve 43 nepřekladových textech; DIN = 55,60)
- J-V-P-N (konjunkce – verbum – pronomen – substantivum)
  - počet realizací v překladu: 1 001,14 ipm, v nepřekladových textech: 862,08 ipm
  - nejčastější realizací v překladu je fráze *že je to pravda* s více než dvojnásobnou frekvencí (3,53 ipm v 70 textech oproti 1,62 ipm ve 25 textech; DIN = 37,09)

  - ostatní realizace nejčastěji popisují nějaký fyzický úkon a začínají spojkou *a* (např. *a zapálil si cigaretu*, *a podala mu ruku*); dosahují však pouze frekvence v řádu jednotek (ipm < 0,9), proto nejsou pro další analýzu vhodné

V **odborné literatuře** se v překladu nejvíce vymykají POS-gramy V-P-A-N (DIN = 14,24), V-R-P-N (DIN = 11,31) a P-V-R-N (DIN = 10,15).

- V-P-A-N (verbum – pronomen – adjektivum – verbum)
  - počet realizací v překladu: 1 454,88 ipm, v nepřekladových textech: 1 087,21 ipm
  - vzhledem k vyššímu ranku této konstrukce (71. v nepřekladech a 31. v překladech) dosahují její realizace v textech pouhých několika málo výskytů (v řádu jednotek), srovnání tedy není obecně vypovídající
  - u nejfrekventovanějších realizací se pak zřetelně projevuje téma textu: v překladech najdeme *poznejme svou mikrovlnnou troubu* nebo *vyjmeme je děrovanou naběračkou*, což ukazuje na zastoupení kuchařek v korpusu, kdežto v nepřekladových textech jsou nejčastějšími zástupci fráze *živí se drobnými živočichy* a *objeví se dialogový panel*, které odkazují k odlišným disciplínám (zoologie a informatika)
- V-R-P-N (verbum – prepozice – pronomen – substantivum)
  - počet realizací v překladu: 1 413,43 ipm, v nepřekladových textech: 1 095,55 ipm
  - nejčastější realizací v překladu je *je v tomto případě* a *byl v té době*, dále následují textově specifické fráze typu *nalejeme do ní těsto*
  - podobně jako u předchozí struktury je počet výskytů jednotlivých realizací příliš nízký na jakékoli zobecnění
- P-V-R-N (pronomen – verbum – prepozice – substantivum)
  - počet realizací v překladu: 2 215,75 ipm, v nepřekladových textech: 1 799,38 ipm
  - ačkoli jde o nízké frekvence, ze seznamu realizací v překladu přesto vyčnívají fráze se strukturou P + V + *v pořádku* (6,08 ipm oproti 3,32 ipm; DIN = 29,36), P + V + *ve skutečnosti* (6,52 ipm oproti 3,45 ipm; DIN = 30,79) a P + V + *k dispozici* (14,36 ipm oproti 11,85 ipm; DIN = 9,58).

POS-gramy typické pro nepřeklady

Podobně jako u překladových textů byly pro bližší pohled vybrány první tři POS-gramy s nejvýraznějším rozdílem v užití v obou textových typech. Pro **beletrii** jsou to následující struktury: N-N-R-N (-15,00), A-N-A-N (-13,70) a R-N-A-N (-13,06).

- N-N-R-N (substantivum – substantivum – prepozice – substantivum)
  - počet realizací v nepřekladových textech: 1 053,15 ipm, v překladu: 785,21 ipm
  - 4-gramy tohoto typu v beletrii zpravidla zahrnují vlastní jméno a odkazují tak k postavě v textu nebo jsou jinak textově specifické, např. nejčastější fráze v překladové beletrii *Vyluštění dekodérem na str.* odkazuje k jediné knize, v níž má čtenář hádat a luštit různé záhady. Druhý nejfrekventovanější 4-gram *návrhu zákona o kontrole* lze také najít v jediném textu, který má politické téma
  - v nepřekladové beletrii, kde má tato struktura převahu, stojí za jejím častým výskytem především román Ladislava Fukse *Myši Natálie Mooshabrové* a spojení *paní Mooshabrová u kredence/u plotny/na lavici/v čele/s pohledem*
- A-N-A-N (adjektivum – substantivum – adjektivum – substantivum)
  - počet realizací v nepřekladových textech: 1 178,40 ipm, v překladu: 899,19 ipm
  - u této struktury opět pozorujeme vliv jediného románu *Myši Natálie Mooshabrové* (nejčastější spojení *cizí člověk černý pes*)
  - podobně i v překladové části lze odhalit vliv několika málo titulů (v románu *Na kratším konci ulice* se často opakuje 4-gram *kratším konci Sluneční třídy* a v románu *Matka Noc* K. Vonneguta Jr. dochází k repeticím fráze [*Železná stráž*] *bílých synů americké ústavy*)
- R-N-A-N (prepozice – substantivum – adjektivum – substantivum)
  - počet realizací v nepřekladových textech: 1 564,66 ipm, v překladu: 1 196,17 ipm
  - jako v případě všech struktur, kde převažují substantiva, i zde jsou nejčastější realizace ovlivněny tématem textu (*na jednotce intenzivní péče* ve 14 překladových textech) nebo jediným titulem (*s erbem olomouckého biskupa* z knihy *Olomoucký bestiář*; *na území Spojených států* v knize *Hřbitov vyzvědačů*)
  - pokud se však podíváme na předložky a jejich vazby na začátku této struktury, objevíme obecnější rozdíly: zatímco v nepřekladových textech vede *v podobě* (7,68 ipm vůči 4,76; DIN = -23,42) a často narazíme i na *v rámci* (4,48 ipm vůči 1,93 ipm; DIN = -39,76), v překladu se nejvíce prosazuje lokalizační/temporální spojení *na konci* (5,54 ipm vůči 4,42; DIN = 11,28)

V **odborné literatuře** se v nepřekladových textech nejvíce odlišují POS-gramy N-N-A-N (DIN = -22,32), N-A-N-R (DIN = -23,80) a N-A-N-N (DIN = -22,32).

- N-N-A-N (substantivum – substantivum – adjektivum – substantivum)
  - počet realizací v nepřekladových textech: 2 059,76 ipm, v překladu: 1 283,37 ipm
  - konkrétní realizace 4-gramu odrážejí především žánr (v nepřekladové části to jsou kuchařky: [*špetka*] *soli špetka mletého pepře*, *polovina sáčku vanilinového cukru*; v překladové pak atlasy hub: *houby velikosti špendlíkové hlavičky* či nejrůznější návody: *jehlice Sada krátkých jehlic*)
  - vzhledem k tomu, že zobrazení textu v korpusu zanedbává původní formátování, tvoří tento typ 4-gramu především struktury psané pod sebou, mezi nimiž by jinak stála interpunkce či jiná členicí znaménka textu (např. odrážky)
  - i z počtu realizací, z nichž většina jsou hapaxy, je ovšem patrné, že tato struktura má v nepřekladových textech větší prominenci
- N-A-N-R (substantivum – adjektivum – substantivum – prepozice)
  - počet realizací v nepřekladových textech: 2 117,56 ipm, v překladu: 1 340,25 ipm
  - vzhledem k tomu, že je tento 4-gram zakončen předložkou, můžeme očekávat, že je součástí delšího n-gramu (např. *objemem stavebních prací na* [*jednoho zaměstnance*] nebo *rozvoje cestovního ruchu v* [*České republice*])
  - v nepřekladové odborné literatuře opět zaznamenáváme výrazně větší bohatost realizací této struktury
  - v překladové odborné literatuře zaujímají čelné místo v tomto 4-gramu především struktury s jednotkou míry či váhy na začátku (např. [*3-10*] *cm vysoká bylina s* [*lodyhami*], [*500*] *g čerstvého těsta na* [*těstoviny*])
- N-A-N-N (substantivum – adjektivum – substantivum – substantivum)
  - počet realizací v nepřekladových textech: 1 868,16 ipm, v překladu: 1 221,72 ipm
  - podobně jako u struktury N-N-A-N i v tomto případě tvoří 4-gramy slova, jež by byla v běžném zápise oddělena odrážkami nebo interpunkcí (např. *g mletý pepř cibule* či *máslo hladká mouka sůl*)
  - realizace odkazují opět především k tématu textu (*houby způsobující skvrnitost listů*, *využití obnovitelných zdrojů energie*, *náčelník generálního štábu generál*)

Shrnutí analýzy POS-gramů

Na základě výše provedené analýzy tohoto typu lze nejen poukázat na rozdíly mezi překladovou a nepřekladovou češtinou, ale také upozornit na **výhody a nevýhody tohoto přístupu**. K výhodám zcela jistě patří důraz na syntagmatičnost a kombinatoriku jazykových jednotek, která má v jazyce zásadní roli. Využití kategorie

slovních druhů umožňuje pozorovat tendence ke zhuštěnějšímu vyjadřování (např. prostřednictvím kupení substantiv za sebou) nebo k nižší informační zátěži (např. preferencí struktur s více synsémantiky). V praxi však analýza POS-gramů přináší i nevýhody. Mezi ty nejpodstatnější patří především fakt, že i výrazné rozdíly na úrovni slovnědruhových kombinací mohou být ve skutečnosti způsobeny několika málo frekventovanými realizacemi, jež jsou pevně spjaty s konkrétním dílem nebo žánrem a o podstatě zkoumaného vzorku (překladech či nepřekladech) nic nevypovídají. Tento efekt se týká především POS-gramů, které obsahují autosémantika, především substantiva, jež mnohdy označují postavu v textu nebo jsou součástí pojmenování specifických pro daný text.

Na druhou stranu lze při pohledu na konkrétní realizace POS-gramů vypozorovat obecnější tendence, jež se však neprojevují na úrovni jednotlivých n-gramů, nýbrž na úrovni sekvencí konkrétních slov s některými variabilními pozicemi, např. výše zmíněné P + V + *v pořádku* nebo *se* + *na* + P + *podívat*. Sečteme-li výskyty takové struktury, kterou bychom mohli označit za koligaci, můžeme najít rozdíly i tam, kde u jednotlivých realizací nebyly patrné.

Lze tedy shrnout, že analýza POS-gramů může ukázat na obecné tendence na vyšší rovině abstrakce, kdežto analýza konkrétních realizací (n-gramů) spíše na specifika konkrétního textu (z hlediska tématu, žánru, idiolektu autora). Pokud nás však budou zajímat konkrétní rozdíly mezi zkoumanými vzorky (nikoli jen obecné tendence), může být užitečné se zaměřit na rovinu v tomto smyslu prostřední – na koligační struktury tvořené z části variabilními pozicemi (např. určitými slovními druhy), z části konkrétními realizacemi. Některé takové struktury, jež jsou pro překlady typické, jsou popsány v rámci analýzy n-gramů v části 4.4.

4.2 Simplifikace

Simplifikace patří mezi původní překladové univerzálie (Bakerová 1993: 243). Výzkum simplifikace v překladu probíhal zatím především na angličtině a španělštině (viz dále), v českém prostředí tento jev doposud na velkých datech zkoumán nebyl (s výjimkou pilotní studie Chlumská & Richterová 2014); existují pouze dílčí kvalitativní studie zahrnující simplifikaci jako jednu z S-univerzálií, tedy ve vztahu překladu k originálu (viz např. Polišenská 2010).

Cílem části 4.2 je proto analyzovat simplifikaci v překladové češtině jako T-univerzálii (viz 2.3.4) a s přihlédnutím k dosavadním výzkumům navrhnout a otestovat hypotézy týkající se možných projevů simplifikace v českých překladových textech.

4.2.1 Popis a dosavadní výzkum univerzálie

Simplifikací podle Bakerové označujeme tendenci podvědomě zjednodušovat jazyk nebo obsah sdělení, příp. obojí (1996: 176). Jedním z často kritizovaných rysů definic Bakerové bývá jejich vágní formulace, je tedy třeba dále pracovat s konkrétními hypotézami badatelů. Bakerová v případě této univerzálie vychází z výzkumu R. Van-

derauwerové z roku 1985, která ve svém výzkumu nizozemských románů přeložených do angličtiny objevila tendence překladatelů zaměňovat potenciálně dvojznačná zájmena za přesnější a srozumitelnější, vynechávat opakující se informace a zjednodušovat složitou syntax pomocí rozdělování vět apod. (Baker 1993: 244). Vanderauwerová však k simplifikaci nepřistupuje jako k překladové univerzálii, pouze popisuje tendence při překladu nizozemských románů.

Sara Laviosová, která se simplifikací zabývala dlouhodobě (1996, 1998a, 1998b), naopak vychází z předpokladu, že simplifikace je vlastností všech překladových textů a tuto teorii ověřuje na anglickém srovnatelném korpusu (*English Comparable Corpus, ECC*), který za tímto účelem sestavila. Tento korpus obsahuje milion slov překladové angličtiny (publicistika a beletrie přeložená z několika jazyků) a srovnatelné množství angličtiny nepřekladové. První studii Laviosová provedla na subkorpusu novinových článků, druhou o dva roky později na subkorpusu prozaických děl. Ve svém výzkumu Laviosová nejprve vychází z předpokladu, že překladová angličtina vykazuje z hlediska simplifikace následující charakteristiky (Laviosa 1998a: 103):

1. ve srovnatelném korpusu překladové a nepřekladové angličtiny je slovní zásoba použitá v překladové části chudší než v části nepřekladové,

2. ve srovnatelném korpusu překladové a nepřekladové angličtiny mají překladové texty nižší podíl lexikálních slov vůči všem slovům v korpusu,

3. ve srovnatelném korpusu překladové a nepřekladové angličtiny mají překladové texty nižší průměrnou délku věty než nepřekladové texty.

Konkrétně se tedy Laviosová ve svých studiích zaměřuje na lexikální variabilitu či rozmanitost (*lexical variety*), informační zátěž (*information load*) a délku věty. Tyto prvky pak testuje např. pomocí TTR (viz dále) nebo zkoumání podílu frekventovaných a málo frekventovaných slov a dochází v publicistických textech k tomu, že 108 nejčastějších slov (který nazývá *list head*) v překladové části korpusu zabírá větší část než v případě stejného počtu nejčastějších slov u nepřekladové části. V překladové beletrii je situace obdobná: seznam nejfrekventovanějších slov zahrnuje 82 lemmat, která tvoří 56,2 % celkové velikosti korpusu, kdežto v nepřekladové beletrii je to 87 lemmat čítajících dohromady 51,6 % objemu korpusu (Laviosa 1998b: 6).

Druhou hypotézu ověřuje pomocí testu lexikální hustoty (*lexical density*), který počítá jako procento lexikálních slov (oproti gramatickým slovům) v textu (viz Stubbs 1986: 33). Dochází k tomu, že překladové texty (publicistické i beletristické) vykazují signifikantně nižší ($p < 0{,}005$) lexikální hustotu než nepřekladové. U třetí hypotézy se u obou textových typů výsledky rozcházejí: zatímco v publicistice jsou v překladu věty průměrně kratší, v beletrii je tomu naopak (i po zohlednění netypických textů ve vzorku).

Laviosová na základě těchto studií formulovala tzv. **klíčové vzorce lexikálního úzu** (*core patterns of lexical use*) neboli globální rysy typické pro publicistickou i beletristickou překladovou angličtinu:

1. Translated texts have a relatively lower percentage of content words versus grammatical words (i.e. their lexical density is lower);

2. The proportion of high frequency words versus low frequency words is relatively higher in translated texts;

3. The list head of a corpus of translated texts accounts for a larger area of the corpus (i.e. the most frequent words are repeated more often);

4. The list head of translated texts contain fewer lemmas. (Laviosa 1998b: 8)

Laviosová sama vyzývala k tomu, aby její zjištění posloužila jako výchozí hypotézy pro budoucí výzkum na dalších žánrech a jazycích. Výzkum Laviosové tak inspiroval další badatele, kteří se o podobný výzkum pokusili například na překladové čínštině (Wang & Qin 2010; Xiao 2010). Navzdory zcela odlišné typologii jazyka došli k obdobným výsledkům jako Laviosová: jejich výzkum mimo jiné ukázal, že překladová čínská beletrie obsahuje v průměru delší věty (podobně jako překladová anglická beletrie).

Výzkum simplifikace probíhal v nedávné době také na španělštině (Corpas Pastor, Mitkov, Afzal & Pekar 2008; Ilisei et al. 2010), kde badatelé použili metody zpracování přirozeného jazyka (NLP). V první studii z roku 2008 pracovali s odbornou literaturou (konkrétně s texty z lékařské a technické sféry) a s profesionálními i studentskými překlady. Zkoumali simplifikaci na několika jazykových rovinách: lexikální, stylistické a syntaktické. Na lexikální rovině se zaměřili na lexikální hustotu (*lexical density*) a na lexikální bohatost (*lexical richness*), v rámci stylistiky porovnávali délku vět, distribuci jednoduchých vět a souvětí[8], srozumitelnost textu (*readability*) a diskurzní částice (*discourse markers*) a na úrovni syntaxe zkoumali POS-gramy.

Zde je třeba upřesnit, jakým způsobem autoři studie počítali lexikální hustotu a lexikální bohatost, neboť v literatuře můžeme narazit na několik různých metod výpočtu. Lexikální hustota se zpravidla počítá jako poměr lexikálních (neboli autosémantických) slov ke všem slovům v textu (viz např. Laviosa 1998a: 104 podle Stubbs 1986: 33). Corpas Pastor a jeho kolegové (2008) však pro výpočet používají poměr mezi typy (unikátními slovními tvary v korpusu) a tokeny (všemi výskyty všech slov v korpusu) neboli TTR. Místo obvyklého zprůměrování výpočtu na úseky o délce zpravidla 1000 slov, tzv. sTTR (viz 4.2.3), počítají průměrné hodnoty pro úseky o délce 6 000 vět. V případě TTR se však jako různá slova (typy) počítají i různé tvary jednoho lexému, což může zkreslovat představu o množství a bohatosti slovní zásoby, proto Corpas Pastor s kolegy navrhuje podle svých slov novou míru, tzv. *lexical richness*, v níž jsou do vzorce za typy místo slovních tvarů dosazena lemmata. Tento index je však pouhou variantou TTR, nejde tedy o nový pohled na bohatost lexikonu.

Výsledky výzkumu potvrdily hypotézu o simplifikaci v překladových textech u několika, ne však u všech ukazatelů. Podle studie vykazují španělské překladové

[8]Za souvětí zde byly v souladu s obvyklým pojetím považovány věty s více než jedním slovesem v určitém tvaru.

texty nižší lexikální hustotu i bohatost, jsou jednodušší z hlediska srozumitelnosti a mají kratší věty. Na druhé straně však u nich byl pozorován nižší výskyt jednoduchých vět a diskurzních částic, jež by podle autorů také měly indikovat snahu o větší srozumitelnost textu. Zajímavý se zdá i fakt, že tendence k simplifikaci se nejvíce projevovaly u technicky zaměřených textů a u profesionálních překladů z oblasti lékařství; u studentských překladů pozorovány nebyly. Autoři tento na první pohled překvapivý výsledek nijak nekomentují, ani jej není možné ověřit na korpusu Jerome, neboť ten studentské překlady neobsahuje. Kdybychom však měli spekulovat o příčině tohoto jevu, bylo by snad možné poukázat na obecnou tendenci začínajících překladatelů v cizím jazyce tíhnout k doslovnosti a kopírování originálu, což se s tendencí zjednodušovat věty i lexikon pravděpodobně vylučuje.

Cílem druhého zmiňovaného výzkumu (Ilisei et al. 2010) bylo zjistit, zda je možné na základě vybraných rysů (vycházejících částečně z výsledků výše popsané studie o simplifikaci) automaticky rozlišit překladové texty od nepřekladových. Inspirací pro autory byla studie Baroni & Bernardini 2006, v níž byly za tímtéž účelem využity ukazatele typu distribuce n-gramů synsémantik, lexikální kotvy, výskyt osobních zájmen a adverbií apod. Španělská studie byla opět provedena na textech z oblasti lékařství a techniky a potvrdila, že překlady je možné na základě vybraných rysů automaticky detekovat s úspěšností až 97,62 % (Ilisei et al. 2010: 510). Ukazatele vycházející ze simplifikace přitom výraznou měrou přispěly k úspěšnosti metody. Mezi rysy, které automatický systém nejvíce využil k určení překladovosti textu, patřily lexikální bohatost, poměr gramatických slov k lexikálním slovům, délka věty, délka slova a některé morfologické atributy jako substantiva, pronomina, verba finita, konjunkce a prepozice.

Z poněkud odlišné perspektivy, než kterou nabízí linie kvantitativního synchronního výzkumu, se na tuto univerzálii podívala Paloposkiová (2001), která zkoumala simplifikaci spolu s protichůdnou tendencí k obohacování textu (*enrichment*) ve finštině z diachronního hlediska – u textů z 19. století. Paloposkiová dochází podobně jako ostatní badatelé k závěru, že univerzalita překladových rysů je kontroverzní koncept, který se obtížně testuje a nelze jej patrně vztahovat na data z diachronní perspektivy. Tendence k simplifikaci může podle Paloposkiové souviset s procesem standardizace jazyka, např. v případě dialektových prvků, jejichž používání může být v určitých časových obdobích vnímáno jako méně prestižní a vhodné. Paloposkiová se domnívá, že simplifikace může být vlastností určitého jazyka v určité době, nikoli jeho univerzálním rysem, proto je podle ní potřeba zkoumat tyto rysy v dlouhodobější, diachronní perspektivě a zohlednit při tom např. změny v oblasti překladatelských norem a vývoj celé jazykové a literární situace v příslušné zemi (viz také Even-Zohárova teorie polysystému, 2.2.1).

Jeden z nejnovějších výzkumů simplifikace (a několika dalších univerzálií) byl proveden na textech přeložených z angličtiny do němčiny (Lapshinova-Koltunski 2015), a to jak ve srovnání s německými texty nepřekladovými, tak i s anglickými zdrojovými texty (tedy z pohledu S-univerzálie i T-univerzálie). Tato studie je výjimečná tím, že mezi překladové texty byly zahrnuty nejen texty přeložené profesionálními překladateli, ale také produkty počítačem asistovaného překladu (CAT) a čistě strojového překladu (MT). Výchozí hypotézy nevybočují z linie dosavadního

výzkumu a vycházejí ze srovnání lexikální hustoty a TTR (délka vět byla kvůli srovnání s originální angličtinou ponechána stranou, neboť německé věty se z důvodů odlišné typologie jazyka – velkého počtu složenin – svou délkou už za normálních okolností liší a výsledky by tak byly neporovnatelné). Výsledky ukázaly, že lexikální hustota není v případě zvolených dat dobrým ukazatelem simplifikace, neboť zastoupení autosémantik je ve všech třech zkoumaných souborech srovnatelné (zdrojové anglické texty, překladové a nepřekladové německé). Oproti tomu poměr typů a tokenů, pro jehož výpočet byl použit normalizovaný vzorec pro sTTR (viz 4.2.3), byl u překladových textů nižší než u textů zdrojových i nepřekladových, přičemž nejvyšší hodnotu sTTR vykazují texty přeložené člověkem, následují produkty strojového překladu a konečně „hybridní“ podoba CAT.

Co se týče výzkumu simplifikace v češtině, kromě již zmíněných dílčích příspěvků v podobě diplomových prací byla první pilotní studie provedena na korpusu Jerome (Chlumská & Richterová 2014). Simplifikace projevující se v rozmanitosti slovní zásoby a srozumitelnosti textu zde byla zkoumána z hlediska TTR, lexikální hustoty a délky vět a poukázala mimo jiné na limity standardně používaných testů. Pro ucelený popis simplifikace v českých překladech, jak jej představuje tato práce, byly proto využity další metody, včetně upraveného výpočtu TTR a indexu srozumitelnosti textu. Jejich přehled, včetně hypotéz, je uveden v následující části, která metodologicky vychází z návrhů F. Zanettina (viz s. 37).

4.2.2 Jazykové indikátory a dílčí hypotézy

Vzhledem k vágní definici simplifikace je nezbytné výchozí hypotézu („překladatelé zjednodušují text po jazykové i obsahové stránce“) rozdělit do několika dílčích, testovatelných hypotéz, které vycházejí z vybraných jazykových indikátorů (viz dále tabulka 4.21). Vzhledem k terminologické nekonzistenci v označování určitých rysů textu či způsobů jejich výpočtu (např. lexikální hustota, lexikální bohatost), je však nejprve nutné definovat a rozlišit tyto pojmy tak, jak jsou užívány v této kapitole.

lexikální rozmanitost odpovídá přibližně anglickému *lexical variety*, které má zpravidla obecný význam (pestrost, různorodost slovní zásoby) a neoznačuje současně žádnou konkrétní míru nebo způsob výpočtu; Zanettin (2013: 22) spojuje *lexical variety* s šíří slovní zásoby (*range of vocabulary*)

lexikální bohatost termín je překladem anglického *lexical richness*, které se zpravidla používá v obecném smyslu pro popis pestrosti a bohatosti slovní zásoby, jež se následně testuje prostřednictvím různých testů. Některé z nich jsou závislé na délce textu (např. TTR, lexikální hustota), jiné s ní nekorelují (např. Yuleova charakteristika či Orlovův koeficient Z)[9]; Corpas Pastor, Mitkov, Afzal & Pekar (2008) však tento termín používají jako označení vlastní upravené míry TTR, která se počítá jako poměr lemmat (nikoli různých slovních tvarů jako u TTR) vůči všem tokenům

[9]Přehledně o těchto indexech lexikální bohatosti referují v příloze *Slovníku Karla Čapka* Cvrček, Čermák & Křen M. (2007: 684–690).

lexikální hustota je překladem anglického *lexical density*, které nejčastěji označuje poměr autosémantik (tradičně zahrnujících substantiva, adjektiva, verba, příp. i adverbia) ku všem tokenům v textu (Laviosa 1998a) a bývá tak indikátorem hutnosti textu; v některých pracích však tento termín odkazuje k výpočtu pestrosti a repetitivnosti lexikálních jednotek v textu v podobě TTR (Corpas Pastor, Mitkov, Afzal & Pekar (2008)

V této práci proto platí následující rozlišení: pro obecné označení charakteru slovní zásoby v textu z hlediska její pestrosti či repetitivnosti jsou zde zaměnitelně využívány termíny lexikální bohatost a lexikální rozmanitost (příp. pestrost). Pro označení konkrétního způsobu výpočtu lexikální bohatosti jako poměru lemmat vůči tokenům (Corpas Pastor, Mitkov, Afzal & Pekar 2008) ponechávám pro odlišení název míry v originále jako *lexical richness*. Jako lexikální hustotu neboli *lexical density* zde označuji výpočet poměru autosémantik vůči všem tokenům v textu (Laviosa 1998a).

Jak vyplývá z přehledu dosavadního výzkumu o simplifikaci v překladu, předmětem analýz, včetně této, bývá především **slovní zásoba překladových textů** a jejich skladba, především s ohledem na **srozumitelnost** a snadnost čtení **textu**. To má několik důvodů – předně lze předpokládat, že tendence zjednodušovat text se bude projevovat právě na těchto dvou rovinách, neboť připouštějí největší variabilitu (např. na úrovni morfologie nelze většinu prostředků nahrazovat jinými, jednoduššími). Zjednodušování zde můžeme chápat jednak jako podvědomou snahu překladatele vycházet cílovému čtenáři vstříc tím, že text učiní snazší na čtení (např. pomocí kratších vět či frekventovanějších slov), jednak jako nezamýšlenou tendenci překladatele volit frekventovanější, obecnější a snáze vybavitelné jazykové prostředky (viz již Levý 1983: 44). Je však třeba současně upozornit na to, že tyto indikátory nemusejí samy o sobě potvrzovat simplifikační tendence. Vezmeme-li např. samotnou délku věty, pouhý fakt, že je kratší, ještě nutně neznamená, že bude zároveň jednodušší na porozumění; může obsahovat mnoho za sebou nakupených autosémantik, cizích slov nebo netypických lexikálních či morfologických prostředků (např. přechodník), které od čtenáře vyžadují větší úsilí při porozumění. Jednou z možností, jak riziko těchto případů eliminovat, je zkoumat více faktorů najednou (více dále a v kapitole 4.3).

Dalším důvodem, proč badatelé zkoumají simplifikaci právě na těchto jazykových indikátorech, je bezesporu fakt, že je lze poměrně snadno operacionalizovat, kvantifikovat a měřit. Na druhou stranu je třeba brát v potaz, že nejčastěji používané testy (jako je TTR nebo *lexical richness*) jsou často náchylné na délku textu, nezohledňují syntagmatiku jazyka a jejich využití i vypovídací hodnota jsou tak omezené. Výběrem různých testů pro **výzkum simplifikace v překladové češtině** jsem se pokusila tato rizika minimalizovat a zohlednit je při interpretaci výsledků. Shrnutí vybraných rysů a metod výpočtu uvádí následující tabulka 4.21. Po ní následují dílčí hypotézy.

SIMPLIFIKACE	jazykové indikátory	formální operátory
lexikon	bohatost slovní zásoby	poměr mezi typy a tokeny (zTTR)
		Yuleův koeficient K
	hutnost vyjadřování	lexikální hustota (*lexical density*)
		frekvenční špička (*list head*)
syntax	srozumitelnost textu	délka vět
		index srozumitelnosti (*readability index*, *RI*)

Tabulka 4.21: Simplifikace – operacionalizace pro výzkum překladové češtiny

Bohatost slovní zásoby a hutnost vyjadřování

Bohatost slovní zásoby se dá kvantitativně zkoumat mnoha různými způsoby; jak jsme však viděli u zmíněných zahraničních prací, většina z nich vychází z poměrů mezi různými typy (a to jak na úrovni lemmat, tak wordů) a tokeny, přičemž platí, že čím vyšší je poměr různých typů v korpusu oproti všem tokenům, tím je slovní zásoba považována za bohatší. Pro účely této studie byla jako hlavní kvantitativní ukazatel lexikální bohatosti a repetitivnosti slovní zásoby v textech vybrána upravená míra TTR nazvaná zTTR (viz 4.2.3), která je univerzálně použitelná při srovnání textů různé délky. Pro doplnění byl pro hlavní zkoumané skupiny (překladovou a nepřekladovou beletrii a překladovou a nepřekladovou odbornou literaturu) vypočítán i Yuleův koeficient K, který rovněž indikuje repetitivnost textu (čím nižší je, tím bohatší je lexikon), ale nevychází přitom z poměru typů a tokenů.

Ukazatelem hutnosti vyjadřování na úrovni lexikonu může být srovnání frekvenční špičky (*list head*) ve zkoumaných souborech, především z hlediska toho, jak velkou část korpusu v počtu tokenů zahrnují. Hypotéza zní, že čím větší procento slova z frekvenční špičky v korpusu celkem tvoří, tím více je text považován za simplifikovaný, neboť díky využití vysoce frekventovaných výrazů, které většinou spadají do skupiny synsémantik, se text snáze čte a informační zátěž textu je nižší. Opět pro doplnění uvádím i výsledky míry *lexical density*, která zohledňuje výskyt autosémantik (substantiv, adjektiv, verb a adverbií) v textech.

Na základě těchto východisek je proto možné formulovat dílčí **hypotézy** o simplifikaci z hlediska lexikální bohatosti a hutnosti vyjadřování:

H_0: Poměr mezi typy a tokeny (měřený pomocí zTTR), Yuleův koeficient K, výskyt autosémantik (měřený pomocí *lexical density*) a objem slov z frekvenční špičky v korpusu se v překladové a nepřekladové češtině neliší.

H_1: Poměr mezi typy a tokeny (měřený pomocí zTTR), výskyt autosémantik (měřený pomocí *lexical density*) je v překladové češtině nižší a Yuleův koeficient K a objem slov z frekvenční špičky v korpusu je v překladové češtině vyšší.

Srozumitelnost textu

Ačkoli se srozumitelností textu zcela jistě souvisejí i ukazatele lexikální bohatosti či hutnosti uvedené výše, v této části je kladen důraz především na syntax. Z hlediska syntaxe může být jedním z ukazatelů simplifikace nižší průměrná délka věty. Větou

se zde myslí souvětí (*sentence*, nikoli *clause*) a její délkou počet slov mezi jedním a druhým koncovým interpunkčním znaménkem (tečka, otazník, vykřičník; nikoli čárka nebo středník).

Pro výpočet srozumitelnosti textu se také používají tzv. indexy srozumitelnosti (*readability index*, RI), které zpravidla zohledňují průměrnou délku vět a slov v dokumentu. Pro tuto studii byl zvolen klasický test ARI (*Automated Readability Index*, viz Smith & Senter 1967).

Hypotézy týkající se srozumitelnosti textu tedy znějí takto:

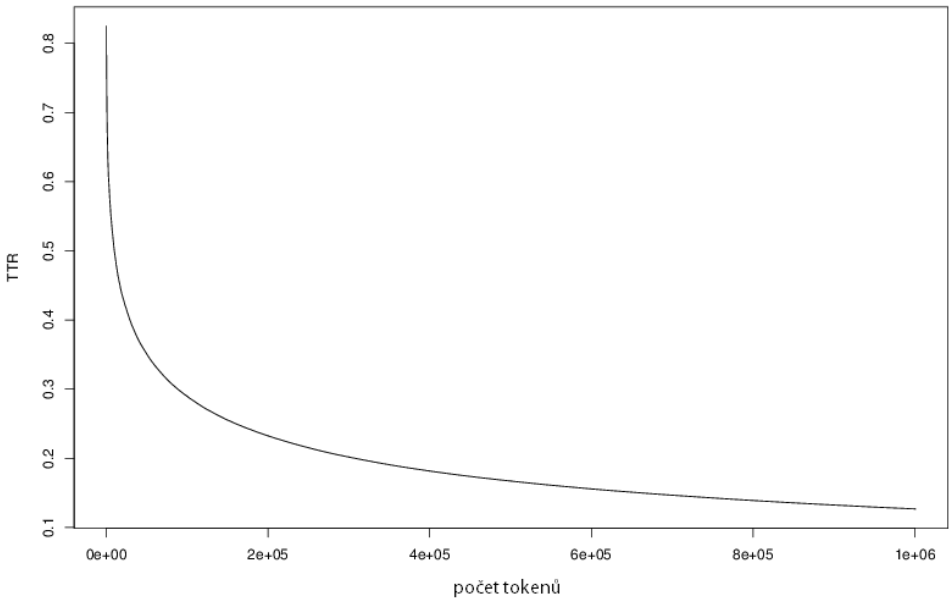
H_0: Průměrná délka věty a hodnota indexu srozumitelnosti (měřená testem ARI) se v překladové a nepřekladové češtině neliší.

H_1: Průměrná délka věty a hodnota indexu srozumitelnosti (měřená testem ARI) je v překladové češtině nižší než v češtině nepřekladové.

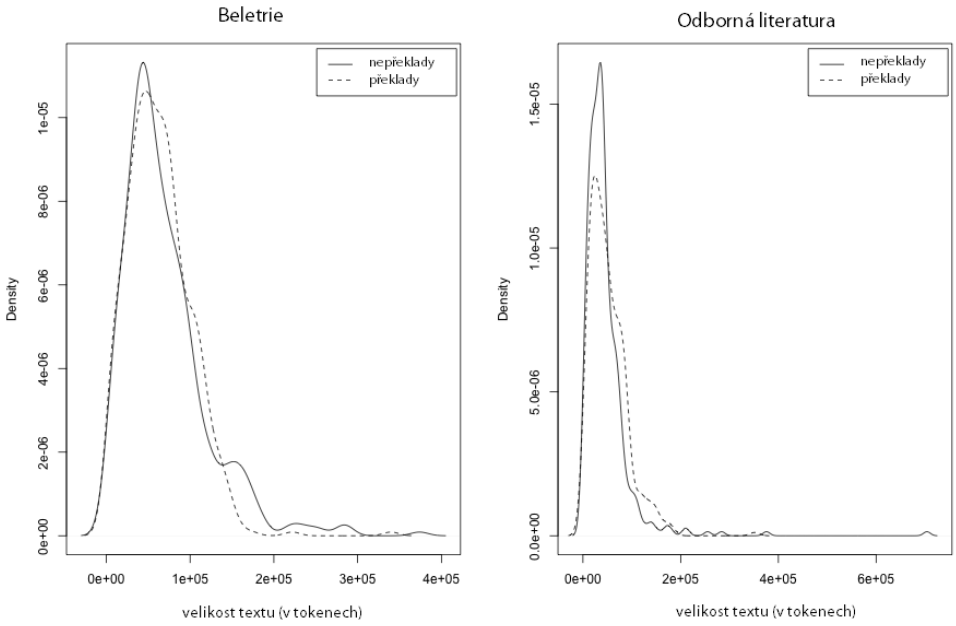
4.2.3 Popis formálních operátorů a srovnání výsledků

Tato část shrnuje výsledky testování hypotéz prostřednictvím vybraných formálních operátorů (viz 4.21). Tam, kde je to nezbytné, je věnována větší pozornost podrobnějšímu popisu a zdůvodnění metody (např. u zTTR).

Poměr mezi typy a tokeny (zTTR)



Obrázek 4.22: TTR v textech různé délky



Obrázek 4.23: Srovnání překladových a nepřekladových textů v korpusu Jerome z hlediska velikosti

Míra TTR je pravděpodobně nejpoužívanější mírou pro popis a srovnání lexikální bohatosti dvou a více textů či korpusů; hlavním důvodem je patrně snadnost jejího výpočtu, kdy se počet různých slov (typů) vydělí počtem všech slov (tokenů). Její velkou nevýhodou je však citlivost na velikost textu. Čím delší je text, tím je hodnota TTR nižší, jak názorně ukazuje graf 4.22.[10]

Míru TTR tak lze bez jakékoli korekce použít pouze v případě textů srovnatelné délky. Pokud však pracujeme s korpusem celých textů, a nikoli vzorků o stejné velikosti, není snadné tomuto požadavku dostát. Ačkoli v případě korpusu Jerome dochází ke zjevné korelaci mezi velikostmi textů v překladové a nepřekladové části ($r = 0{,}9387$)[11], u určitých velikostí se tyto populace přece jen liší (v délce 30 000–39 000 tokenů převažují nepřekladové texty, kdežto v délce 70 000–79 000 tokenů texty překladové), viz graf 4.23.

Velikost textu však není jediným faktorem, který ovlivňuje výši TTR – tento index vykazuje jiné hodnoty i u textů různých textových typů (viz výše graf 4.1). Z toho vyplývá, že i texty stejné délky, ovšem různého textového typu (beletrie, odborná literatura či publicistika), budou dosahovat zcela odlišných hodnot TTR.

Ve snaze odstranit nesrovnatelnost TTR u různě dlouhých textů byla představena upravená verze této míry, tzv. **standardizované TTR** neboli **sTTR** (Scott 2006). Výpočet sTTR není založen na poměru tokenů a typů v *celém* textu,

[10]Grafy, postup i výsledky v této části o zTTR jsou převzaty z článku Cvrček & Chlumská (v tisku), který vyjde v roce 2015 v časopise *Russian Linguistics*.

[11]Relativně vysoký korelační koeficient můžeme připisovat tomu, že všechny texty psané česky (ať už původní či ne) mají podobnou distribuci délek, jak ukazuje graf 4.23.

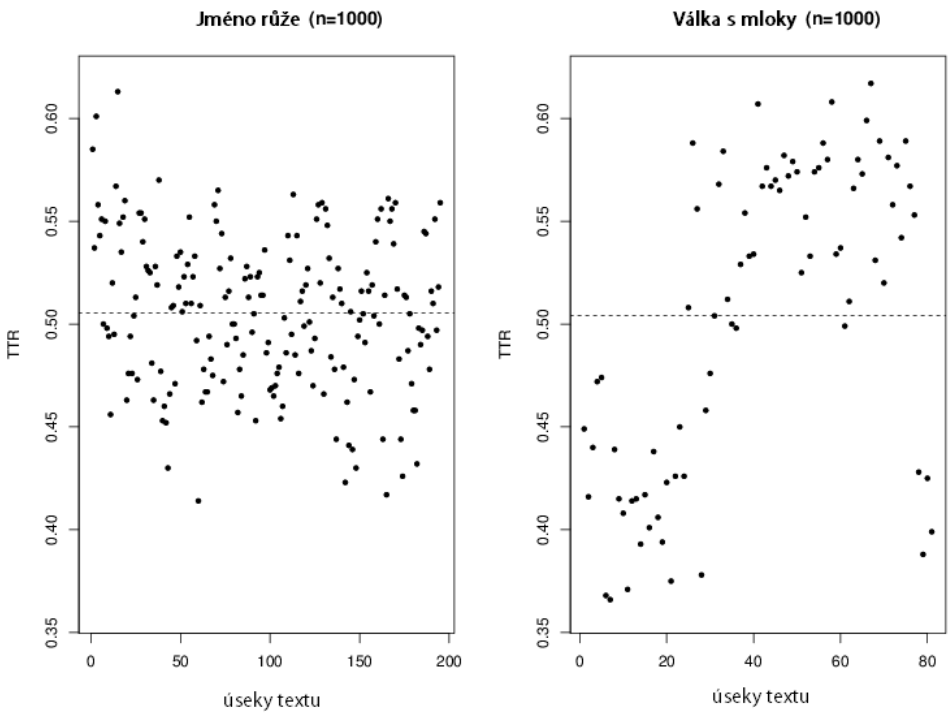
nýbrž vychází z průměru hodnot TTR pro úseky textu o arbitrárně zvolené délce n slov (obvykle však 1 000 slov), jež po sobě následují. Výsledná hodnota sTTR je tak normalizována a může být v této podobě použita při srovnávání různě dlouhých textů.

Ačkoli sTTR řeší problém citlivosti TTR na délku textu, přináší s sebou jiné problematické rysy, kvůli kterým byla v tomto výzkumu upřednostněna varianta výpočtu zTTR (viz dále). Rozdělováním textu na stejně velké úseky a průměrováním naměřených hodnot dochází k zanedbání *vnitrotextové variability*. Upravená míra sTTR vychází z předpokladu, že všechny úseky jsou si rovny nebo přinejmenším podobny, co se týče frekvenční distribuce slov. To však většinou není pravda. Jeden úsek textu o délce 1 000 slov bude zpravidla obsahovat určité množství synsémantik (jako jsou předložky a spojky), ovšem zrovna nemusí zahrnovat některá méně frekventovaná autosémantika, jež však hrají důležitější roli při určování velikosti lexikonu a jeho bohatosti. V důsledku toho lze říci, že míra sTTR nadhodnocuje běžná (nejčastěji synsémantická) slova, kdežto méně častá autosémantika podhodnocuje.

Dalším problematickým rysem sTTR je fakt, že počítá s aritmetickým průměrem ze všech hodnot naměřených v jednotlivých úsecích textu, jakkoli mohou být některé z nich netypické či dokonce extrémní. Dva texty tak mohou mít totožnou hodnotu sTTR, ale rozptyl hodnot pro jednotlivé úseky se může podstatně lišit. Příkladem výrazně odlišné vnitrotextové variability mohou být například následující dva romány: *Jméno růže* Umberta Eca (v českém překladu) a *Válka s mloky* Karla Čapka. Ačkoli jde zcela jistě o z mnoha hledisek rozdílné texty, oba by byly pravděpodobně zařazeny a porovnávány v rámci stejného textového typu beletrie. Graf 4.24 zobrazuje prostřednictvím teček hodnoty TTR pro jednotlivé úseky o délce 1 000 slov, přičemž přerušovaná čára představuje průměrné TTR daného textu, tedy sTTR. Hodnoty TTR pro jednotlivé úseky Ecova románu se v celém románu pohybují v podobné vzdálenosti od průměru, kdežto Čapkův text vykazuje jiný vzorec – v úsecích na začátku a na konci románu jsou hodnoty TTR nižší než uprostřed. K lepší vizualizaci rozdílu v rozptylu je použit krabicový graf[12] (viz 4.25).

Oba texty mají takřka identickou průměrnou hodnotu (sTTR Eco = 0,5054 and sTTR Čapek = 0,5041), z hlediska sTTR by tak byly vyhodnoceny jako stejně či obdobně lexikálně bohaté. Rozdíl v rozptylu je u této míry zcela zanedbán. Zatímco Ecův román má poměrně rovnoměrnou distribuci typů v celém textu a většina úseků dosahuje obdobných hodnot TTR (variační koeficient je 7,3 %), Čapkův román, jenž bývá označován za román-fejeton (tedy jakýsi pomezní žánr), nemá podobně pevnou strukturu – text nabývá v průběhu různých podob, od žurnalistického stylu až po klasickou beletrii. To se odráží ve dvojnásobně vyšším variačním koeficientu (14,8 %). Index sTTR tak nezohledňuje vnitřní dynamiku textu, která může mít

[12]Boxplot neboli krabicový graf je jedním ze způsobů grafické vizualizace numerických dat pomocí jejich kvartilů. Střední část diagramu je shora ohraničena 3. kvartilem, zdola 1. kvartilem a mezi nimi se nachází linie vymezující medián. Linie vycházející ze střední části diagramu kolmo nahoru a dolů vyjadřují variabilitu dat pod prvním a nad třetím kvartilem. Odlehlé hodnoty, tzv. outliery, jsou vykresleny jako jednotlivé body a označují hodnoty, jež se výrazně vymykají trendu ve vzorku; dosahují více než 1,5násobku hodnoty horního kvartilu v grafu – horní vodorovná čára – nebo hodnoty 1,5krát menší než hodnota spodního kvartilu – dolní vodorovná čára.



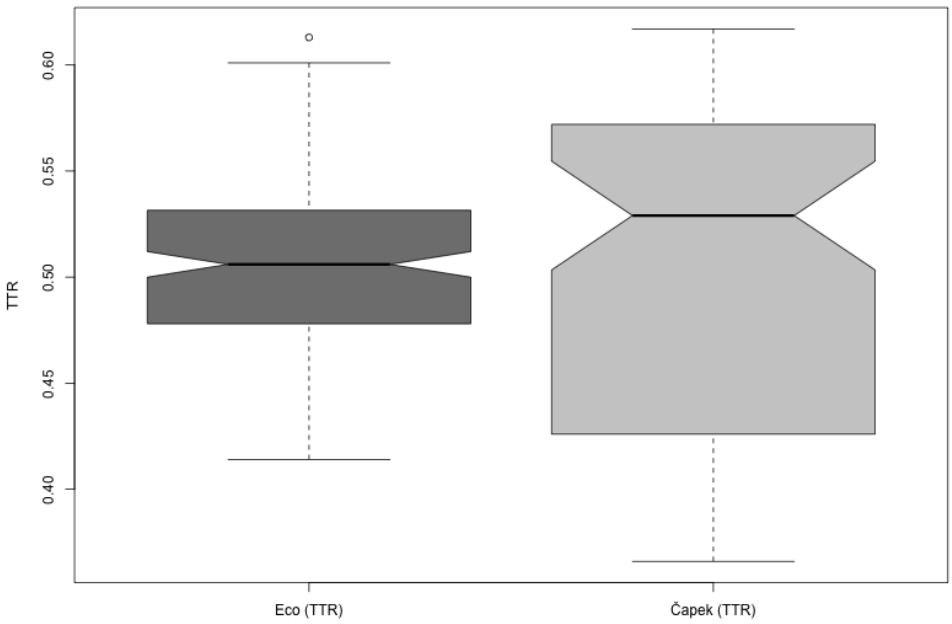
Obrázek 4.24: Proměny TTR v rámci jediného textu na příkladu románů *Jméno růže* a *Válka s mloky*

různé tempo podle toho, jak rychle jsou v textu představována nová témata a nové postavy (a slova s nimi spojená). Pokud má text nestejnoměrnou dynamiku, bude mít pravděpodobně větší rozptyl hodnot TTR v jednotlivých úsecích a tím méně spolehlivou hodnotu sTTR.

Vnitrotextová variabilita (zde znázorněna pomocí rozptylu hodnot TTR) však není jediným faktorem, který míru sTTR diskvalifikuje jako spolehlivý ukazatel lexikální bohatosti. Problematická je i metoda dělení textu na jednotlivé úseky. Zprůměrováním hodnot TTR nelze odhalit, zda mají tyto po sobě jdoucí úseky podobný lexikon. Představme si hypotetickou situaci, v níž mají dva texty identickou délku a identické sTTR (slova zde nahrazují písmena a texty jsou rozděleny do stejně velkých úseků o třech slovech/písmenech):

TextA	a b c	a b c	a b c	a b c	a b c	a b c	a b c	a b c
TextB	a b c	d e f	g h i	j k l	m n o	p q r	s t u	v w x

V obou hypotetických textech se hodnoty TTR pohybují blízko průměru (3 typy na 3 tokeny), avšak v textu A jsou úseky textu identické a text jako celek je velmi repetitivní. Oproti tomu text B, v němž také najdeme 3 typy na 3 tokeny v každém úseku, se skládá z úseků s různými typy (takovým textem by mohla být třeba kniha tematicky nepodobných povídek, v nichž vystupují pokaždé jiné postavy). Navzdory



Obrázek 4.25: Vnitrotextová disperze v románu *Jméno růže* a *Válka s mloky*

intuitivnímu předpokladu, že text B je lexikálně bohatší, budou hodnoty sTTR pro oba texty stejné (sTTR = 3). Tento paradox vyplývá z toho, že u sTTR je lexikální bohatost textu pojímána jako průměrná hodnota jeho částí, ačkoli v případě uvedeného modelu má text A celkově pouhé 3 různé typy, kdežto text B 24 typů.

Ačkoli je sTTR mezi badateli stále velmi využívaná, vykazuje nesporné nedostatky. Proto byla v této práci pro výzkum lexikální bohatosti překladové češtiny použita nová **varianta TTR nazvaná zTTR** (Cvrček & Chlumská, v tisku), která vychází ze srovnání naměřených TTR hodnot s referenčními hodnotami. Aby bylo zTTR srovnatelné pro texty nestejné délky, musejí být tyto referenční hodnoty dvojího typu: jednak průměrné TTR populace textů dané délky, jednak směrodatná odchylka (dále značena jako s) TTR v rámci stejné populace textů. Porovnáme-li pak naměřené TTR textu s distribucemi TTR ve velkém vzorku stejně velkých textů, zjistíme, zda a jak se jeho hodnota vymyká. V ideálním případě by k získání referenčních hodnot byl potřeba velký a reprezentativní vzorek textů pro každou možnou délku textu. Z tohoto vzorku by se poté vypočítala průměrná hodnota a směrodatná odchylka, která by (jako očekávaná hodnota) sloužila pro porovnání s pozorovanou hodnotou TTR zkoumaného textu. Tak by bylo možné zjistit, jak ne/typická je hodnota TTR ve srovnání s ostatními texty stejné délky.

Tento postup je samozřejmě v praxi neproveditelný, neboť žádný korpus, byť by byl sebevětší, nebude obsahovat dostatečný počet referenčních textů pro každou délku textu. Je proto nezbytné zvolit alternativní metodu výpočtu referenčních hodnot: simulovat populaci textů dané délky tím, že rozdělíme referenční vzorek na

úseky o požadované délce. Referenční vzorek b tomto případě tvořily texty z korpusů SYN2000, SYN2005 a SYN2010, přičemž ze vzorku byly vyloučeny texty, které jsou součástí korpusu Jerome. Vzhledem k tomu, že pilotní experimenty prokázaly, že TTR se liší nejen v závislosti na délce textu, ale i na textovém typu (viz graf 4.1), byl tento faktor zohledněn i při výpočtu referenčních hodnot – pro beletrii a odbornou literaturu byly hodnoty počítány pouze z textů stejného textového typu. Tak byly získány referenční hodnoty pro každou velikost textu (z rozmezí velikostí obsažených v korpusu Jerome) s tím, že u kratších textů byl zvolen krok o délce 100 tokenů a u delších 500 tokenů, viz 4.26.

Velikost textu (v tokenech)	beletrie průměrné TTR	s	odborná literatura průměrné TTR	s
500	0,5933	0,05566	0,6156	0,05930
600	0,5778	0,05495	0,5990	0,05925
700	0,5647	0,05441	0,5852	0,05930
–	–	–	–	–
199 000	0,1657	0,02667	0,1672	0,02817
199 500	0,1655	0,02680	0,1671	0,02816
200 000	0,1650	0,02684	0,1670	0,02846

Tabulka 4.26: Příklad výpočtu zTTR pro různě dlouhé texty

Na základě obdobné tabulky byl samotný index zTTR pro každý zkoumaný text počítán následujícím způsobem:

zTTR = (naměřené TTR − průměrné TTR v dané délce) / s

Jak je patrné už z jejího názvu, tato míra je inspirována mírou z-score, avšak výpočet zTTR nelze považovat za klasickou normalizaci. Vzhledem k tomu, že data použitá pro výpočet referenčních hodnot nemají normální rozdělení, nelze zTTR interpretovat jako z-score (jinými slovy hodnoty nekorespondují s percentily populace jako u z-score, kdy např. z $\leq$ -2 odpovídá 2,3 % případů). Index zTTR však přesto plní svůj účel a umožňuje srovnání textů nestejné délky, neboť uvádí vzdálenost mezi naměřeným TTR zkoumaného textu a průměrným TTR (textů stejné délky) v počtu směrodatných odchylek. Hodnoty zTTR jsou porovnatelné pouze mezi sebou, slouží tedy jako ukazatel při srovnání textů, nikoli jako reprezentativní hodnota sama o sobě. Platí přitom následující:

zTTR = 0 ... průměrná hodnota

zTTR < 0 ... podprůměrná hodnota (lexikálně chudší)

zTTR > 0 ... nadprůměrná hodnota (lexikálně bohatší)

Názornější ukázkou rozdílu v přístupech sTTR a zTTR může být porovnání obou hodnot u již zmíněných románů *Jméno růže* a *Válka s mloky*. Tabulka 4.27 ukazuje, že zatímco hodnota sTTR je pro obě díla takřka totožná (a indikuje, že díla mají srovnatelnou lexikální bohatost), hodnoty zTTR se s ohledem na vnitrotextovou variabilitu a dynamiku díla liší: Ecův román se pohybuje pod průměrem hodnot typických pro tuto délku textu, kdežto Čapkův text se jeví jako nadprůměrný.

Text	tokeny	typy	TTR	sTTR	zTTR
Eco	195 679	28,976	0,1481	0,5054	**-0,7011**
Čapek	81 758	18,394	0,225	0,5041	**0,3523**

Tabulka 4.27: Rozdílné hodnoty TTR, sTTR a zTTR na příkladu románů *Válka s mloky* a *Jméno růže*

Index zTTR se tak snaží kombinovat výhody TTR a sTTR: není závislý na délce textu, zohledňuje vnitrotextovou variabilitu a informační dynamiku zkoumaného textu a zároveň umožňuje porovnávat texty různých textových typů, neboť referenční hodnoty jsou počítány pro každý typ zvlášť. Má samozřejmě i své nevýhody, z nichž patrně tou nejzávažnější je náročný proces získání referenčních hodnot, jež budou pro každý jazyk specifické (vzhledem k typologickým rozdílům mezi jazyky).

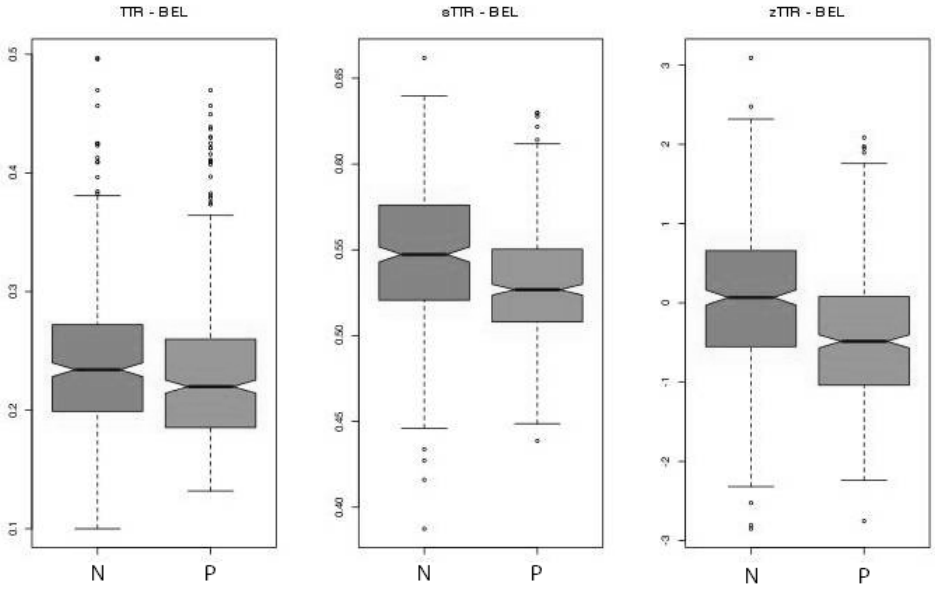
Vypočítáme-li pro srovnání všechny zmíněné indexy (TTR, sTTR i zTTR) pro **překladové a nepřekladové texty v korpusu Jerome**, dostaneme výsledky, které jsou přehledně zobrazeny v grafech 4.28 a 4.29. Jak již bylo řečeno, číselné hodnoty jednotlivých indexů nelze porovnávat mezi sebou, vždy pouze v rámci porovnávaných souborů, proto zde výsledky zobrazuji pomocí krabicového grafu. Z grafu 4.28 vyplývá, že v případě beletrie ukazují všechny tři indexy podobný výsledek – překladové texty vykazují nižší poměr typů a tokenů než texty nepřekladové. U odborné literatury (graf 4.29) je situace poněkud odlišná – rozdíl mezi překlady a nepřeklady je patrný pouze na základě hodnot TTR a zTTR, hodnoty sTTR vychází srovnatelné. To potvrzuje i tabulka 4.30, která shrnuje výsledky testů statistické signifikance pozorovaných rozdílů. Pomocí Mann-Whitneyho U testu (provedeného v R pomocí funkce wilcox.test) byly potvrzeny všechny rozdíly jako statisticky signifikantní s výjimkou hodnoty sTTR u odborné literatury.

Textový typ	Wilcox (U-test)		p-hodnota	
	sTTR	zTTR	sTTR	zTTR
beletrie	111 974,5	115 993	$p < 0{,}001$	$p < 0{,}001$
odborná	56 876,5	67 333	$\mathbf{p = 0.6775}$	$p < 0{,}001$

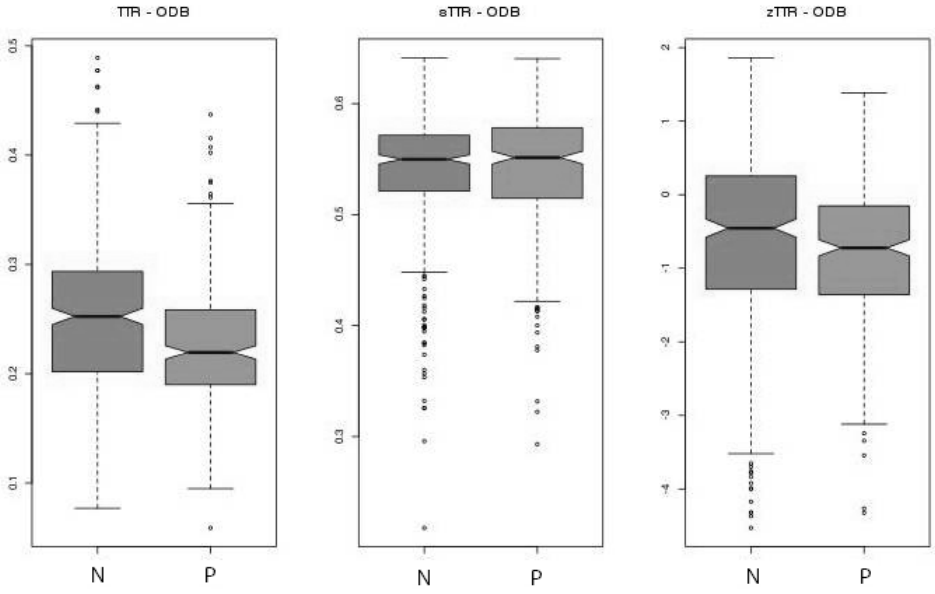
Tabulka 4.30: Výsledky testování statistické signifikance pro hodnoty sTTR a zTTR

Ačkoli zde s ohledem na výše uvedenou argumentaci budou v této práci považovány za **relevantní pouze výsledky zTTR**, výsledky zbylých dvou měr si zaslouží komentář. Jak je možné, že i přes odlišné přístupy vykazují hodnoty v beletrii podobné výsledky? A proč tomu tak není i v odborné literatuře? TTR přináší platné výsledky pouze tehdy, jsou-li délky textů ve zkoumaných souborech srovnatelné, což u korpusu Jerome přibližně platí (viz 4.23). Oproti tomu sTTR lze jako spolehlivý ukazatel použít pouze v tom případě, kdy je informační dynamika (tedy distribuce typů) v textech srovnatelná, což zjevně platí pro beletrii v korpusu Jerome, ovšem nikoli pro odbornou literaturu.

Vzhledem k tomu, že samotná statistická signifikance sama o sobě nevypovídá o výzkumné relevanci rozdílu (viz výše 4.1.1), byla pro hodnoty zTTR vypočítána také síla účinku (*effect size*) prostřednictvím Wendtova vzorce (jenž vychází z Man-



Obrázek 4.28: Srovnání TTR, sTTR a zTTR – beletrie



Obrázek 4.29: Srovnání TTR, sTTR a zTTR – odborná literatura

Whitneyho U testu). Podle výsledků (r = 0,326 pro beletrii a r = 0,16 pro odbornou literaturu) můžeme hovořit o malé až střední síle účinku, což znamená, že překladové texty mají skutečně tendenci vykazovat nižší hodnoty zTTR, avšak rozdíly nejsou příliš výrazné.

Závěrem je třeba dodat, že index zTTR sice řeší nedostatky míry TTR i sTTR, přesto se jedná jen o jeden z možných ukazatelů bohatosti lexikonu v textech. Je nezbytné jej doplnit o další míry a interpretovat jejich výsledky jako celek, chceme-li přesvědčivě hovořit o projevech simplifikace v překladové češtině.

Yuleův koeficient K

Další z měr, která se používá pro vyhodnocení lexikální bohatosti, je Yuleova charakteristika neboli koeficient K (Yule 1944). Tato míra charakterizuje distribuci slov v textu (jinými slovy repetitivnost textu) a na rozdíl od TTR je považována za nezávislou na délce textu. Koeficient K byl vypočítán pro čtyři hlavní skupiny textů: nepřekladovou a překladovou beletrii a nepřekladovou a překladovou odbornou literaturu (nikoli tedy pro každý text zvlášť a následně průměrován jako v případě ostatních testů). Pro výpočet byla, na rozdíl od originálního Yuleova K, které využívá jen substantiva, jako vstupní data použita všechna autosémantika (substantiva, adjektiva, verba a adverbia), neboť ta jsou považována za hlavní ukazatele bohatosti lexikonu. Výsledky shrnuje tabulka 4.31.

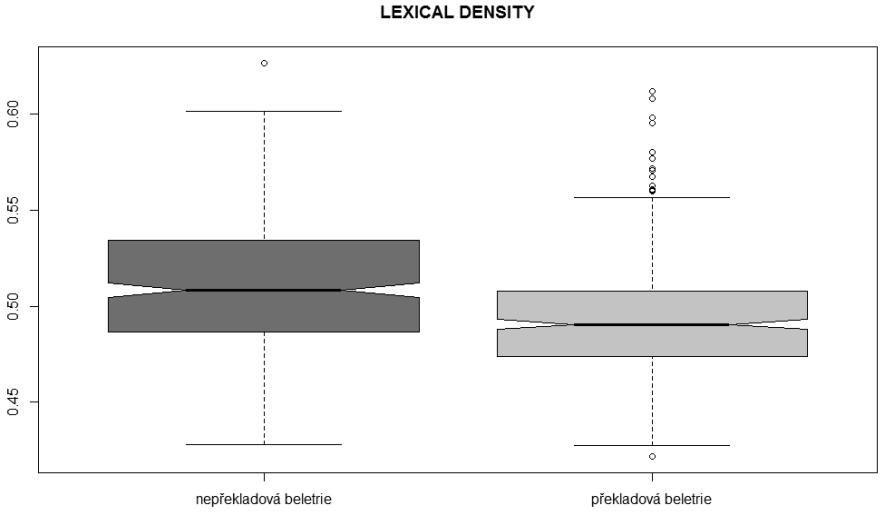
Koeficent K	nepřeklady	překlady
beletrie	0,40418	0,39535
odborná	0,61848	0,62183

Tabulka 4.31: Yuleova charakteristika K

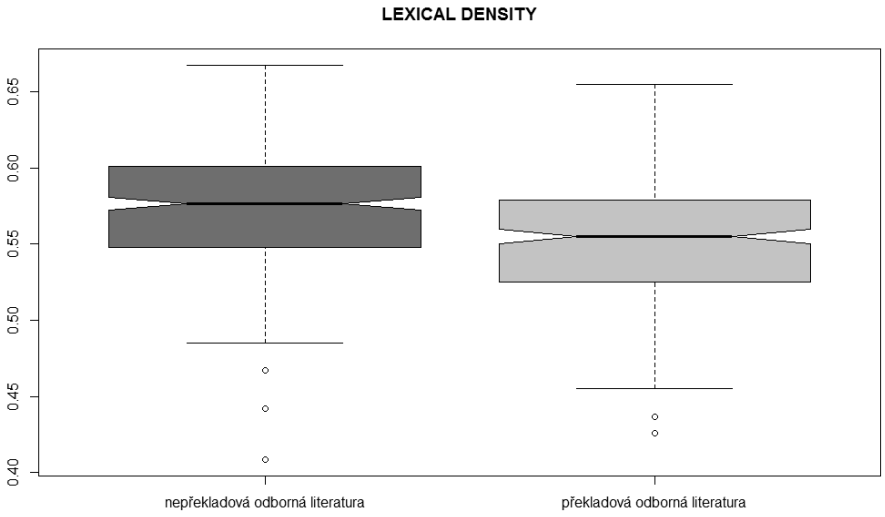
Čím vyšší hodnoty koeficient K dosahuje, s tím větší pravděpodobností můžeme očekávat, že se v textu zopakuje totéž slovo; naopak čím nižší je, tím menší je repetitivnost. Na rozdíl od zTTR nepotvrzuje Yuleova charakteristika trend nižší lexikální rozmanitosti překladových děl, neboť výsledné hodnoty jsou takřka totožné. Vyjádříme-li rozdíl mezi zkoumanými soubory procentuálně jako odchylku od průměru, vyjde nám $\pm 1,1$ % pro beletrii (jako více repetitivní vychází nepřekladová) a $\pm 0,27$ % pro odbornou literaturu (jako více repetitivní vychází překlady). Rozdíly jsou prakticky zanedbatelné; lze tedy konstatovat, že tento test nižší lexikální bohatost u překladů nepotvrdil.

Test *lexical density* (LD)

Dalším z projevů simplifikace může být nižší hutnost vyjádření, která se na úrovni lexikonu může projevovat nižším výskytem autosémantik. Test lexikální hustoty (*lexical density*) je proto založen na poměru autosémantik vůči všem slovům v textu. Vzhledem k tomu, že vzorec počítá pouze s tokeny (všemi výskyty všech autosémantik/slov), nikoli typy (různými autosémantiky/slovy), neměla by případná různá velikost textu ovlivňovat výsledek jako v případě TTR.



Obrázek 4.32: Srovnání lexikální hustoty – beletrie



Obrázek 4.33: Srovnání lexikální hustoty – odborná literatura

Graf 4.32 shrnuje rozdíl v lexikální hustotě beletristických textů. Podobně jako v případě zTTR i zde dosahují nepřekladové texty vyšších hodnot, tzn. podíl autosémantik je v nich vyšší. Obsahují v průměru 51,09 % autosémantik, kdežto překladové texty 49,28 %. Vizualizace pomocí krabicového grafu ukazuje, že rozdíl není jen v prostřední hodnotě, ale také v rozptylu hodnot – překladové texty jsou si jako celek podobnější, byť v souboru najdeme více odlehlých hodnot. Podobností překladových textů se zabývá další z univerzálií nazývaná *levelling-out* nebo konvergence (viz dále 4.3).

U odborné literatury (graf 4.33) je trend obdobný: překladové texty mají v průměru nižší lexikální hustotu (55,18 % autosémantik) než nepřekladové (57,18 %); rozdíl je u beletrie i odborné literatury statisticky signifikantní ($p < 0{,}001$ na základě Mann-Whitneyho U-testu). Lze také konstatovat, že odborné texty jsou bez ohledu na zdrojový jazyk lexikálně hutnější než texty beletristické (obsahují min. 55 % autosémantik). Výrazný rozdíl v rozptylu však u odborné literatury nepozorujeme.

Frekvenční špička

Podíváme-li se na to, kolik místa v textu zabírají slova z frekvenční špičky, dostaneme další z možných ukazatelů lexikální hutnosti zkoumaného textu. Výchozí hypotéza zní, že ve frekvenční špičce se vyskytují především synsémantika a velmi krátká slova, která přispívají ke srozumitelnosti textu, a proto platí, že čím větší je jejich zastoupení (v tokenech) v textu, tím je text lexikálně jednodušší a snazší na porozumění. Následující tabulky 4.34 a 4.35 ukazují zastoupení frekvenční špičky – v podobě prvních deseti, sta a tisíce nejfrekventovanějších lemmat (rank) – ve zkoumaných souborech.

Z výsledků srovnání vyplývá, že v překladech u obou textových typů zabírají slova z frekvenční špičky větší část textu než u nepřekladových děl. Vezmeme-li jako příklad prvních deset lemmat v překladové literatuře, jejich výskyt v tokenech dohromady činí 18,92 % v beletrii a 15,43 % v odborné literatuře, zatímco prvních deset lemmat v nepřekladové texty zabírá v beletrii 17,86 % a v odborné literatuře 14,51 %.

Všechny rozdíly byly na základě testu chí-kvadrát vyhodnoceny jako statisticky signifikantní, ale s ohledem na míru DIN, která měří velikost účinku, nejde o nikterak dramatické rozdíly. Na základě tohoto ukazatele je tedy možné konstatovat, že trend je obdobný jako u ostatních měr (s výjimkou Yuleovy charakteristiky) a překladové texty mají tendenci být na základě zkoumaných parametrů jednodušší, ne však příliš výrazně.

RANK	nepřeklady	%	překlady	%	test chí-kvadrát	DIN
1.-10.	4 743 165	17,86	5 035 976	18,92	p <0,001	2,99
1.-100.	9 390 433	35,37	9 798 588	36,81	p <0,001	2,13
1.-1000.	14 323 279	53,95	14 904 829	56,00	p <0,001	1,99

Tabulka 4.34: Zastoupení frekvenční špičky v korpusu – beletrie

RANK	nepřeklady	%	překlady	%	test chí-kvadrát	DIN
1.-10.	2 314 201	14,51	2 459 974	15,43	p <0,001	3,05
1.-100.	4 676 868	29,32	5 027 881	31,53	p <0,001	3,62
1.-1000.	7 940 083	49,78	8 316 500	52,15	p <0,001	2,32

Tabulka 4.35: Zastoupení frekvenční špičky v korpusu – odborná literatura

Délka věty

Srovnání interpunkce v překladové a nepřekladové češtině (viz 4.1.1) ukázalo, že překladové texty obsahují vyšší počet interpunkčních znamének označujících konec věty/souvětí. Tento rozdíl by se měl projevit i v počtu a délce vět (větším počtem kratších vět v překladech). Následující tabulka 4.36 s průměry naměřených hodnot zmíněný předpoklad potvrzuje: v překladové beletrii v korpusu Jerome jsou věty v průměru o dvě slova kratší než v nepřekladové, zatímco v odborné literatuře jde o rozdíl jednoho slova. Všechny rozdíly jsou statisticky signifikantní (proveden Mann-Whitneyův U-test, $p < 0{,}001$), variační koeficient (VK) v délce věty je také přibližně srovnatelný, můžeme tedy počítat se střední hodnotou.

DÉLKA VĚTY	nepřeklady			překlady		
(průměrná hodnota)	počet vět	délka vět	VK	počet vět	délka vět	VK
beletrie	4 462,75	16,23	31,85%	4 526,51	14,06	25,46%
odborná	2 374,05	17,86	27,21%	3 171,21	16,76	27,80%

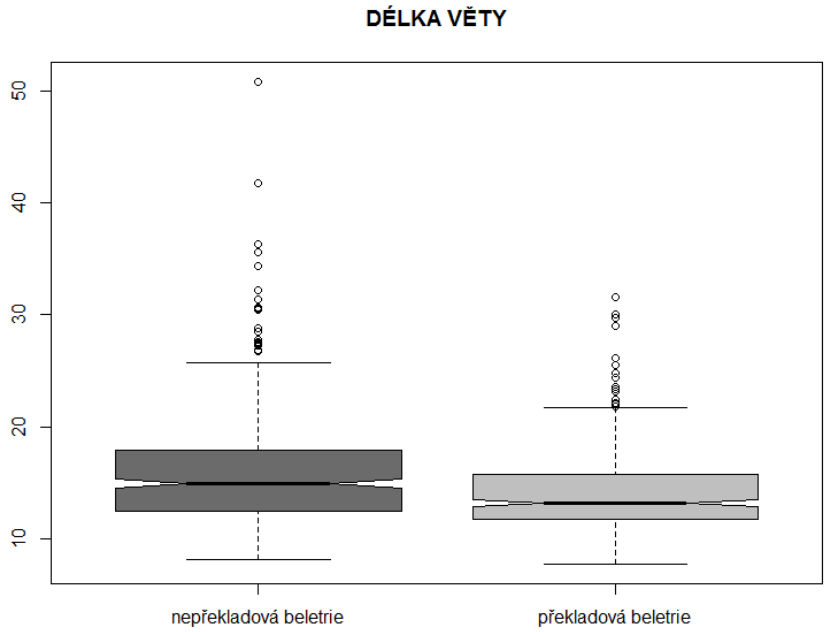
Tabulka 4.36: Průměrná délka vět (ve slovech)

Vykreslíme-li průměrnou délku věty v beletristických textech do krabicového grafu, dostaneme obrázek 4.37, který vypovídá nejen o rozdílu v mediánu, ale také o rozptylu hodnot a tedy o tom, zda jsou v tomto ohledu oba zkoumané soubory homogenní. Ze srovnání vyplývá, že překladové beletristické texty jsou si opět podobnější; ponecháme-li stranou odlehlé hodnoty, mají překladové texty průměrnou délku věty od 7,77 do 21,70 slov, kdežto nepřekladové od 8,12 do 25,73 slov.

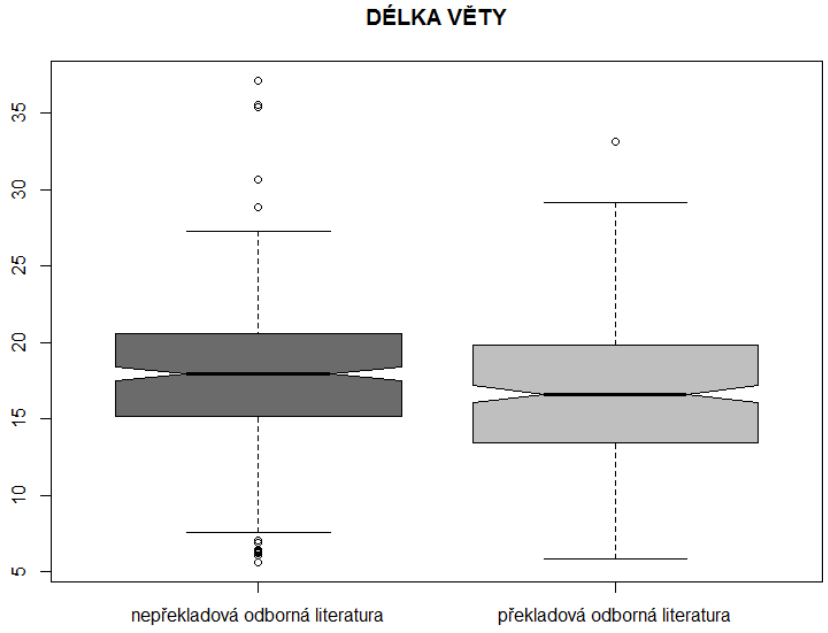
U odborné literatury pozorujeme obdobný trend v tom smyslu, že medián u překladových textů dosahuje hodnoty 16,63 slov ve větě, kdežto u nepřekladových 17,95. Rozptyl hodnot je však jiný – průměrná délka vět se v překladech pohybuje od 5,84 slov do 29,10, kdežto u nepřekladových textů od 7,60 do 27,29. Podobně jako v testu *lexical density* tak mají překladové odborné texty větší rozptyl než texty beletristické. Svou roli v tom patrně hraje samotná povaha odborné literatury, jak je pojímána v této práci – subkorpus odborné literatury obsahuje mnoho různých žánrů/disciplín, jež se mohou lišit (např. akademický text z medicíny oproti populární kuchařce), kdežto beletrii lze považovat za homogennější celek.

Index srozumitelnosti ARI

Posledním testem, který byl proveden v rámci analýzy simplifikačních tendencí v překladové češtině, je index srozumitelnosti neboli *Automated Readability Index*,



Obrázek 4.37: Srovnání délky věty ve slovech – beletrie



Obrázek 4.38: Srovnání délky věty ve slovech – odborná literatura

ARI. Tento index (Smith & Senter 1967) primárně slouží k vyhodnocení srozumitelnosti anglicky psaného textu, přičemž výslednou hodnotu lze interpretovat ve vztahu k různým stupňům amerického školství a odhadnout tak obtížnost textu pro žáky a studenty (*grade level*, GL). Výpočet ARI využívá průměrné délky věty (w/s) a slova (s/w) a vypadá následovně:

$$\text{GL} = 0{,}50\ (\text{w/s}) + 4{,}71\ (\text{s/w})\ \text{-}\ 21{,}43$$

Výpočet lze podle autorů zjednodušit, přičemž zdůraznění délky slova ve vzorci (násobkem devíti) údajně zajišťuje realističtější odhad relevance dvou počítaných faktorů (Smith & Senter 1967: 8):

$$\mathbf{ARI = (w/s) + 9\ (s/w)}$$

Čím vyšší hodnotu GL nebo ARI naměříme, tím je text složitější (v původním využití to znamená, že je určený pro studenty vyššího stupně). Ačkoli byl vzorec pro výpočet ARI vytvořen pro angličtinu (z čehož vychází i hodnota konstant, jež se v něm používají), parametry jako délka věty a slova mohou být považovány za univerzální, přijmeme-li původní premisu dávající do souvislosti délku a srozumitelnost/jednoduchost. Ačkoli o použitých konstantách bychom mohli zcela jistě polemizovat, výsledná hodnota ARI v případě češtiny neslouží jako vodítko pro zařazení textu do konkrétní skupiny podle srozumitelnosti, nýbrž pouze jako srovnávací měřítko pro překladové a nepřekladové texty, proto zde diskuzi nad povahou vzorce pro výpočet ARI ponechávám stranou.

ARI	nepřeklady		překlady	
	ARI	VK	ARI	VK
beletrie	58,84	12,90%	55,31	10,28%
odborná	66,66	10,15%	64,28	10,28%

Tabulka 4.39: Srovnání indexu srozumitelnosti textu ARI

Vypočítáme-li index ARI pro všechny texty ve zkoumaných souborech a porovnáme-li průměrné hodnoty (variační koeficient vyšel pro porovnávané dvojice souborů takřka stejně), zjistíme, že překladové texty v obou textových typech mají nižší hodnotu ARI, jsou tedy z hlediska tohoto indexu jednodušší a srozumitelnější (viz tabulka 4.39). Z tabulky rovněž vyplývá, že odborné texty mají celkově vyšší hodnotu ARI, což znamená, že obsahují delší věty i slova. Vyšší průměrnou délku věty v odborné literatuře potvrzuje i výše uvedený graf 4.38.

Lze tedy konstatovat, že překladové texty se na základě tohoto indexu jeví srozumitelnější a jednodušší, byť rozdíly opět nejsou příliš výrazné. Výsledky této analýzy tak doplňují další informaci do mozaiky testů simplifikace v překladové češtině.

4.2.4 Vliv zdrojového jazyka na výsledky testů

Ačkoli cílem této práce je popsat překladovou češtinu tak, jak se s ní setká běžný čtenář, tedy vč. výrazné převahy textů přeložených z angličtiny, při hledání možné interpretace výsledků z korpusu Jerome mohou být cenným pomocným údajem i **testy provedené na menším subkorpusu** vyváženém z hlediska zdrojového jazyka (viz 3.2.3). Pro doplnění zde proto uvádím i výsledky testování jazykových indikátorů (lexikální bohatost, hutnost vyjádření, srozumitelnost textu) na tomto subkorpusu prostřednictvím operátorů zTTR, LD a délka věty.

zTTR

Výpočet zTTR u textů vybraných do vyváženého subkorpusu přinesl odlišné výsledky než při testování korpusu Jerome. Jak ukazuje graf 4.40 zobrazující beletrii, mezi překladovými a nepřekladovými texty není prakticky žádný rozdíl kromě poněkud většího rozptylu u překladových textů, který je v protikladu k výsledkům testu zTTR u beletrie v celém korpusu Jerome (viz graf 4.28).

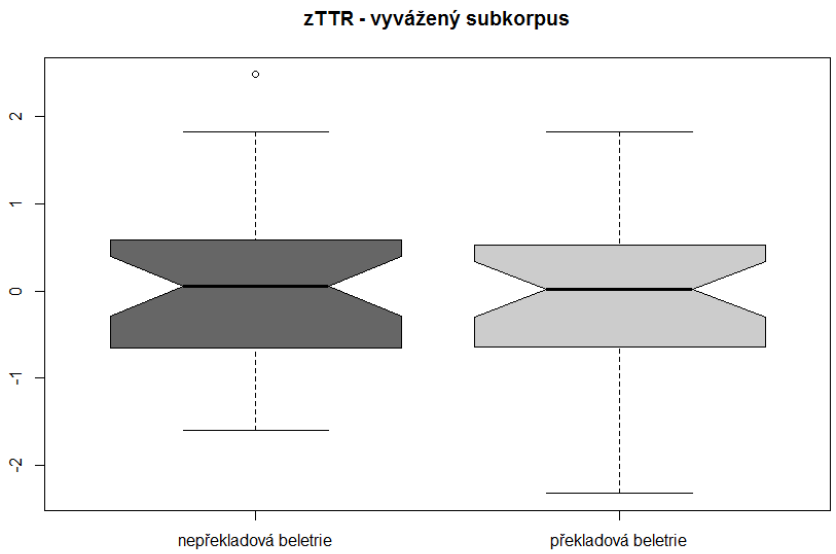
U odborné literatury (viz graf 4.41) sice pozorujeme nižší medián hodnoty zTTR, ovšem vzhledem k tomu, že se výseky (*notches*[13]) obou krabicových grafů překrývají, rozdíl není statisticky signifikantní. Rozptyl hodnot je opět větší u překladové literatury a značí menší homogenitu překladových textů, což je v přímém rozporu s hypotézami univerzálie označované jako konvergence neboli *levelling-out*, viz kapitola 4.3.

Ačkoli je třeba brát v potaz, že vyvážený subkorpus má značnou nevýhodu ve své velikosti – v beletrii je porovnáváno pouhých 32 textů v každém souboru, v odborné literatuře dokonce jen 16 oproti textům v řádu stovek v korpusu Jerome (každý jednotlivý text tak může ovlivnit výsledek mnohem zásadnějším způsobem), tyto výsledky naznačují, že texty přeložené z angličtiny, které zabírají většinu překladové části korpusu Jerome, se chovají jinak než ostatní překladové texty. Aby byl rozdíl dobře patrný, grafy 4.42 a 4.43 znovu zobrazují srovnání indexu zTTR v korpusu Jerome, tentokrát však s rozlišením zdrojového jazyka u překladových textů.

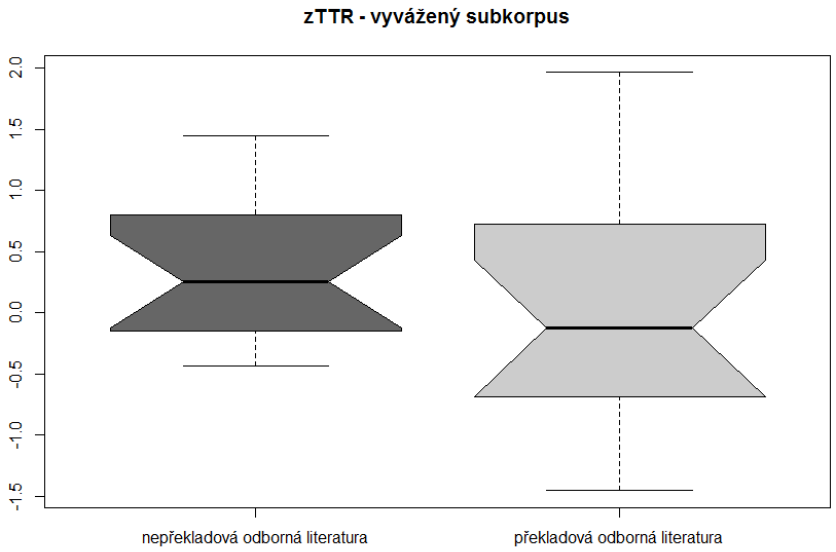
Výsledky srovnání ukazují poměrně překvapivé rozdíly – texty přeložené z angličtiny mají skutečně jiné parametry, ovšem tendence se liší v závislosti na textovém typu. Zatímco v beletrii vykazují překlady z angličtiny nepatrně nižší hodnoty než texty přeložené z jiných jazyků, u odborné literatury je tomu naopak. V obou textových typech sice překlady v součtu dosahují nižších hodnot zTTR než nepřeklady, avšak překlady z angličtiny ukazují v každém z nich jiný trend.

Vzhledem k tomu, jak pestrá je paleta textů zařazených do odborné literatury (od akademických pojednání až po populárně naučné texty v různých disciplínách), můžeme jedno z možných vysvětlení hledat v nerovnoměrném zastoupení těchto podskupin v rámci jednotlivých zdrojových jazyků. V překladech z angličtiny jsou nejvíce zastoupeny disciplíny PSY (18 textů), MED (12), EXC (12), HOU (11) a AMU (8), kdežto v překladech z ostatních jazyků převažuje HOU (27), REL (22),

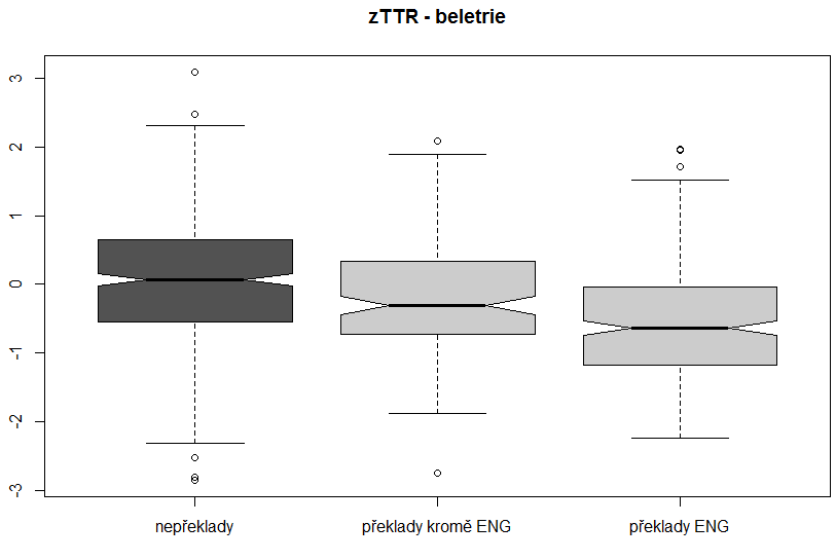
[13]Tyto výseky zobrazují konfidenční interval pro medián – pokud se nepřekrývají, značí to 95 % jistotu, že mediány se statisticky signifikantně liší.



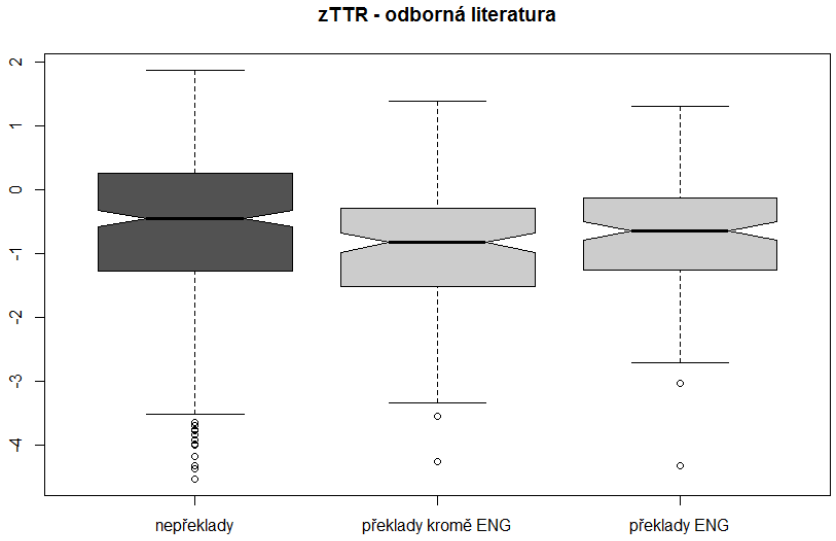
Obrázek 4.40: Srovnání zTTR u vyváženého subkorpusu – beletrie



Obrázek 4.41: Srovnání zTTR u vyváženého subkorpusu – odborná literatura



Obrázek 4.42: Srovnání zTTR v korpusu Jerome podle zdrojového jazyka – beletrie



Obrázek 4.43: Srovnání zTTR v korpusu Jerome podle zdrojového jazyka – odborná literatura

PHI (15) a PSY (14). Opět se tak potvrzuje zjištění, že odborná literatura tak, jak je chápána v tradici Českého národního korpusu, představuje poměrně heterogenní soubor textů a je třeba tuto charakteristiku zohledňovat ve výzkumu.

LD

Dalším zkoumaným ukazatelem je lexikální hustota. Výsledky z vyváženého subkorpusu tentokrát vypadají podobně jako v případě celého korpusu Jerome – v překladech beletrie bylo naměřeno v průměru 49,08 % autosémantik a v nepřekladech 50,29 %, rozdíl však není v tomto případě vzhledem k malému počtu dat statisticky signifikantní, viz graf 4.44. Obdobný závěr můžeme učinit i na základě grafu 4.45, který ukazuje situaci v odborné literatuře. V souladu s obecnou tendencí vykazují odborné texty také celkově vyšší podíl autosémantik než texty beletristické (nepřeklady v průměru 57,27 % a překlady 55,06 %).

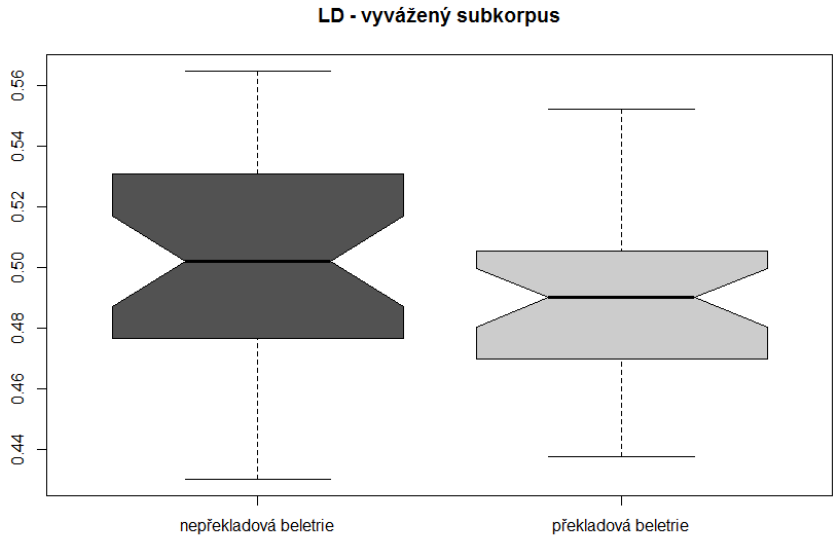
Srovnáme-li opět vliv překladů z angličtiny v celém korpusu Jerome, dostaneme grafy 4.46 a 4.47. Ty potvrzují, že překlady z angličtiny vykazují jiné vlastnosti v beletrii a v odborné literatuře. V beletrii je jejich lexikální hustota nižší (48,92 %) než u překladů z jiných jazyků (49,92 %), zatímco v odborné literatuře se překlady z angličtiny jeví z hlediska hutnosti textu v průměru bohatší (55,84 % oproti ostatním překladům 54,51 %). Rozdíly nejsou dramatické, ovšem vzhledem k velkému počtu dat v korpusu Jerome statisticky signifikantní. Důvodem těchto tendencí může být stejně jako v případě výsledků zTTR odlišná povaha textů zařazovaných do textových typů beletrie a odborná literatura.

Odlišnosti můžeme pozorovat i na úrovni rozptylu hodnot – zatímco beletristické texty přeložené z angličtiny jsou si z hlediska LD o něco podobnější (mají menší rozptyl), odborné překlady z angličtiny mají spíše opačnou tendenci, jak ukazují vousy krabicového grafu 4.47. Blíže se otázce homogennosti/variability překladů věnuje část 4.3.

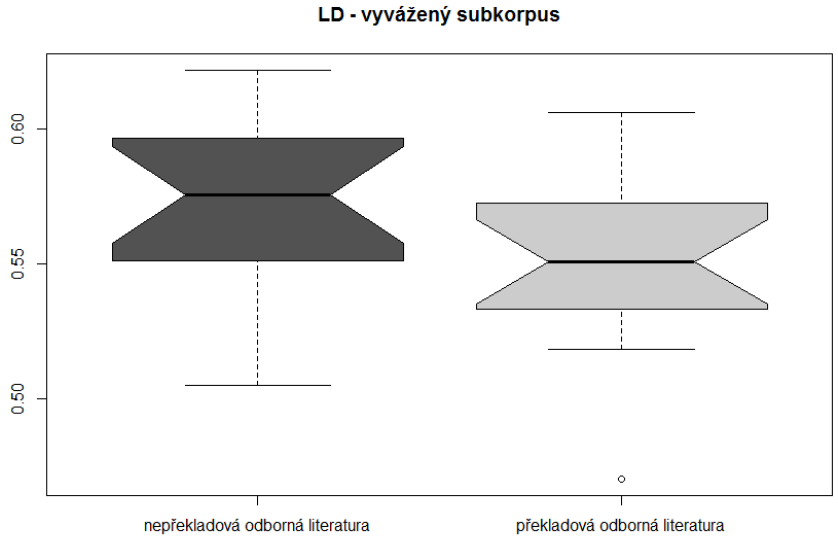
Délka věty

Posledním testovaným ukazatelem u vyváženého subkorpusu byla délka věty. V beletrii činila průměrná délka věty 14,56 slov u nepřekladů a 14,29 slov u překladů – rozdíl není vzhledem k malému počtu textů statisticky signifikantní. V odborné literatuře je výsledek opačný – překlady dosahují v průměru 19,07 slov ve větě, kdežto nepřekladové texty 18,68, opět se však jedná o zanedbatelný rozdíl.

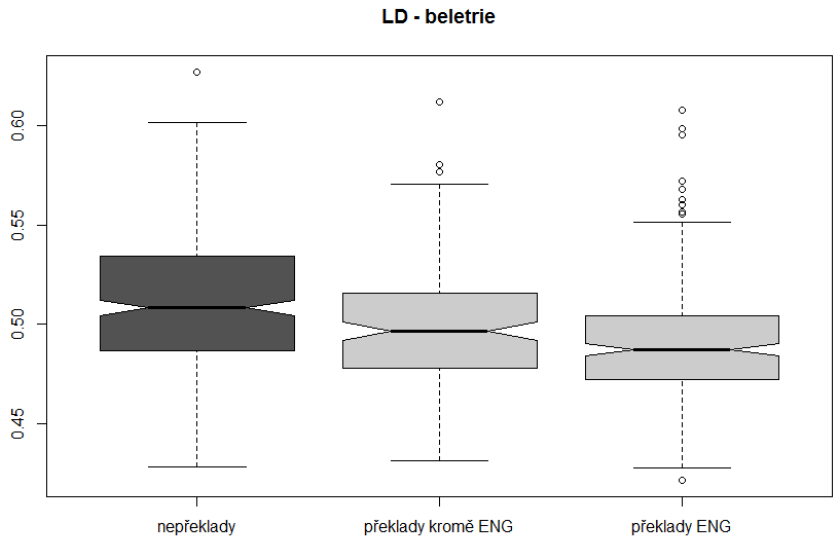
Podíváme-li se do korpusu Jerome zvlášť na překlady z angličtiny (viz grafy 4.48 a 4.49), zjistíme, že v beletrii mají tendenci mít kratší věty než ostatní překlady, kdežto v odborné literatuře je trend u všech překladů srovnatelný. Podobně jako na základě jiných měr (např. LD, viz graf 4.46) se beletristické překlady z angličtiny jeví homogennější (s menším rozptylem), kdežto odborné překlady vypadají v grafu podobně bez ohledu na zdrojový jazyk.



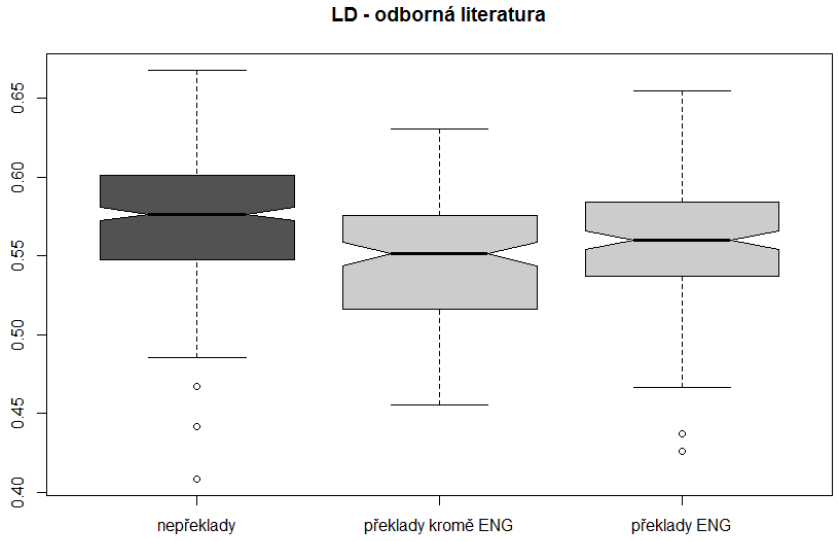
Obrázek 4.44: Srovnání LD u vyváženého subkorpusu – beletrie



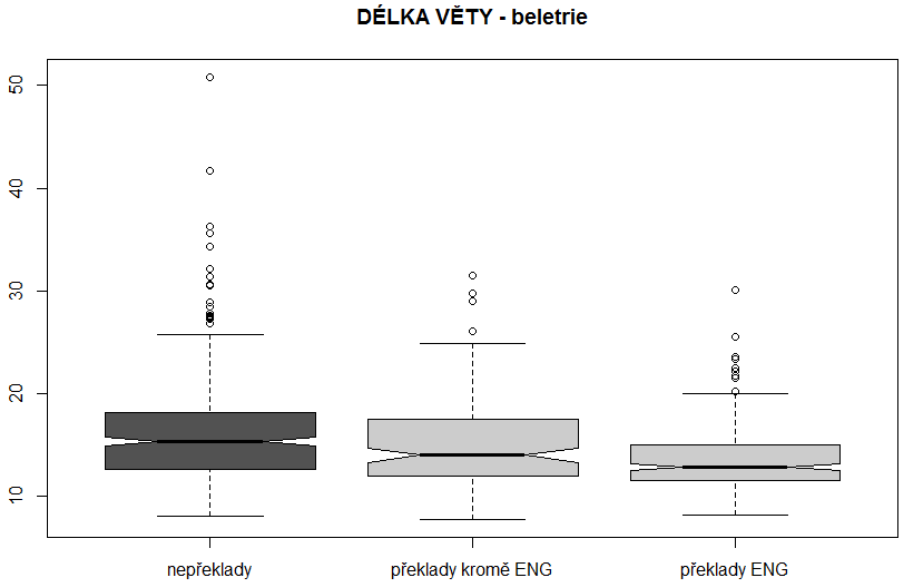
Obrázek 4.45: Srovnání LD u vyváženého subkorpusu – odborná literatura



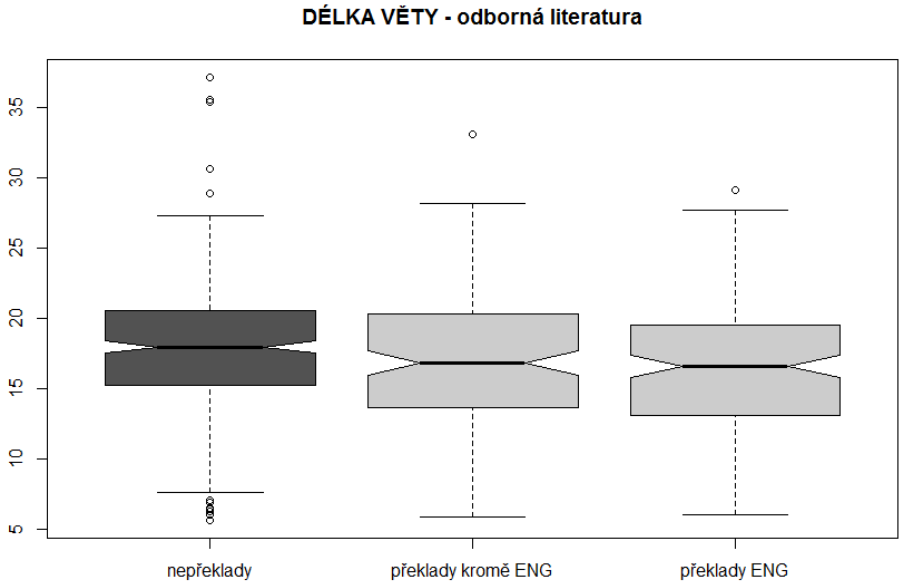
Obrázek 4.46: Srovnání zTTR v korpusu Jerome podle zdrojového jazyka – beletrie



Obrázek 4.47: Srovnání zTTR v korpusu Jerome podle zdrojového jazyka – odborná literatura



Obrázek 4.48: Srovnání délky vět v korpusu Jerome podle zdrojového jazyka – beletrie



Obrázek 4.49: Srovnání délky vět v korpusu Jerome podle zdrojového jazyka – odborná literatura

Shrnutí

Ačkoli byl vyvážený subkorpus sestaven právě pro potřeby ověřování výsledků z celého korpusu Jerome, tento fakt se negativně odrazil v jeho o řád menší velikosti, která si při testování vybírá svou daň – vzhledem k nízkému počtu zkoumaných textů v souborech (32 a 32 textů v beletrii a 16 a 16 v odborné literatuře) hraje každá dílčí odchylka mnohem větší roli. Je tak obtížné určit, které odlišnosti lze přikládat vlastnostem překladového jazyka a které naopak vyplývají např. z idiolektu autora nebo tématu textu. Provedené testy jsou z hlediska statistické signifikance neprůkazné, neboť rozdíly mezi zkoumanými soubory jsou nepatrné a dat je příliš málo na to, aby se dala vyloučit statistická chyba. Můžeme však konstatovat, že kdyby byly rozdíly mezi překlady a nepřeklady opravdu výrazné, projevily by se pravděpodobně i na tomto malém vyváženém korpusu.

O vlivu zdrojového jazyka však můžeme usuzovat i na základě porovnání textů **z celého korpusu Jerome**, pokud zvlášť prozkoumáme **texty přeložené z angličtiny**, jež v korpusu výrazně převažují, a všechny překlady z jiných jazyků. Toto srovnání přineslo výsledky v podobě grafů 4.42, 4.43, 4.46, 4.47, 4.48 a 4.49, které naznačují, že překlady z angličtiny mají skutečně jiné vlastnosti než ostatní překlady. U všech tří provedených testů (zTTR, LD a průměrná délka věty) dosahovaly překlady z angličtiny v beletrii nižších hodnot než překlady z ostatních jazyků a z hlediska rozptylu hodnot se ve dvou testech ze tří (kromě zTTR) jevily jako homogennější celek než ostatní překlady. V odborné literatuře však výsledky srovnání dopadly opačně – překlady z angličtiny dosahují u zTTR a LD nepatrně vyšších hodnot, jen v případě průměrné délky věty jsou na stejné úrovni jako překlady z ostatních jazyků. Rozptyl dat u překladů z angličtiny je také opačný – kromě menšího rozptylu u zTTR je srovnatelný či větší.

Důvodem pro rozdílné výsledky překladů z angličtiny v textových typech by mohla být nejen již zmíněná odlišnost beletrie a odborné literatury, vyplývající z větší pestrosti zařazovaných textů v odborné literatuře, ale také fakt, že překlady z angličtiny jsou v každém z textových typů porovnávány s jinou skupinou textů co do zdrojového jazyka. Zatímco v beletrii převažují v ostatních překladech texty přeložené z francouzštiny a němčiny (oba jazyky jsou zastoupeny rovnoměrně s více než 2 mil. pozic z celkových 8,3 mil.) a celkem tuto skupinu tvoří 20 jazyků, v odborné literatuře němčina výrazně převažuje (téměř 4 mil. pozic, což tvoří více než 50 % všech textů), následovaná francouzštinou (1,3 mil. pozic z celkových necelých 8 mil. pozic)[14].

Pro kontrolu, do jaké míry ovlivňují německé překlady skupinu ostatních překladů, byly pro testy zTTR, LD a délku věty dopočítány i výsledky zvlášť pro překlady z němčiny v porovnání s angličtinou. Trend se však nijak nezměnil: v beletrii dosahují ve všech testech překlady z angličtiny nepatrně nižších hodnot než z němčiny a v odborné literatuře naopak o něco vyšších. Zdá se tedy, že za rozdíly mezi překlady stojí především odlišnost textových typů a nerovnoměrně zastoupené žánry, případně obojí v kombinaci s jiným poměrem jazyků v beletristické a odborné překladové literatuře.

[14]Podrobný popis korpusu Jerome z hlediska počtu jazyků viz kapitola 3.2.2.

4.2.5 Shrnutí výzkumu simplifikace

Výzkum simplifikace provedený v rámci této práce na korpusu Jerome se zaměřil na dvě hlavní oblasti, v nichž se může simplifikace projevovat: bohatost slovní zásoby, vč. hutnosti vyjadřování, a srozumitelnost textu. Pro každý z těchto jazykových indikátorů byly zvoleny dva formální operátory, reprezentující různé metody testování a měření zmíněných indikátorů. Cílem výběru šestice testů bylo minimalizovat vliv jedné metody na výslednou interpretaci a nahlédnout možné simplifikační tendence z různých úhlů.

Co se týče bohatosti lexikonu a hutnosti vyjadřování, které byly testovány pomocí zTTR, Yuleova koeficientu, LD a frekvenční špičky, lze konstatovat, že na základě výsledků třech operátorů ze čtyř (kromě Yuleova koeficientu) je možné zamítnout nulovou hypotézu a dojít k následujícímu závěru:

> **Poměr mezi typy a tokeny (měřený pomocí zTTR), výskyt autosémantik (měřený pomocí *lexical density*) je v překladové češtině nižší a objem slov z frekvenční špičky v korpusu je v překladové češtině vyšší, což nasvědčuje simplifikačním tendencím v překladové češtině. Yuleův koeficient hypotézu o simplifikaci nepotvrdil.**

Z hlediska srozumitelnosti textu, která byla operacionalizována pomocí průměrné délky věty a indexu srozumitelnosti ARI, je možné zamítnout nulovou hypotézu ve všech bodech, platí tedy, že:

> **Průměrná délka věty a hodnota indexu srozumitelnosti (měřená testem ARI) je v překladové češtině nižší než v češtině nepřekladové.**

Ve všech případech, kdy se simplifikační tendence potvrdily, byly rozdíly mezi překladovými a nepřekladovými texty statisticky signifikantní, ovšem s ohledem na velikost účinku nejde o příliš výrazné rozdíly. Trend je z hlediska všech prokázaných testů obdobný – překladová čeština se ve srovnání s nepřekladovou jeví jako simplifikovanější, ovšem rozdíly na těchto makrorovinách, jako je slovní zásoba jako celek nebo srozumitelnost textu, nejsou pro běžného čtenáře pravděpodobně vůbec postřehnutelné.

Pokud bychom měli tyto rozdíly nějak kvantifikovat a lépe ilustrovat, jako referenční hodnota se nabízí například **rozdíl mezi textovými typy**, tedy beletrií a odbornou literaturou. Ačkoli všechna měření probíhala odděleně pro každý textový typ zvlášť, následné porovnání výsledků jasně ukazuje, že větší rozdíly nalezneme mezi beletrií a odbornou literaturou než mezi překlady a nepřeklady. Příkladem mohou být například naměřené hodnoty u testu *lexical density*, které se mezi textovými typy liší rozdílem až 6 procentních bodů (49,28, resp. 51,09 % v beletrii a 55,18, resp. 57,18 % v odborné literatuře), kdežto mezi překlady a nepřeklady jde o rozdíl jednoho až dvou procent.

Lze tedy shrnout, že k simplifikaci v překladové češtině z hlediska testovaného souboru a použitých metod dochází, nicméně rozdíly na zmíněných rovinách nejsou natolik výrazné, aby bylo možné je z hlediska čtenáře jednoduše identifikovat.

4.3 Konvergence (*levelling out*)

Další univerzálií, jejímuž výzkumu se v překladové češtině doposud nikdo systematicky nevěnoval, je konvergence neboli *levelling out*. V označení této univerzálie nepanuje vždy shoda (viz dále), zde jsou však tyto názvy používány synonymicky, přičemž přednost dostává termín **konvergence**, neboť přesněji odráží podstatu této univerzálie a lze s ním v češtině lépe pracovat (na rozdíl od obtížně přeložitelného *levelling out*[15]).

Spolu se simplifikací patří tato univerzálie pod názvem *levelling out* mezi původní překladové rysy Mony Bakerové, ale zároveň mezi nejméně prozkoumané. Cílem části 4.3 je stručně shrnout dosavadní výzkum této překladové univerzálie, představit jednu z metod, jak je možné ji analyzovat v překladové češtině, a konečně popsat, zda ke konvergenci v češtině dochází a za jakých podmínek.

4.3.1 Popis a dosavadní výzkum univerzálie

Bakerová (1996: 176–7) popisuje tuto univerzálii velmi obecnou definicí:

> „Levelling out [...] the tendency of translated text to gravitate around the centre of any continuum rather than move towards the fringes."

Termín *levelling out* si Bakerová vypůjčila od translatoložky Miriam Shlesingerové, která jej poprvé použila v roce 1989 ve své magisterské diplomové práci s názvem *Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts on the Oral-Literate Continuum*. Označovala jím posuny na rovině mluvenosti/psanosti v simultánně tlumočených textech (angličtina – hebrejština). Ve snaze vystihnout lépe podstatu této univerzálie, zavedla pro ni Laviosová (2002: 71) nové označení konvergence a definovala ji jako tendenci překladových textů být si navzájem podobnější (a jako skupina homogennější) z hlediska určitých lingvistických rysů. Dobrým ukazatelem homogennosti skupiny může být např. rozptyl, jak bylo vidět u krabicových grafů v části o simplifikaci. Právě tento indikátor použila Laviosová ve vlastním výzkumu simplifikace (1996, 1998a, 1998b) a zjistila, že překladové texty mají ve zkoumaných hodnotách (TTR, LD a délka věty) menší rozptyl a jsou si tedy bližší než texty nepřekladové.

Tento jev by bylo možné popisovat také jako vedlejší efekt jiné univerzálie, normalizace (neboli konvencionalizace), která značí tendenci volit v překladu především rysy typické pro cílový jazyk, tedy text určitým způsobem neutralizovat, takže výsledný překlad nevybočuje po lingvistické stránce z řady. Bakerová i Laviosová však prosazují konvergenci jako samostatnou univerzálii a takto je chápána i v této práci. Hlavním důvodem je fakt, že příčinou větší homogenity překladových textů může být nejen normalizace, ale také další jevy související s jinými univerzáliemi (např. již zmiňovanou simplifikací).

[15] Sloveso *level out* bychom v češtině mohli přeložit jako „ustálit se", „vyrovnat" či „nivelovat". Termín nivelizace je však v translatologii již zatížen jiným významem, i proto zde pro pojmenování univerzálie volím označení konvergence.

Laviosová proto ve výsledku rozlišuje mezi původním termínem *levelling out* a konvergencí. První jmenovaný termín podle ní označuje tendenci jednotlivých překladových textů pohybovat se blíže středu na nějaké předem definované škále (např. mluvenost/psanost), kdežto konvergence odkazuje k vyšší míře shlukování překladových textů okolo střední hodnoty, ať už měříme jakýkoli lingvisticky relevantní rys nebo jejich kombinaci. I z toho důvodu je v této práci termín konvergence preferován.

Jak již bylo řečeno, konvergence nepatří mezi nejoblíbenější témata korpusově-translatologických výzkumů. Jedním z důkazů budiž fakt, že ani jedna ze dvaceti aktuálních studií, které z metodologického hlediska analyzoval Zanettin (2013), se této univerzálii nevěnovala. Olohanová (2004: 100) to přičítá tomu, že ji nelze tak snadno operacionalizovat a měřit. Badatelé, kteří se jí ve svých pracích věnovali, se proto většinou zaměřili na podobné rysy, které zkoumala už Laviosová (lexikální hustotu, průměrnou délku věty, TTR apod.) a které zpravidla využívají ve výzkumu jiné univerzálie.

To je i případ výzkumného týmu Corpas Pastor, Mitkov, Afzal & Pekar (2008), o jehož výzkumu překladové španělštiny zde již byla řeč v souvislosti se simplifikací. Tito badatelé porovnali překladové a nepřekladové španělské texty z hlediska lexikálních, stylistických i syntaktických rysů domnělé simplifikace (*lexical density*, *lexical richness*, indexy srozumitelnosti, délku věty, podíl jednoduchých vět a souvětí a další) a zjišťovali, zda jsou si překlady podobnější. K výpočtu využili míry pro výzkum podobnosti mezi soubory (nepárový t-test pro jednotlivé rysy a dále test chí-kvadrát pro skupinu rysů jako celek). Poté porovnali p-hodnoty mezi jednotlivými soubory, přičemž menší p-hodnoty značily větší pravděpodobnost, že mezi zkoumanými dvojicemi je rozdíl. Pomineme-li samotnou metodu srovnání, která nic nevypovídá o relevanci případného rozdílu (pouze o jeho statistické signifikanci), výsledek srovnání konvergenci mezi překladovými texty nepotvrdil – pouze v případě jediného zkoumaného rysu si byly překlady podobnější (ve výskytu diskurzních částic).

Obdobný postup zvolila i Lapshinova-Koltunski (2015), která rovněž spolu s konvergencí zkoumala v němčině i další univerzálie (viz výše oddíl o simplifikaci) a využila pak zkoumané rysy k analýze konvergence. Na rozdíl od španělského týmu však badatelka dospěla k jiným závěrům: pomocí Pearsonova testu chí-kvadrát spočítala odchylky mezi překladovými a nepřekladovými texty a zjistila, že překlady jsou si skutečně podobnější než nepřekladové texty. Nejvýznamnější roli v určení podobnosti pak sehrály rysy zkoumané v rámci simplifikace.

Výzkum polského vědce Łukasze Grabowského (2012) se na první pohled může jevit podobně – jako většina badatelů, i on zkoumal dohromady simplifikaci a *levelling out*, ovšem pro testování druhé jmenované univerzálie zvolil metodu **multivariační analýzy dat** (*Principal Component Analysis*, PCA, viz dále). Ta, jak její název napovídá, umožňuje srovnání vícerozměrných dat, tedy laicky řečeno analýzu mnoha pozorování u mnoha objektů zároveň. Grabowski pracoval s malými jednojazyčnými srovnatelnými korpusy beletristických děl o celkové velikosti 500 tisíc pozic, porovnával tedy pouze deset překladových a deset nepřekladových děl. Jako vstupní data pro PCA zvolil frekvence a distribuce 1 000 nejfrekventovanějších slov, z nichž na základě PCA lze určit ta, která nejvíce přispívají k rozdílu mezi

překladovými a nepřekladovými texty. Výsledný graf vytvořený na základě PCA potvrdil Grabowského hypotézu, že polské překladové texty jsou si podobnější a jsou v grafu vykresleny u sebe, kdežto nepřekladové texty jsou rozesety po větší ploše grafu a tvoří tak méně homogenní celek. Z tisíce nejfrekventovanějších lemmat pak Grabowski identifikoval šest slov, která byla z hlediska podobnosti překladů a nepřekladů nejvíce určující: spojku *a*, předložku *v* a *na*, negativní částici *ne*, ukazovací zájmeno *to* a zvratné zájmeno *se* (2012: 178). O relevanci tohoto výsledku by se však dalo diskutovat, neboť se jedná bez výjimky o slova ze souboru nejvíce frekventovaná, u nichž budou veškeré rozdíly statisticky signifikantní – o rozdílech mezi překlady a nepřeklady to však samo o sobě nic neříká.

Ačkoli vypovídací hodnota Grabowského výzkumu je vzhledem k volbě dat značně omezená (malá velikost korpusu, pouze jeden textový typ – beletrie, výběr rysů pro analýzu), metoda PCA, kterou zvolil, se jeví jako vhodný způsob zkoumání konvergence v překladech, neboť umožňuje přehledně porovnat a zobrazit podobnosti a odlišnosti zkoumaných dat na základě mnoha pozorovaných znaků a současně vyhodnotit, které znaky jsou pro toto rozlišení nejpodstatnější, tj. které z naměřených rysů jsou pro zkoumanou univerzálii nejvíce určující (podrobněji o PCA viz 4.3.3).

Podobně jako je tomu u jednodušších statistických metod, i metoda PCA stojí na výběru vhodných jazykových rysů, které do ní vstupují jako proměnné. Na rozdíl od polského výzkumu byly pro **výzkum konvergence v češtině** vybrány rysy, které zahrnují poměrně širokou škálu morfologických, lexikálních i stylistických rysů a neomezují se tak pouze na frekvenční charakteristiku nejčastější slovní zásoby jako v případě Grabowského výzkumu. Následující část 4.3.2 zahrnuje popis těchto rysů spolu s odůvodněním jejich výběru.

4.3.2 Jazykové indikátory a dílčí hypotézy

Jak vyplývá z výše uvedeného přehledu dosavadního výzkumu konvergence, zpravidla se tato univerzálie testuje pomocí hodnot naměřených kvůli ověřování simplifikace. Na podobnost či odlišnost překladových textů však mohou mít vliv i jiné rysy než ty úzce související s bohatostí lexikonu či srozumitelností textu. Z toho důvodu jsem rysy – **vstupní data pro PCA** – vybírala i na základě jiných kritérií, včetně informací z odborné literatury zabývající se srovnáním češtiny a angličtiny (jakožto nejvíce zastoupeným zdrojovým jazykem) a na základě zjištění, jež vyplynula z analýzy POS-gramů a n-gramů.

Tato vstupní data tvoří dva soubory rysů – soubor A zahrnuje vybrané rysy naměřené v rámci výzkumu simplifikace, kdežto soubor B se skládá z nově naměřených hodnot, k jejichž vyhledání bylo využito také morfologické značkování korpusu Jerome. Cílem měření na souboru A je potvrdit či vyvrátit zjištění Lapshinové-Koltunski (2015), že pro identifikaci konvergence se nejlépe osvědčují měření provedená v rámci výzkumu simplifikace. Soubor B pak slouží jako jakýsi ověřovací soubor dat, jenž má ukázat, zda se konvergence projevuje i v případě měření, která přímo nesouvisejí se simplifikací.

Soubor A

Do souboru A byly z naměřených hodnot vybrány ty, které by neměly být závislé na délce textu, u průměrných hodnot byl naměřen obdobný variační koeficient. Hodnota LD značí poměr mezi autosémantiky vůči všem tokenům v textu. Průměrná délka věty a průměrná délka slova (VETA, SLOVO) představují dva vstupní parametry pro index srozumitelnosti ARI, ovšem doporučuje se, aby do PCA vstupovala spíše základní měření, neboť mohou souviset i s dalšími parametry a jejich významnost se tak posoudí lépe, jsou-li uvedena samostatně, a nikoli jako výsledek již provedeného výpočtu. Trojici hodnot doplňuje hodnota zTTR (viz 4.2.3).

LD	hodnota *lexical density*, tedy poměr autosémantik (N, A, V, D) vůči všem pozicím v textu
VETA	průměrná délka věty (ve slovech)
SLOVO	průměrná délka slova (v písmenech)
zTTR	hodnota zTTR

Soubor B

Do souboru B bylo vybráno celkem 16 rysů, jejichž přehled je uveden v rámečku na straně 123. Všechny rysy byly měřeny pro každý text v korpusu Jerome zvlášť a absolutní frekvence byly normalizovány vzhledem k velikosti textu. Cílem bylo vytvořit seznam takových jevů, které mohou poukazovat na variabilitu na rovině morfologické, lexikální, syntaktické i stylistické, přičemž ve výběru jsou záměrně zahrnuty rysy, o nichž je známo, že mohou být ovlivněny procesem překladu (nejčastěji interferencí vlivem zdrojového jazyka, v případě korpusu Jerome angličtiny). Pakliže k těmto jevům v překladech skutečně dochází ve větší míře, tak by se tento vliv měl projevit i v PCA a překladové texty by se sobě měly více podobat.

Několik rysů (PAS, MIN, PRIT, ROZ, VIDN, INF, MOD) se týká sloves, jakožto významných autosémantik. Jejich výběr byl inspirován seznamem lingvistických rysů relevantních pro multidimenzionální analýzu textu (Biber 1995: 95). Biber uvádí celkový seznam 67 rysů pro angličtinu, 58 pro korejštinu a 65 pro somálštinu, jejich výběr však zdůvodňuje pouze velmi obecně: „[...] identifying all linguistic features that might have functional associations (including lexical classes, grammatical categories and syntactic constructions)“. Pro potřeby testování konvergence v češtině tak byly výběrově zvoleny pouze ty rysy, které lze na základě morfologického značkování z korpusu zjistit (nelze např. hledat slovesa podle sémantiky nebo funkce jako pomocná či plnovýznamová) a které mohou mít souvislost s překladem, příp. rozdílem mezi textovými typy, který také hraje podstatnou roli.

Jedním ze zvolených rysů je pasivum. Přemíra pasivních konstrukcí bývá označována za jeden z rysů interference z angličtiny (tedy přebírání rysu typického pro zdrojový jazyk). Čeština kromě analytického pasiva, které lze v podobě pasivního

participia vyhledat podle morfologické značky a je formální obdobou anglického pasiva, disponuje i pasivem zvratným, které je v mnoha případech přirozenější variantou. Je tedy možné, že české překlady (z nichž většina pochází z angličtiny) budou vlivem interference vykazovat větší procento analytických pasivních konstrukcí než texty nepřekladové.

PAS	počet pasivních participií (např. *vytvořen, zkoumáno*)
MIN	počet sloves v minulém čase (aproximace, viz dále)
PRIT	počet sloves v přítomném čase (aproximace, viz dále)
ROZ	počet sloves v rozkazovacím způsobu
VIDN	počet sloves v nedokonavém vidu
INF	počet sloves v infinitivním tvaru
BY	počet kondicionálových tvarů od slovesa *být* (*by*, vč. *aby*, *kdyby*, ve všech tvarech)
MOD	počet modálních sloves *moci*, *muset*, *smět*, *chtít*
LZE	počet výskytů modálního predikativa *lze*
UKAZ	počet ukazovacích zájmen (např. *ten*, *tento*, *takový* ve všech tvarech)
NEJAKY	počet zájmene *nějaký* ve všech pádech, rodech i číslech
CAST	počet částic
SPOJ	počet podřadicích spojek
ZDAT	počet sloves *zdát*, *vypadat* a *připadat* ve všech tvarech
TAK	počet částic *tak*, *tedy*, *přece*, *vždyť*
URCITE	počet modálních částic *určitě*, *jistě*, *snad*, *možná*, *asi*

Sloveso v infinitivním tvaru (INF) se na základě výzkumu slovních druhů (viz Verba v kapitole 4.1) ukázalo jako jeden z rysů, který se v překladech a nepřekladech může lišit. Zda-li se jedná o celkově významný ukazatel, může pomoci odhalit právě metoda PCA. Zbývající slovesné tvary reprezentují buď jevy, na nichž se může projevit interference z angličtiny (např. imperativ ROZ jako ekvivalent časté anglické konstrukce s *Let's*) nebo jevy, které naopak v angličtině nemají svůj přímý protějšek (např. vid zastoupený v seznamu videm nedokonavým VIDN). Seznam doplňují z hlediska překladu neutrální slovesné tvary, ovšem s jasně definovanou jazykovou funkcí: sloveso v minulém a přítomném čase (MIN a PRIT), kondicionál (BY) a modální sloveso (MOD) kromě slovesa *mít*, u nějž nelze snadno z formálního

hlediska zjistit, kdy vystupuje jako plnovýznamové a kdy jako modální (k problematice vyhledání takových rysů viz dále).

Na základě úvahy, že ukazovací zájmeno *ten/ta/to* může být do jisté míry protějškem určitého členu v jazycích, které jím disponují (viz část Pronomina ve 4.1.1), byl do seznamu přidán i počet zájmen ukazovacích v textu. Obdobná hypotéza vedla i k zařazení neurčitého zájmene *nějaký* ve všech jeho tvarech jako případná interference vlivem členu neurčitého.

Dalším naměřeným rysem byl celkový počet částic v textu (CAST). Ačkoli na vyšší rovině zkoumaných souborů (překlady a nepřeklady v beletrii a odborné literatuře) k výrazným rozdílům ve výskytu částic nedochází (viz tabulky 4.2 a 4.3), je možné, že na úrovni jednotlivých textů se počet částic jakožto slovního druhu, který nemá v angličtině jednoznačný protějšek, projeví jako relevantní indikátor. Z hlediska syntaxe byl jako ukazatel zvolen počet podřadicích spojek v textu (SPOJ).

Poslední tři ukazatele byly vyhledány na základě výčtu. Výběr sloves *zdát*, *vypadat* a *připadat* (ZDAT) je motivován anglickými slovesy *seem* a *appear*, jejichž konstrukce lze do češtiny překládat nejen pomocí výše zmíněných sloves, ale také pomocí částic. Nadužívání uvedených sloves tak může opět poukazovat na interferenci z angličtiny a tedy odlišit překlady od nepřekladových textů.

Výčet (TAK) zahrnující *tak*, *tedy* se opírá o fakt, že tyto výrazy nemají v angličtině explicitní protějšek (Dušková 1988: 163) a je tedy možné, že překladatelé budou mít tendenci si na tyto jazykové prostředky nevzpomenout. Seznam doplňují částice *přece* a *vždyť*, jež lze navíc považovat i za jeden z možných projevů explicitace (zvýraznění nevyřčené informace v textu).

Posledním sledovaným rysem je počet modálních částic *určitě*, *jistě*, *snad*, *možná* a *asi* (URCITE), která v angličtině odpovídají modálním konstrukcím. Motivace pro jejich zařazení se tedy opět odvíjí od možného vlivu angličtiny jako nejčastějšího zdrojového jazyka v korpusu Jerome.

Ačkoli se může zdát, že rys CAST je nadmnožinou rysů TAK a URCITE, ve skutečnosti jsou některé z částic tagovány jako spojky (kterými také mohou být, což je pro účely PCA zanedbáno) nebo adverbia, např. *tak*, *vždyť*, *jistě*, *možná*. Zbývající výrazy tagované jako částice pak zabírají v celé skupině částic max. jednotky procent (např. *tedy* 5 %, *snad* 3 %, přece 0,2 %). Množiny se tak zcela nepřekrývají a výčty mají své opodstatnění.

Vyhledání konkrétních tvarů v korpusu na základě morfologických kritérií není vždy přímočarý úkon. Na rozdíl od tradiční morfologie, které často pracuje se složenými tvary, **morfologické značkování v korpusu** je omezeno na jednotlivá slova. Hledáme-li minulý čas v podobě značky pro l-ové participium, musíme kupříkladu nějak zohlednit fakt, že se toto participium podílí i na vytváření kondicionálových tvarů (*dělal jsem* v. *dělal bych*). Obdobně i značka pro kondicionálový tvar slovesa *být* nezahrnuje podřadící spojky *aby* a *kdyby*, které se však na vytváření kondicionálu podílejí až v polovině případů. Proto bylo nezbytné v takových případech přistoupit k aproximaci hledaného jevu a zkombinovat více podmínek hledání. Tento postup se týkal následujících rysů:

BY počet kondicionálových tvarů slovesa *být* plus počet spojek *aby* a *kdyby* ve všech tvarech

MIN počet l-ových participií minus počet všech výskytů kondicionálu (BY)

PRIT počet nedokonavých sloves v přítomném čase minus počet sloves v 1 a 2. os. min. času (eliminace případů, kdy *být* v přít. čase patří ke složenému tvaru min. času)

Na závěr je třeba zdůraznit, že navzdory snaze o promyšlenou volbu rysů bude jakýkoli seznam vždy do značné míry arbitrárním a subjektivním výběrem. Výhodou metody PCA však je mimo jiné i fakt, že dokáže rozeznat, které ze zkoumaných rysů spolu korelují, které mají na rozlišení mezi soubory velký vliv a které zanedbatelný. Lze tedy říci, že ze všech rysů, jejichž zařazení je vždy diskutabilní, by tato metoda měla odhalit ty, které nejlépe poslouží svému účelu – totiž rozlišení mezi překladovými a nepřekladovými texty.

Hypotézy o konvergenci

Hypotézy týkající se konvergence tak znějí následovně:

H_0: Překladové a nepřekladové texty se z hlediska metody PCA na základě uvedených rysů (souboru A a souboru B) neliší.

H_1: Překladové texty jsou si z hlediska metody PCA na základě uvedených rysů (souboru A a souboru B) podobnější než texty nepřekladové.

4.3.3 Popis formálních operátorů a srovnání výsledků

Metoda *Principal Component Analysis*, **PCA** (česky analýza hlavních komponentů[16]), je spolu s tzv. faktorovou analýzou jednou z nejstarších a nejvíce používaných metod vícerozměrné analýzy. U analýz tohoto typu nejde o vztahy mezi závislými a nezávislými proměnnými, ale o vztahy uvnitř jedné skupiny proměnných (Volín 2007: 296). Cílem PCA je jednak snížení počtu původních proměnných, ale také identifikace těch proměnných, které nejlépe vystihují chování souboru.

Klíčovým termínem je zde hlavní komponent (*principal component*, PC), který představuje jakýsi číselný deskriptor tvořený kombinací původních proměnných. Dochází tedy k transformaci původních naměřených proměnných (v tomto případě hodnot v souboru A a B) na nové, nekorelované proměnné, hlavní komponenty. Základní charakteristikou hlavního komponentu je jeho míra variability neboli rozptyl. Hlavní komponenty jsou pak seřazeny podle důležitosti, od největšího rozptylu k nejmenšímu. Nejvíce informací o variabilitě je obsaženo v prvním komponentu a nejméně v posledním. První dva, příp. tři hlavní komponenty se využívají především jako techniky zobrazení vícerozměrných dat v dvourozměrných grafech.

[16]Vzhledem k tomu, že česká zkratka AHK není v literatuře zdaleka tak zavedena jako PCA, preferuji zde anglický název metody.

Výsledky PCA pro soubor A

Nejprve byla PCA provedena u souboru proměnných, které vzešly z výzkumu simplifikace, opět odděleně pro oba textové typy. Tabulka 4.50 pro **beletrii** shrnuje komponentní zátěže s ohledem na původní proměnné, vyjadřuje tedy míru korelace komponentů s původními proměnnými. Z tabulky vyplývá, že nejdůležitější první komponent (PC1) je tvořen pouze proměnnou VETA (tedy průměrnou délkou věty ve slovech). Druhý nejdůležitější komponent (PC2) je pak tvořen kombinací zTTR a délky slova v písmenech, ovšem parametr zTTR se na PC2 podílí větší měrou (hodnota -0,976 oproti -0,205; znaménka zde značí pouze polaritu logiky komponentu, viz Volín 2007: 304).

Soubor A – beletrie	PC1	PC2	PC3	PC4
LD				0,997
VETA	**-0,997**			
SLOVO		-0,205	-0,974	
zTTR		**-0,976**	0,209	

Tabulka 4.50: Komponentní zátěže pro soubor A – beletrie

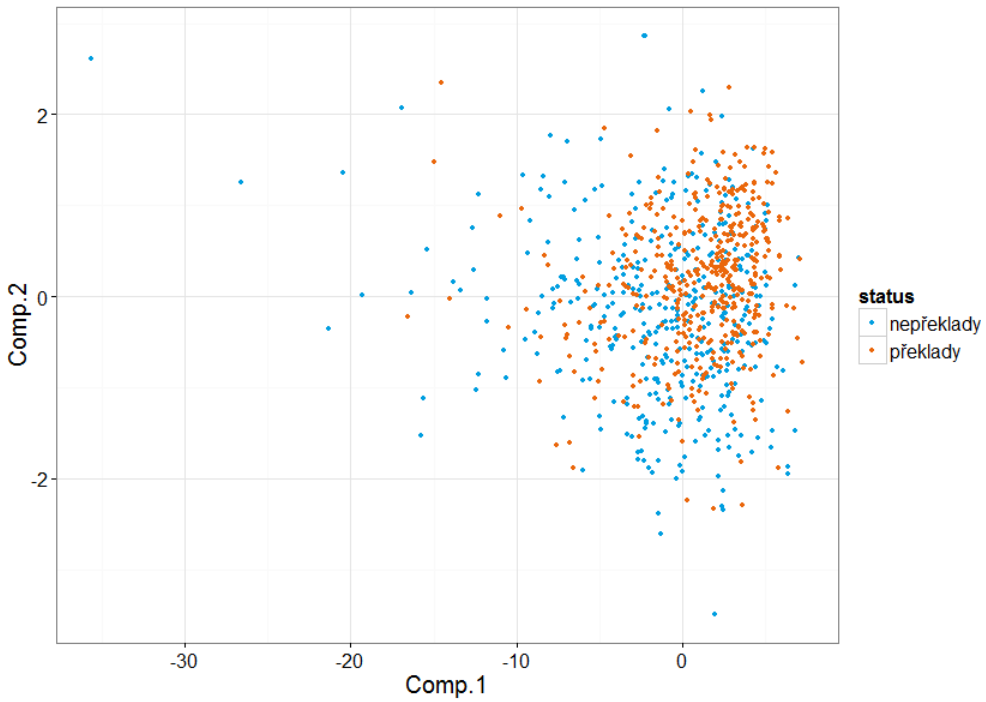
Tabulka 4.51 pak ukazuje především to, jakou míru rozptylu jednotlivé komponenty pokrývají: k vysvětlení 96,2 % variability v datech by stačilo využít pouze první komponent PC1. Vzhledem k tomu, že je tvořen pouze parametrem VETA, můžeme říci, že ukazatel délky věty se jeví jako zásadní, hovoříme-li o rozptylu v beletristických překladech a nepřekladech.

Soubor A – beletrie	PC1	PC2	PC3	PC4
směrodatná odchylka	4,546	0,880	0,220	0,015
podíl zachyceného rozptylu	0,962	0,036	0,002	0,000
kumulativní podíl zachyceného rozptylu	0,962	0,998	1,000	1,000

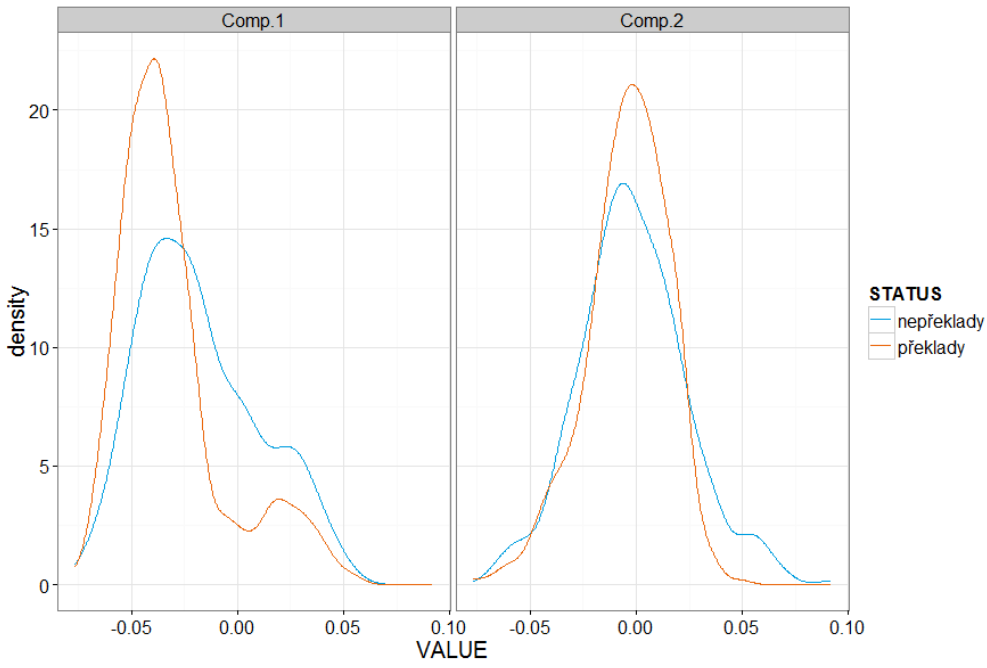
Tabulka 4.51: Významnost komponentů pro soubor A – beletrie

Z výše uvedených souhrnných tabulek však není možné vyčíst, zda lze na základě hlavních komponentů rozeznat překlady od nepřekladů, tedy zda dochází na základě PCA u souboru A ke konvergenci. Odpověď na tuto otázku mohou dát následující grafy 4.52 a 4.53. První z nich zobrazuje na ose x hodnoty PC1 a na ose y hodnoty PC2, které dohromady dokážou zachytit 99,8 % variability v datech (viz tabulka 4.51). Jednotlivé barevné body značí texty nepřekladové (modrá) a překladové (oranžová).

Jak je patrné už z grafu 4.52, beletristické překlady i nepřeklady sice vykazují podobný trend, tedy nacházejí se ve stejné části grafu a nelze je od sebe jasně oddělit, ovšem z hlediska rozptylu zde jistý rozdíl je. Překladové texty mají **menší rozptyl**, jsou více shluknuty u sebe, kdežto nepřekladové texty jsou v grafu více rozptýleny. Tento závěr potvrzuje i graf 4.53, který zobrazuje přímo rozptyl dat, stejně jako F-test porovnávající rozptyly obou souborů (PC1: $F = 2{,}074$, $p < 0{,}001$, PC2: $F = 1{,}467$, $p < 0{,}001$; nepřeklady mají větší rozptyl).



Obrázek 4.52: Výsledky analýzy PCA pro soubor A – beletrie



Obrázek 4.53: Rozptyl hodnot pro soubor A – beletrie

Provedeme-li PCA pro **soubor textů odborné literatury**, dostaneme z hlediska hlavních komponentů obdobné výsledky. Jak vyplývá z tabulky 4.54, jako klíčové parametry pro určení variability v datech se ukázaly být VETA (pro první hlavní komponent) a zTTR (pro druhý), přičemž samotná průměrná délka věty by stačila pro zachycení 94,8 % původní variability, viz podíl zachyceného rozptylu v tabulce 4.55. To je však dáno i faktem, že do PCA vstoupily pouze čtyři různé proměnné.

Soubor A – odborná literatura	PC1	PC2	PC3	PC4
LD			-0,101	0,995
VETA	**-0,996**			
SLOVO			-0,993	-0,101
zTTR		**0,995**		

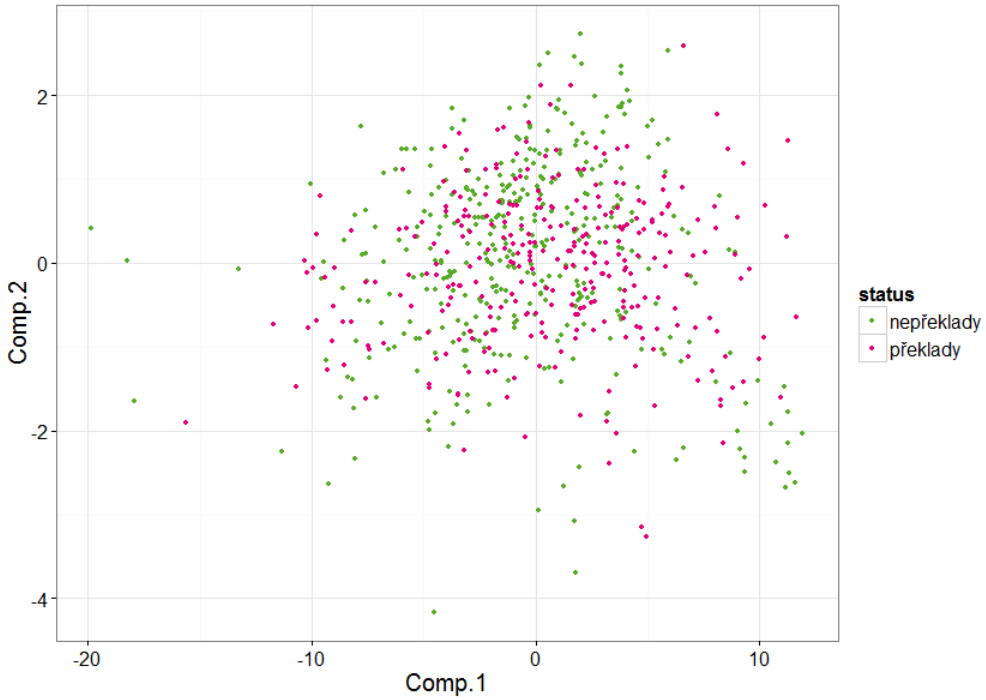
Tabulka 4.54: Komponentní zátěže pro soubor A – odborná literatura

Souhrnné informace o provedené PCA tak vypadají obdobně jako v případě beletrie, jak je to však s rozptylem dat, který je pro konvergenci klíčovým ukazatelem? Podíváme-li se na graf 4.56, na první pohled vidíme, že texty překladové (růžová) nemají tendenci držet více pohromadě, jsou obdobně rozptýlené jako texty nepřekladové (zelená). Přesnější obrázek nabízí graf 4.57, z nějž je patrné, že rozptyly obou zkoumaných souborů jsou obdobné, přičemž odborné nepřekladové texty se v případě PC1 jeví naopak podobnější. Provedený F-test však rozdíl mezi rozptyly potvrdil pouze u PC2 ($F = 1{,}693$, $p < 0{,}001$; větší rozptyl mají nepřeklady), který je řádově méně významným ukazatelem. V tomto případě tedy o konvergenci u překladových textů mluvit nelze. Opět se tedy ukazuje zásadní rozdíl mezi textovými typy, který naznačovaly už dílčí testy provedené v rámci simplifikace. Vzhledem ke složení hlavních komponentů není divu, že odlišné výsledky pro beletrii i odbornou literaturu korelují s grafy 4.37 a 4.38, které zobrazovaly průměrnou délku věty.

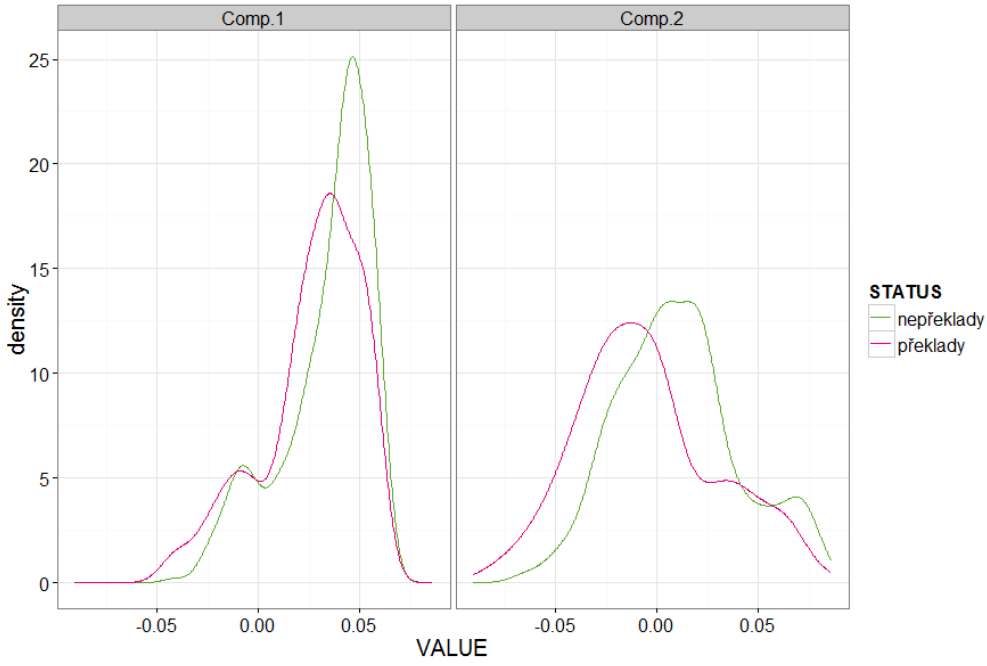
Soubor A – odborná literatura	PC1	PC2	PC3	PC4
směrodatná odchylka	4,800	1,067	0,350	0,020
podíl zachyceného rozptylu	0,948	0,047	0,005	0,000
kumulativní podíl zachyceného rozptylu	0,948	0,995	1,000	1,000

Tabulka 4.55: Významnost komponentů pro soubor A – odborná literatura

Z hlediska PCA lze tedy konstatovat, že k popisu variability v datech na základě souboru A by stačil jeden, příp. dva původní parametry, neboť i v rámci provedené PCA byly využity pouze tyto proměnné. Rozdíl mezi překlady a nepřeklady se neprojevil v tom smyslu, že by oba soubory textů dosahovaly výrazně jiných hodnot (z hlediska grafického zobrazení by se seskupily k sobě a byly jasně rozeznatelné). Z pohledu konvergence a jejího hlavního ukazatele rozptylu však můžeme u překladů a nepřekladů pozorovat odlišný trend v beletrii a odborné literatuře. Zatímco v beletrii ke konvergenci dochází, tj. překlady jsou si podobnější a mají tendenci pohybovat se více okolo střední naměřené hodnoty, v odborné literatuře jsou rozptyly obou souborů srovnatelné. Podstatnou roli zde patrně hraje větší homogennost beletristických textů a specifika složení subkorpusu odborné literatury.



Obrázek 4.56: Výsledky analýzy PCA pro soubor A – odborná literatura



Obrázek 4.57: Rozptyl hodnot pro soubor A – odborná literatura

Výsledky PCA pro soubor B

Stejným postupem byly získány výsledky pro soubor B, zahrnující 16 původních proměnných. Pro **soubor beletristických textů** se v PCA ukázaly být stěžejní především parametry MIN a INF, které tvoří spolu s dalšími rysy první a druhý hlavní komponent, viz tabulka 4.58. Dohromady pak tyto dva komponenty pokrývají 89 % rozptylu v datech, jak ukazuje tabulka 4.59. Na rozdíl od souboru A je v klíčových komponentech zastoupeno více původních proměnných, přičemž většina z nich se týká sloves. Ani jeden ze zamýšlených „překladových“ rysů (tedy těch, které by od sebe měly odlišovat překlady a nepřeklady) se neprosadil jako ukazatel výraznější variability v datech.

Podíváme-li se na grafické znázornění výsledků PCA (viz graf 4.60), můžeme opět pozorovat tendenci překladových textů shlukovat se blíže k sobě, kdežto nepřekladové texty zabírají v grafu větší plochu a jejich rozptyl je větší. To potvrzuje i graf 4.61, z nějž vyplývá, že s ohledem na PC1 i PC2 mají překlady tendenci dosahovat obdobných hodnot, zatímco nepřeklady se více různí. Provedený F-test tento trend potvrzuje (PC1: $F = 1{,}189$, $p = 0{,}039$, PC2: $F = 1{,}734$, $p < 0{,}001$, nepřeklady mají větší rozptyl). Výsledek je tak konzistentní s výsledkem PCA pro soubor A (beletrie).

Co se týče **odborné literatury**, složení hlavních komponentů se příliš neliší od beletrie. Opět hrají nejvýznamnější úlohu slovesné parametry, konkrétně infinitiv (INF) a přítomný čas (PRIT) pro PC1 a minulý čas (MIN) pro PC2, viz tabulka 4.64. Z hlediska zachyceného rozptylu dospějeme k celkové hodnotě 88,3 % pro první dva nejdůležitější komponenty, viz tabulka 4.65. Podíváme-li se na grafické zobrazení výsledků v grafu 4.62 na s. 133, můžeme pozorovat celkově větší rozptyl textů v obou souborech, zapříčiněný pravděpodobně větší heterogenitou textů v odborné literatuře. Srovnáme-li rozptyly překladů a nepřekladů, dostaneme graf 4.63, který ukazuje obdobnou tendenci jako v případě odborné literatury u souboru A: překladové texty nevykazují nižší variabilitu, naopak. F-test potvrdil, že v odborné literatuře (na základě PCA souboru B) jsou si naopak o něco podobnější nepřeklady (PC1: $F = 0{,}817$, $p = 0{,}031$, PC2: $F = 0{,}766$, $p = 0{,}007$, překlady mají větší rozptyl).

Z údajů v tabulkách tedy vyplývá, že nejlepším ukazatelem variability jsou obecné charakteristiky týkající se sloves (počet sloves v minulém čase, přítomném čase a v infinitivu), nikoli vybrané rysy související s překladem či možnou interferencí z angličtiny. Metoda PCA sice nerozlišuje, do jaké míry se překlady a nepřeklady v jednotlivých původních proměnných liší (může tedy docházet k dílčím odlišnostem), ale samotný fakt, že tyto rysy netvoří součást nejdůležitějších komponentů, naznačuje, že v nich nedochází k výrazným rozdílům.

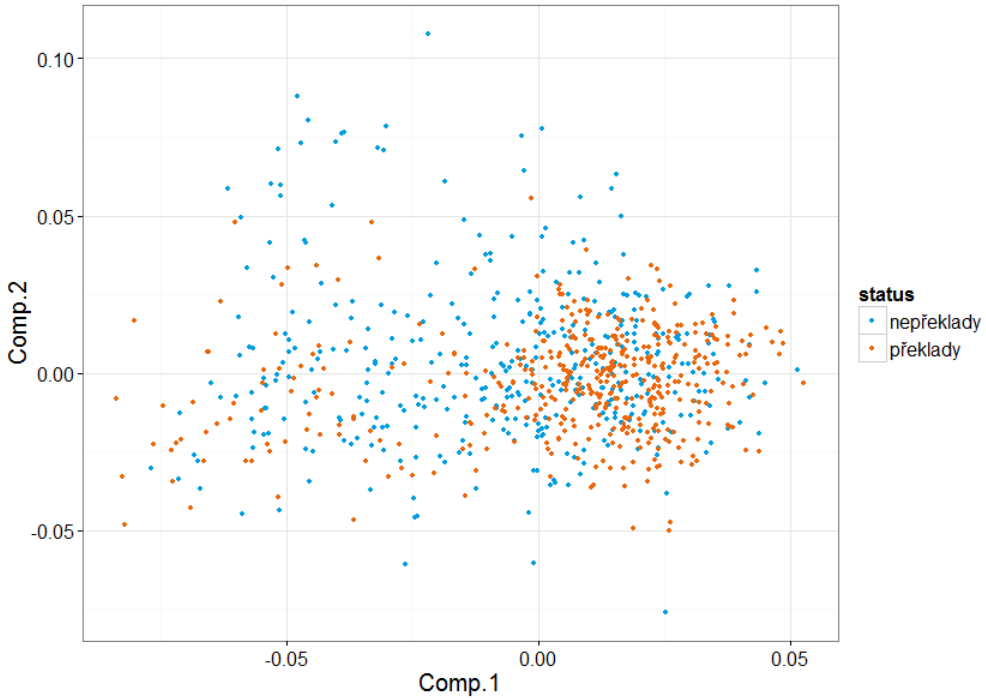
Na základě analýzy souboru B tak můžeme konstatovat, že ke konvergenci u překladů dochází pouze v beletrii, kde překladové texty skutečně nabývají obdobných hodnot a tvoří homogennější celek než nepřekladové texty. U odborné literatury je tomu ovšem naopak – překladové texty mají větší rozptyl a podobnější jsou si texty nepřekladové.

Soubor B	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
PAS				0,147				-0,277	0,912	-0,185	-0,136					
MIN	**0,875**	-0,159	0,171	-0,363	0,180											
PRIT	-0,480	-0,346	0,279	-0,632	0,368		-0,114									
ROZ				-0,132		-0,227	0,284	0,856	0,279	0,106	-0,159					
VIDN		-0,144	-0,182	0,112	0,352	-0,689	0,310	-0,310		0,232		-0,292				
INF		**-0,853**	0,218	0,367	-0,294											
MOD					0,113	-0,286	0,129			-0,262	0,101	0,888				
UKAZ		-0,215	-0,650	-0,425	-0,417	0,170	0,336	-0,141			-0,106					
NEJAKY													-0,184	0,722	0,656	
BY			-0,172		0,173			0,121	-0,135	-0,881		-0,321	0,123			
CAST			-0,301		-0,166	-0,401	-0,780				-0,226		0,192			
SPOJ		-0,200	-0,487	0,285	0,608	0,424	-0,155	0,126		0,190						
ZDAT													-0,143	0,633	-0,750	
TAK			-0,143				-0,139	0,117	0,191		0,931	-0,102				
URCITE							-0,148			-0,116			-0,939	-0,247		
LZE																0,995

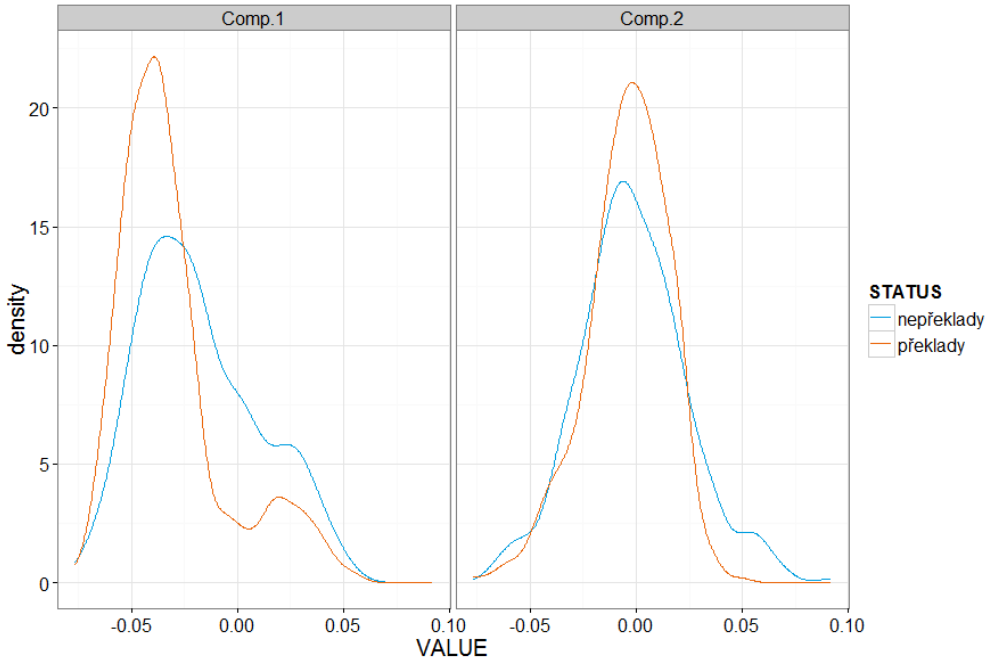
Tabulka 4.58: Komponentní zátěže pro soubor B – beletrie

Soubor B	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
sm.odch.	0,028	0,022	0,007	0,006	0,006	0,003	0,003	0,002	0,002	0,002	0,001	0,001	0,001	0,001	0,000	0,000
rozptyl	0,564	0,326	0,035	0,027	0,022	0,008	0,007	0,004	0,002	0,002	0,001	0,001	0,000	0,000	0,000	0,000
kum. rozptyl	0,564	0,890	0,926	0,953	0,975	0,983	0,989	0,993	0,996	0,997	0,999	0,999	1,000	1,000	1,000	1,000

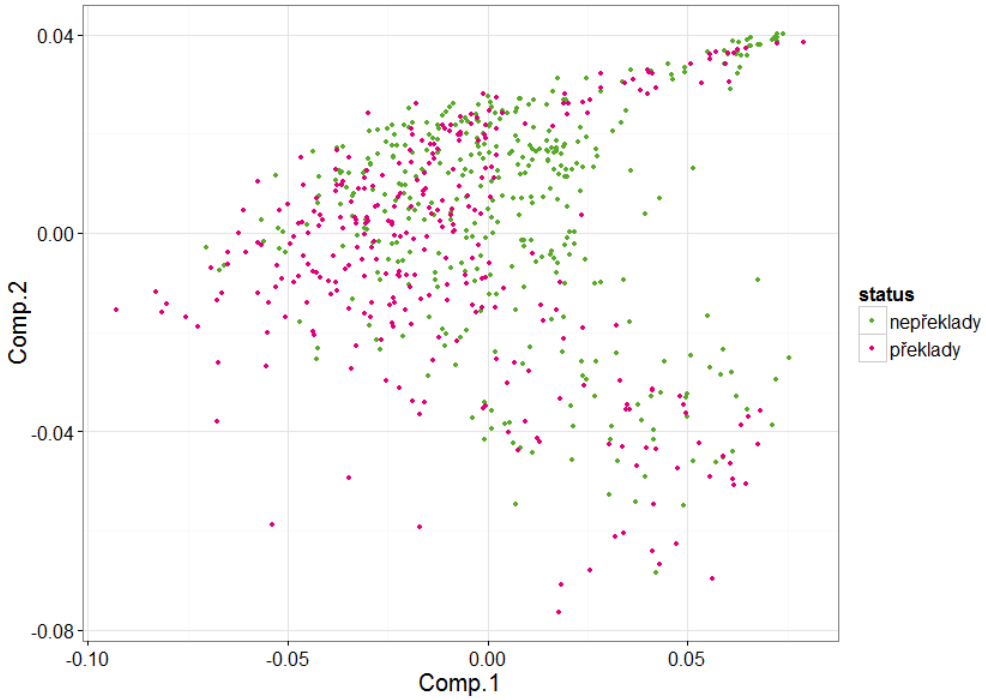
Tabulka 4.59: Významnost komponentů pro soubor B – beletrie



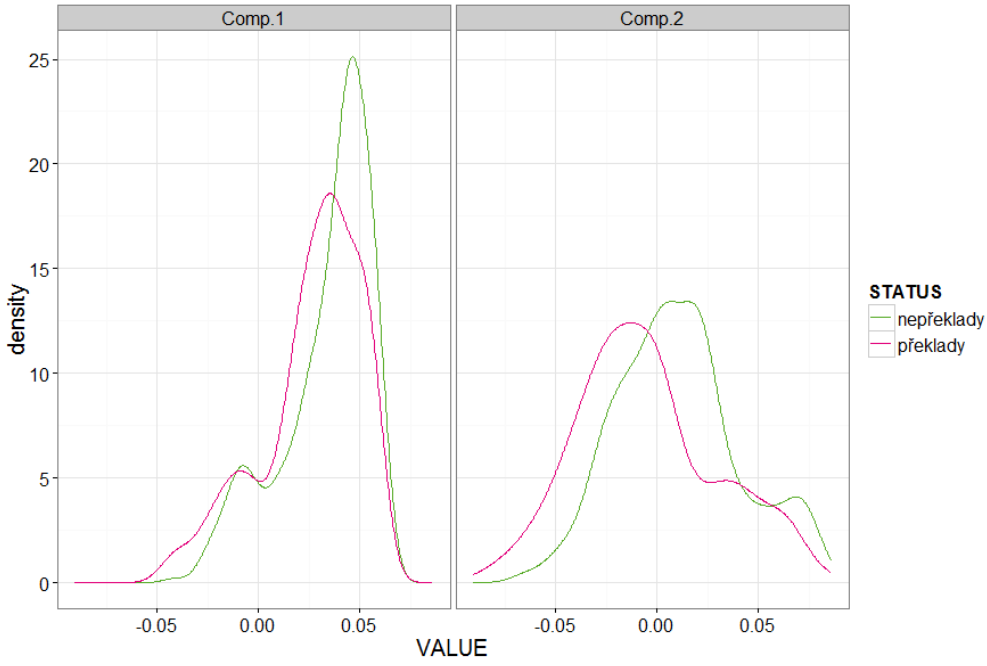
Obrázek 4.60: Výsledky analýzy PCA pro soubor B – beletrie



Obrázek 4.61: Rozptyl hodnot pro soubor B – beletrie



Obrázek 4.62: Výsledky analýzy PCA pro soubor B – odborná literatura



Obrázek 4.63: Rozptyl hodnot pro soubor B – odborná literatura

Soubor B	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
PAS			-0,161	0,237	-0,299	0,744	-0,192	0,145	-0,439			0,106				
MIN	0,195	**-0,862**	-0,126	0,190		-0,235		0,233	-0,177							
PRIT	**-0,598**	0,255	-0,341	0,189	0,249	-0,286		0,386	-0,306		-0,147					
ROZ			0,789	0,503	0,280			0,130								
VIDN	-0,172		0,357	-0,348	-0,638	-0,192	0,270	0,272	-0,221	0,239	-0,119					
INF	**-0,698**	-0,353		0,231	-0,216	0,114		-0,400	0,305							
MOD			0,106		-0,184				-0,115	-0,800	0,486	-0,142		-0,121		
UKAZ	-0,169	-0,149		-0,396	0,330	0,481	0,381	0,406	0,356							
NEJAKY															-0,376	0,924
BY			0,130	-0,150			-0,167	-0,136	0,134	-0,469	-0,784	0,110	-0,206			
CAST				-0,207	0,259		0,474	-0,555	-0,527				-0,233			
SPOJ	-0,212	-0,185	0,245	-0,451	0,297		-0,673		-0,271	0,145	0,14					
ZDAT															-0,925	-0,377
TAK							0,106			-0,117	-0,227	-0,455	0,828			
URCITE								-0,170		-0,110		0,844	0,449	-0,143		
LZE									-0,104	0,123	-0,106	-0,113		-0,972		

Tabulka 4.64: Komponentní zátěže pro soubor B – odborná literatura

Soubor B	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
sm.odch.	0,033	0,024	0,010	0,006	0,005	0,004	0,004	0,003	0,002	0,002	0,001	0,001	0,001	0,001	0,000	0,000
rozptyl	0,590	0,293	0,053	0,021	0,014	0,010	0,007	0,005	0,003	0,001	0,001	0,001	0,000	0,000	0,000	0,000
kum. rozptyl	0,590	0,883	0,936	0,957	0,972	0,981	0,988	0,993	0,996	0,998	0,998	0,999	1,000	1,000	1,000	1,000

Tabulka 4.65: Významnost komponentů pro soubor B – odborná literatura

4.3.4 Shrnutí výzkumu konvergence

Na základě analýzy PCA lze postulovat několik závěrů. První se týká pozorovaných parametrů, které vstupovaly do PCA jako **původní proměnné**. V souboru A, který zahrnoval čtyři hodnoty naměřené v rámci výzkumu simplifikace, se jako klíčový ukazatel variability v datech projevila především průměrná délka věty a dále hodnota zTTR. V souboru B, který obsahoval celkem 16 různých hodnot z oblasti morfologie, lexikologie i syntaxe, se jako nejdůležitější parametry (tvořící základ hlavních komponentů) ukázaly slovesné charakteristiky: počet sloves v minulém a přítomném čase a počet sloves v infinitivu.

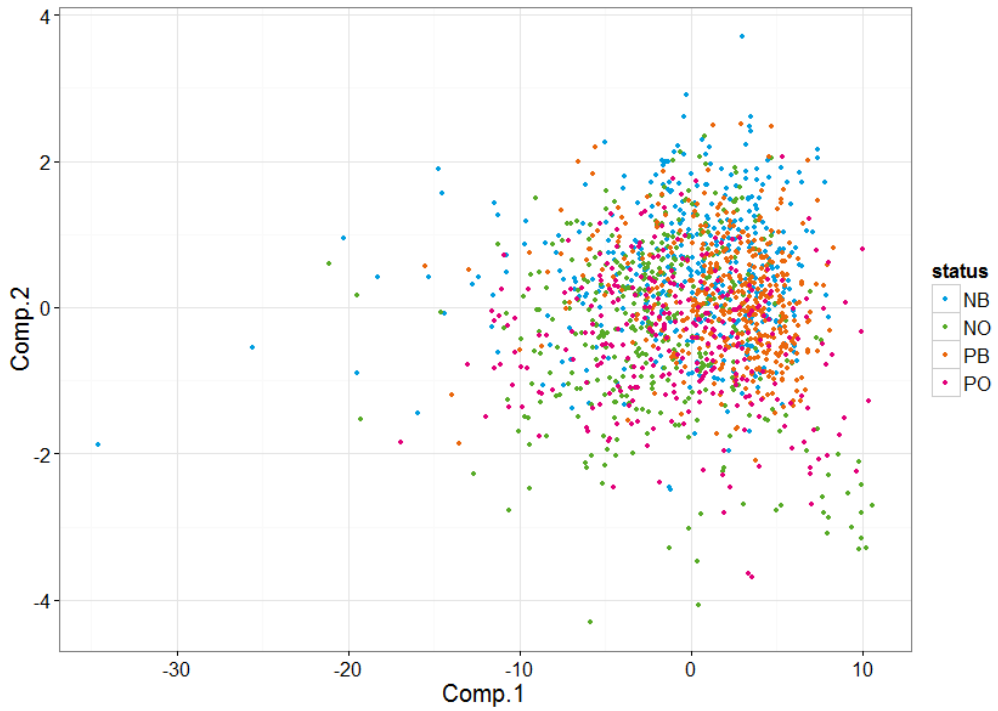
Druhý závěr se vztahuje k samotné výzkumné otázce, zda u překladových textů dochází ke konvergenci či nikoli. Samotné **testování hypotézy** proběhlo na základě PCA, přičemž odlišnost či podobnost překladů a nepřekladů z hlediska variability následně potvrdil provedený F-test, který slouží k porovnání rozptylů dvou vzorků. S ohledem na dílčí hypotézu tak platí následující výrok:

> **Beletristické překladové texty jsou si z hlediska metody PCA na základě uvedených rysů (souboru A a souboru B) podobnější než texty nepřekladové. V odborné literatuře na základě téže analýzy ke konvergenci u překladových textů nedochází.**

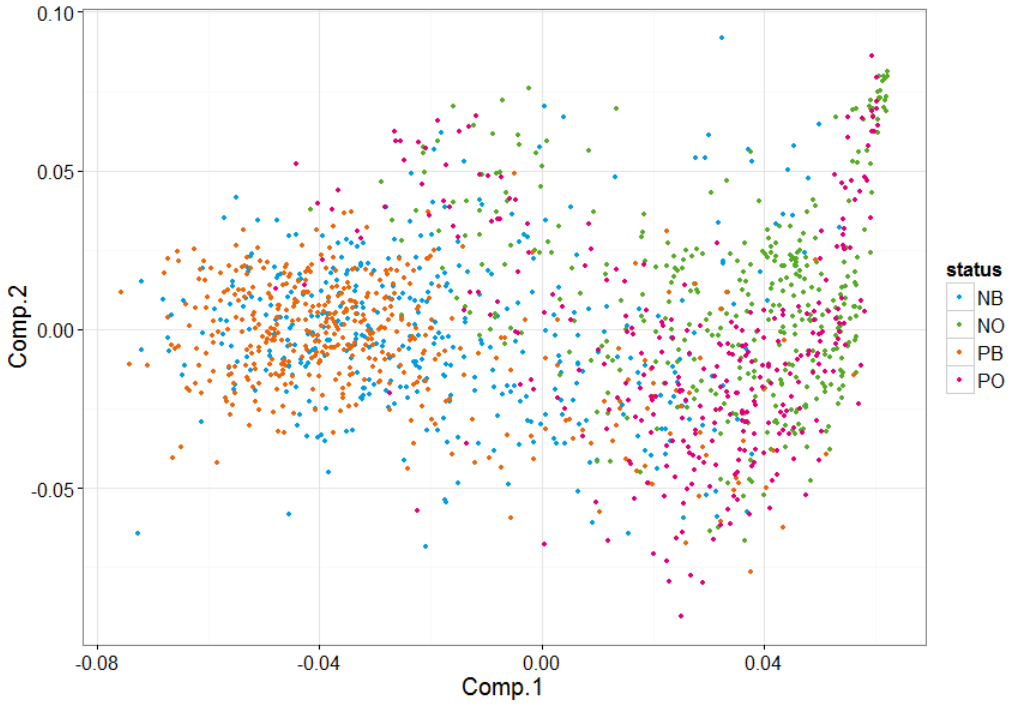
Dále se analýzou potvrdilo, že parametry získané v rámci výzkumu simplifikace, soubor A, je možné využít jako vstupní data pro PCA a že mohou být jedním z indikátorů konvergence. Totéž ovšem platí i pro rozsáhlejší, a tedy spolehlivější soubor B, který měl ukázat, zda se podobnost překladových textů z hlediska rozptylu projevuje i u jiných proměnných než u těch, jež se bezprostředně vztahují k simplifikaci. I tento předpoklad se tímto potvrdil.

A konečně třetí a poslední závěr lze formulovat s ohledem na **odlišnost výsledků v obou textových typech**. Metoda PCA opět poukázala na fakt, že texty beletristické mají jiné vlastnosti než texty odborné. Aby byl rozdíl co nejlépe patrný, provedla jsem pro doplnění PCA na celém korpusu Jerome bez žánrového rozlišení a výsledky graficky zobrazila, viz graf 4.66 a 4.67. V případě souboru A tvoří všechny texty dohromady jeden velký shluk, v němž lze jen stěží pozorovat určitou tendenci odborných textů, jež se vyskytují spíše ve spodní části grafu, zatímco beletristické texty nacházíme spíše nad hodnotou 0 osy y. Osa y však značí druhý hlavní komponent (PC2 tvořený hodnotou zTTR), který nemá takovou výpovědní hodnotu. Výsledek je i přesto konzistentní se závěry výzkumu simplifikace na základě zTTR (viz 4.2.3), kde se rovněž projevuje rozdíl mezi beletrií a odbornou literaturou.

U souboru B je odlišný trend obou textových typů vidět mnohem zřetelněji. Zatímco beletristické texty (modrá, oranžová) se bez ohledu na status (překlad/nepřeklad) shlukují v levé části grafu, odborné (zelená, růžová) nacházíme vpravo, přičemž jakousi dělicí linií může být hodnota 0 na ose x. Hlavní komponenty tvoří jako v případě oddělených analýz opět především parametry MIN, INF a PRIT. Lze tak znovu konstatovat, že z pohledu této analýzy je rozdíl mezi textovými typy pro uživatele jazyka mnohem snáze zachytitelný než rozdíl mezi překladovými a nepřekladovými texty.



Obrázek 4.66: Výsledky analýzy PCA pro soubor A – celý korpus Jerome



Obrázek 4.67: Výsledky analýzy PCA pro soubor B – celý korpus Jerome

4.4 (Ne)typické slovní kombinace v překladech

Na rozdíl od předchozích dvou podkapitol, které se zabývaly vždy jednou konkrétní překladovou univerzálií (simplifikací a konvergencí), je poslední část této kapitoly věnována obecnějšímu tématu slovních kombinací v překladovém jazyce. Nejprve je třeba říci, že toto téma je tak široké, že by bylo jistě možné mu věnovat samostatnou dizertační práci; s ohledem na kapacitu této kapitoly je tedy tato část spíše sondou než zevrubným popisem této problematiky. Translatologické studie týkající se specifických víceslovných jednotek v překladu se zpravidla vztahují k jedné ze dvou univerzálií – normalizaci a *unique items hypothesis*, proto bude nejprve následovat stručné shrnutí dosavadního výzkumu v této oblasti. Poté se na základě analýzy n-gramů pokusím identifikovat a blíže popsat takové slovní kombinace, které jsou pro překladový jazyk charakteristické – ať už se jedná o jevy v češtině typické (tedy časté, široce používané) nebo naopak o kombinace netypické, které v nepřekladové češtině nenacházíme tak často.

Výraz **slovní kombinace** je zde přitom volen záměrně jako neutrální zastřešující termín (*word combination* podle Mauranen 2000: 131), jenž může zahrnovat kolokace, koligace, *lexical bundles* a další typy lexikálních struktur. V obdobném smyslu se v angličtině pro ustálené slovní kombinace často používá také termín (*lexical*) *pattern*, pro který v češtině neexistuje jednoznačně zavedený ekvivalent (lze uvažovat o ekvivalentech ustálená struktura či lexikální vzorec). Mauranenová definuje výraz *pattern* jako opakovanou konfiguraci gramatických a lexikálních jednotek okolo základového lexikálního slova. Pokrývá tím tak všechny možné druhy slovních kombinací (kolokace, koligace atd.), přičemž důraz je kladen právě na ustálenost a opakovanost výskytu. V obdobném smyslu jsou tyto výrazy používány i zde v této části. Blíže se kritériím výběru zkoumaných slovních kombinací věnuje část 4.4.2.

Vzhledem k tomu, že v této části nejde o testování konkrétní univerzálie, nýbrž o popis specifických slovních kombinací (s přihlédnutím k dosavadnímu translatologickému výzkumu), neřídí se struktura této části přesně Zanettinovým teoretickým rámcem (od jazykových indikátorů a hypotéz k formálním operátorům), ačkoli v popisu a analýze vybraných jednotek lze samozřejmě vysledovat obdobný postup.

4.4.1 Popis a dosavadní výzkum jevu

Jak již bylo naznačeno, analýzu slovních kombinací v překladovém jazyce by bylo možné provádět pod hlavičkou různých univerzálií v závislosti na tom, jak široce si badatel definuje jazykové indikátory vztahující se k jednotlivým překladovým rysům a na které roviny jazyka se chce zaměřit. Zde však bude pozornost věnována dvěma univerzáliím, pro něž je výzkum syntagmatických jednotek a ustálených struktur klíčový a které přinesly zajímavé výsledky i podněty k diskuzi.

Normalizace

První z nich je jedna z původních bakerovských univerzálií: normalizace (někdy označovaná jako standardizace nebo konvencionalizace). Ta bývá obecně definována

jako „tendency to exaggerate features of the target language and to conform to its typical patterns“ (Baker 1996: 183). Už v samotné definici zaznívá výraz *pattern*, jenž je zde chápán velmi široce jako prakticky jakýkoli jazykový jev opakovaně se vyskytující a zavedený v cílovém jazyce. Překladový jazyk se podle této hypotézy jeví jako „normálnější“ než jazyk nepřekladový, což se projevuje nadužíváním klišé a typických gramatických či kolokačních struktur či přizpůsobováním interpunkce cílové normě. Jinými slovy překlady jsou považovány za méně příznakové, méně kreativní a na druhou stranu více konvenční a standardizované.

Normalizací z hlediska lexikálního se zabýval už Toury (1980: 130), když při svém výzkumu zjistil, že dvojčleny coby typický prvek v hebrejštině se mnohem častěji vyskytují v překladech než v původních hebrejských textech. Také Vanderauwerová (1985: 108), o jejímž výzkumu zde byla řeč zejména v souvislosti se simplifikací, dochází k tomu, že překladatelé beletrie do nizozemštiny se zdráhají využívat netypických, osobitých či nápadných jazykových prostředků. Stejně tak Malkmjærová (1998) na základě analýzy několika anglických překladů jediného dánského textu zjistila, že většina překladatelů vytvořila lexikálně konvenčnější cílový text, ačkoli zdrojový text se po této stránce vymykal dánským normám. Lexikální stránku překladového jazyka zkoumala i Puurtinenová, která se omezila na jeden konkrétní žánr, totiž finskou dětskou literaturu. Na základě výzkumu klíčových slov na přibližně milionovém jednojazyčném srovnatelném korpusu došla k závěru, že překladatelé mají tendenci volit častěji neutrální, standardnější výrazy a vyhýbat se hovorovým a jinak příznakovým slovům (k podobnému závěru odkazují i dílčí závěry analýzy sloves v překladové češtině, viz s. 76 a s. 77).

S ohledem na **víceslovné jednotky**, jež jsou zde hlavním předmětem zájmu, se na hypotézu normalizovanějších překladů podívala např. Bakerová (2004), která se zaměřila především na metodologické otázky. Využila k tomu přibližně šestimilionový korpus překladové angličtiny a stejně velký korpus angličtiny nepřekladové (subkorpus BNC). Výchozím předpokladem její studie bylo, že překladatelé mají tendenci být konzervativnější v používání jazykových prostředků a produkovat tak jednotnější texty, neboť hlavním požadavkem čtenářů na překladový text je plynulost (Baker 2004: 173). Jelikož k plynulosti přispívá častý výskyt zavedených a ustálených frází a slovních spojení, cílem studie bylo ověřit, zda se v překladech tyto jevy vyskytují častěji než v nepřekladových anglických textech.

Klíčovým úkolem bylo samozřejmě identifikovat tyto ustálené fráze neboli vzorce (*patterns*), k čemuž Bakerová zvolila metodu extrakce n-gramů (3-, 4-, 5-gramů). Jejich celkový počet neuvádí, ale po ručním třídění zbylo přibližně padesát lexikálních vzorců vybraných na základě dvou kritérií (1: ustálenost a opakující se výskyt v textech oproti frázím, které se vztahují ke konkrétnímu tématu nebo textu, 2: vyřazení většiny výrazů času a místa, neboť podle Bakerové vykazují obdobný výskyt), přičemž preferovány byly výrazy metatextové (jako např. *in other words*, *that is to say* atd.) a diskurzivní částice strukturující text (*when it comes to*, *on the other hand* atd.). Výsledky srovnání potvrdily, že mnoho z těchto výrazů se častěji objevuje v překladové angličtině. Bakerová ve studii dále rozebírá možné problémy při zpracování dat touto metodou a zdůrazňuje zásadní úlohu badatele při interpretaci výsledků.

Obdobný cíl jako Bakerová, tedy popsat odlišnost či podobnost překladového jazyka z hlediska ustálených slovních kombinací, si ve své studii vytyčila i Bernardiniová (2007), jejímž cílem bylo jednak prozkoumat **kolokační vzorce** v překladové italštině a jednak poukázat na výhody kombinace srovnatelného, paralelního a referenčního korpusu. Vzhledem k tomu, že Bernardiniová měla k dispozici velmi malý anglicko-italský recipročí korpus (přibližně 200 tisíc slov), veškeré výpočty frekvence a distribuce (viz dále) prováděla na referenčních korpusech (BNC a Repubblica) a recipročí korpus sloužil pouze jako zdroj seznamu kolokací a k jejich pozdější podrobnější analýze.

Na rozdíl od identifikace kolokací na základě frekvence a/nebo výskytu jednotlivých slovních kombinací (ve smyslu tokenů) se Bernardiniová rozhodla pro alternativní extrakci kolokací jakožto typů (Bernardini 2007: 3). Využila k tomu předem identifikované kombinace slovních druhů (na základě jiných studií a kolokačního slovníku), jako např. sekvence dvou substantiv, mezi nimiž může stát předložka či spojka *N prep/conj N*. Nejprve v malém recipročním korpusu pro každý jazyk zvlášť vyhledala všechny slovní kombinace odpovídající těmto POS-vzorcům, čímž dostala seznam kolokací vyskytujících se v překladu a nepřekladu, a poté **v referenčních korpusech** vypočítala jejich frekvenci a seřadila je na základě asociační míry MI (*Mutual Information*). Tím získala jakýsi repozitář zavedených kolokací zvoleného typu v daném jazyce. K další analýze pak byly využity jen ty kombinace s hodnotou MI vyšší než 2 a frekvencí vyšší než 1. Výsledné seznamy v rámci každého POS-vzorce pak Bernardiniová porovnala s ohledem na výskyt v překladech a nepřekladech a dostala údaje pro každý vzorec z hlediska hodnoty MI. V případě příkladové konstrukce *N prep/conj N* byl rozdíl v hodnotě MI statisticky signifikantní ve prospěch překladové italštiny. Na základě toho pak Bernardiniová postuluje závěr, že italští překladatelé mají tendenci využívat tento kolokační vzorec častěji než pisatelé původních italských děl. Toto zjištění dále ověřuje na malém paralelním korpusu v rámci recipročního korpusu, kde identifikuje posuny, k nimž dochází ve prospěch většího výskytu kolokačních struktur v překladech.

Postup Bernardiniové představuje zajímavý způsob, jak se vypořádat s nedostatkem dat – malý recipročí korpus zde slouží jen jako vodítko k výběru POS-vzorců a jako ověřovací paralelní korpus pro kvalitativní sondu v případě konkrétního POS-vzorce. Možný problém lze spatřovat v tom, že referenční korpusy nejsou stejného žánru: recipročí korpus je beletristický, kdežto korpus Repubblica obsahuje pouze publicistiku a BNC více žánrů. Bernardiniová argumentuje tím, že referenční korpusy slouží jen jako zrcadlo toho, co je v jazyce vnímáno jako zavedené a typické, ovšem zde je třeba namítnout, že co je typické pro publicistiku, nemusí platit pro beletrii. Jak vyplynulo z dosavadních zkoumání na překladové češtině, žánr (či textový typ) hraje při výzkumu velkou roli, proto je vhodné zkoumat texty různých žánrů zvlášť. Na druhou stranu je třeba říci, že výhody recipročního, byť malého korpusu (zejména možnost využití paralelní části pro analýzu zdrojových textů) jsou v tomto ohledu s velkými jednojazyčnými korpusy nesrovnatelné a umožňují přesnější interpretaci a explanaci zjištěných jevů.

Kolokace v překladovém jazyce si jako téma své studie zvolila také Dayrellová (2007), jejímž cílem bylo ověřit, zda jsou kolokační vzorce (*collocational patterns*)

v překladu méně různorodé (tj. zda slova mají méně různých kolokátů) a zda jsou tyto kolokáty rovnoměrně distribuovány nebo zda převažuje několik málo vysoce frekventovaných kolokátů. Dayrellová zvolila kolokace se základovým substantivem. Za tímto účelem vybrala ze svého milionového srovnatelného korpusu brazilské portugalštiny všechna substantiva s frekvencí nad 200 a z nich zvolila deset, která se srovnatelně vyskytují v překladové i nepřekladové části. Poté vyhledala jejich kolokáty (v rozmezí 4 pozic vpravo i vlevo, MI > 3) a porovnala jejich výskyt a distribuci. V 60 % případů (tedy u šesti slov z deseti) pozorovala Dayrellová u překladů menší počet kolokátů a v 70 % vykazovaly překladové texty tendenci k menšímu počtu frekventovanějších kolokátů, což by ukazovalo na normalizaci (ve smyslu preference frekventovanějších kolokátů na úkor méně častých a netypických). Sama Dayrellová však uznává, že studie byla provedena na malém počtu slov a výsledky mohou být ovlivněny i velikostí a složením korpusu, případně žánrem (pouze beletrie).

Unique items hypothesis

Univerzálie s názvem *unique items hypothesis* (někdy se k ní odkazuje také jako k *under-representation*) by se dala do češtiny přeložit jako hypotéza o jedinečných jednotkách. Její autorkou je Sonja Tirkkonen-Conditová (2000, 2002, 2004), která se ve svém výzkumu zaměřuje na finštinu. S odkazem na předpokládanou standardizaci a větší „normálnost“ překladových textů vychází z hypotézy, že překladatelé zcela nevyužívají jazykový potenciál cílového jazyka. Ve své teorii se Tirkkonen-Conditová inspirovala závěry výzkumu Reissové (1971), která ve své teorii kritiky překladu hovoří o jednoduchém testu, jenž je založený na tzv. chybějících slovech. Tím se myslí nejfrekventovanější slova v cílovém jazyce, která nemají přímočarý překladový protějšek ve zdrojovém jazyce. Jejich výskyt v překladech pak může poukázat na to, nakolik je překladatel fundovaný a schopný vybavit si tyto výrazy, i když pro ně ve zdrojovém textu nenachází žádný předobraz.

Tirkkonen-Conditová na tomto základě předkládá hypotézu, že překladové texty budou mít menší výskyt takových jazykových prostředků, které postrádají jednoznačné jazykové protějšky ve zdrojovém jazyce, jež by mohly posloužit jako překladové ekvivalenty. Tyto **jedinečné jednotky** nejsou nepřeložitelné a dokonce mohou být velmi frekventované a v cílovém jazyce typické; jsou však netypické a jedinečné ve smyslu svého překladového potenciálu, neboť se nedají v obou jazycích vyjádřit obdobně. Nabízí se samozřejmě hned klíčová otázka, jakým způsobem lze tyto jedinečné jednotky určit. Tirkkonen-Conditová si uvědomuje obtížnost tohoto úkolu a upozorňuje, že tyto jednotky budou vždy vázány na konkrétní jazykové páry. Sama je ve svých studiích vybírá intuitivně (2002: 214) a zahrnuje mezi ně jak konkrétní lexikální jednotky a idiomy, tak i nejrůznější částice a gramatické prostředky. Ve finštině tak jde především o určitá slovesa, která v sobě zahrnují význam dostatečnosti (*ehtii*, *jaksaa*, *uskaltaa* atd.) a klitika (částice *-kin* a *-han*). V obou případech – u sloves i částic – se hypotéza o jedinečných jednotkách potvrzuje, výskyt těchto slov se v překladech liší nejen co do frekvence, ale také co do kolokačního potenciálu. Pro výzkum v dalších jazycích navrhuje Tirkkonen-Conditová jako jedinečnou jednotku

například deminutiva, která jsou velmi frekventovaná a různorodá ve slovanských jazycích (především v ruštině) oproti angličtině.

Obdobný výzkum provedla na finštině i Mauranenová (2000), která se podobně jako Bakerová (viz výše) zaměřila na **víceslovné výrazy s metatextovou funkcí** (vztahují se k textu a organizují jej). Srovnání provedla na textech akademických a populárně naučných a došla k následujícím závěrům: výrazy vztahující se k textu samotnému se podle očekávání častěji vyskytují v akademických textech a také častěji v textech překladových než nepřekladových. Víceslovné kombinace v překladech Mauranenová charakterizuje jako méně jednoznačné a méně stabilní (*strange strings in translated language*), přičemž u některých přibližných synonym (*near-synonyms*) popisuje tendenci, kdy jeden z výrazů je typický pro překlady a druhý pro nepřeklady. Dále Mauranenová uvádí ve shodě s hypotézou o jedinečných jednotkách, že specifický finský výraz *toisaalta* (jenž by se dal přibližně přeložit jako „na druhé straně“) a jeho kombinace je v překladech do finštiny daleko méně používán (a v překladech z finštiny vynecháván).

4.4.2 Výběr slovních kombinací a jejich analýza

Jak vyplývá z uvedeného shrnutí, k výzkumu víceslovných jednotek v překladovém jazyce přistupují badatelé různě, a to jak z hlediska metody, tak i výchozí hypotézy/univerzálie. Pro účely výzkumu slovních kombinací v překladové češtině v této práci byl zvolen postup, jenž vychází z nově provedené kvantitativní analýzy n-gramů. Jeho cílem je identifikovat a popsat takové slovní kombinace (ustálené lexikální vzorce), které jsou typické pro překladový jazyk a mají přitom obecnou platnost (tj. nejsou spjaty pouze s konkrétním textem či tématem). Při výběru relevantních kombinací byly využity i výsledky analýzy POS-gramů (viz kapitola 4.1.2).

Výhoda tohoto postupu plyne ze zpracování velkého množství dat a ze snahy o co možná nejobjektivnější srovnání n-gramů v překladových a nepřekladových textech. Z hlediska zmiňovaných univerzálií pak můžeme zaměřit pozornost buď na výskyt a povahu ustálených struktur, jejichž nadužívání by se dalo vztáhnout k normalizaci, nebo na výrazy, které jsou v překladech naopak podužívané a zároveň tvoří nedílnou součást jazykového repertoáru češtiny (v závislosti na jejich překladovém potenciálu bychom je následně mohli zařadit mezi jedinečné jednotky).

Analýza slovních kombinací na základě n-gramů

Jako základ slovních kombinací byly zvoleny 4-gramy (zdůvodnění výběru právě této délky n-gramu viz s. 83 a graf 4.18). Postup **extrakce 4-gramů** z korpusu Jerome byl následující:

1. V obou zkoumaných částech korpusu Jerome (beletrii a odborné literatuře) byly vyhledány všechny sekvence čtyř po sobě jdoucích slov, s vyloučením interpunkce a zohledněním velikosti písmen, jež mají frekvencí 2 a vyšší (tj. bez hapaxů).

2. Následně byl porovnán výskyt těchto 4-gramů v překladových a nepřekladových textech, přičemž v dalším výběru zůstaly pouze ty, které se vyskytly alespoň pětkrát v jednom z porovnávaných souborů (tato hranice je arbitrární s cílem lépe odfiltrovat nepříliš časté kombinace).

3. 4-gramy byly poté seřazeny podle velikosti rozdílu ve výskytu v překladech a nepřekladech (na základě poměru frekvencí) a roztříděny do skupin (4-gramy unikátní pro jeden ze souborů, 4-gramy prominentně se vyskytující v jednom ze souborů apod., viz dále)

4. V každé skupině pak byly identifikovány takové 4-gramy, které splňují následující kritéria:

   - nezahrnují vlastní jméno (týká se především beletrie a n-gramů vztahujících se ke konkrétním postavám či reáliím, např. *zeptal se pan Prag*, *Váš oddaný Karel Čapek*)
   - neobsahují pouze či převážně synsémantika (např. *co se u nás*, *jak by se na*)
   - neodkazují k jedinému textu či tématu (např. *moučkový a vanilkový cukr*, *v našich klimatických podmínkách*, *po druhé světové válce*)

Bližší pozornost je pak věnována výběrově jen těm nejzajímavějším slovním kombinacím, neboť cílem tohoto postupu není vyčerpávajícím způsobem popsat 4-gramy v překladové a nepřekladové literatuře, nýbrž zachytit v datech takové struktury a lexikální vzorce, které mají co nejvíce všeobecnou platnost (nejsou tolik podmíněny složením korpusu a konkrétním textem) a mohou poukazovat na odlišné trendy v překladovém jazyce. Jde především o víceslovné jednotky idiomatické povahy[17] (viz Tirkkonen-Condit 2002), kombinace s textově strukturující funkcí (viz Mauranen 2000, Baker 2004) a kombinace poukazující na určitý specifický vzorec, který je vyděluje od ostatních 4-gramů (např. repetitivnost prvků, viz dále). Rozdíly jsou statisticky testovány pomocí testu chí-kvadrát.

Podíváme-li se nejprve na **shrnutí počtu 4-gramů**, dostaneme následující přehled (viz tabulka 4.68). Na první pohled patrný rozdíl v počtu 4-gramů v beletrii a odborné literatuře pramení z nestejné velikosti subkorpusů (beletristický je takřka dvojnásobně větší). Tento údaj v absolutní frekvenci zde však není relevantním faktorem, podstatnější jsou poměry v rámci vymezených skupin. Skupiny označené jako A a E zahrnují ty n-gramy, které mají v druhém ze souborů nulový výskyt (nebo se vyskytly pouze jednou a propadly tedy sítem jako hapaxy). Skupiny B a D zahrnují horních 5 % 4-gramů, seřadíme-li je podle poměru frekvencí (výsledkem jsou tedy slova s největším rozdílem ve výskytu, jež se však objevují v obou souborech). Zbývající společné n-gramy pak tvoří skupinu C.

Jak vyplývá z přehledu, v beletrii najdeme dvakrát více společných n-gramů než v odborné literatuře ($\chi^2 = 4\ 856{,}12$, $p < 0{,}001$), což můžeme chápat jako další potvrzení toho, že beletrie tvoří homogennější celek, kdežto v rámci odborné literatury

[17] Slovní kombinace zde označuji za frazémy ve shodě se Slovníkem české frazeologie a idiomatiky (Čermák, Hronek & Machač 2009).

4-gramy	beletrie počet typů	%	odborná počet typů	%
A – pouze v překladech	10 730	**25,44**	7 837	31,42
B – výrazně v překladech	1 324	3,14	404	1,62
C – společné	22 865	**54,21**	6 629	**26,58**
D – výrazně v nepřekladech	1 322	3,13	478	1,92
E – pouze v nepřekladech	5 939	14,08	9 593	38,46
celkem	42 180	100,00	24 941	100,00

Tabulka 4.68: Přehled počtu n-gramů v nepřekladech a překladech

narazíme na rozdíly mezi žánry (v tomto smyslu obory, disciplínami), které nejsou v překladech a nepřekladech zastoupeny stejně. Z hlediska srovnání překladových a nepřekladových textů stojí za povšimnutí, že 4-gramů, které se vyskytují výhradně v překladech je v beletrii statisticky signifikantně více než u nepřekladů ve stejné kategorii, 25,44 % oproti 14,08 % ($\chi^2 = 22{,}13$, $p < 0{,}001$), což by poukazovalo na větší variabilitu 4-gramů v překladech. To je opačný trend než v případě samostatných substantiv (viz tabulka 4.5 v kapitole 4.1), ale stejný jako u verb (viz tabulka 4.7). U odborné literatury je situace opačná (tedy ve shodě s trendem u substantiv): výhradně v nepřekladech najdeme více 4-gramů (38,46 % oproti 31,42 %, $\chi^2 = 271{,}62$, $p < 0{,}001$). Důvody těchto tendencí nelze na základě provedených analýz vysledovat, opět ale narážíme na odlišnou charakteristiku beletrie a odborné literatury.

U všech skupin (kromě společných 4-gramů) byla provedena **bližší analýza první stovky nejfrekventovanějších kombinací**, v níž byly na základě výše uvedených kritérií identifikovány relevantní 4-gramy. Rozlišování 4-gramů ze skupin A a B a skupin D a E se v okamžiku analýzy ukázalo jako nepraktické, neboť i ty relevantní 4-gramy, které byly nalezeny pouze v jednom ze souborů (skupiny A, resp. E), se v určité obměně vyskytly ve skupinách B, resp. D, čímž se jen potvrdila jejich prominence v překladových či nepřekladových textech.

Podíváme-li se nejprve na ty slovní kombinace, jež se vyskytly **prominentně v překladové beletrii**, mezi častými kombinacemi najdeme především určení času a místa (např. *u sebe v pokoji*, *O deset minut později*, *na druhé straně místnosti*), které odrážejí zejména žánr textu (popisné scény v románech apod.). Kromě těchto výrazů se potvrdil i mnohem častější výskyt sloves *podívat se* a *zadívat se* v různých kombinacích, který byl patrný již z analýzy POS-gramů (viz kapitola 4.1.2) a souvisí zřejmě rovněž s tematickým zaměřením textů zahrnutých do překladové části. Mezi potenciálně relevantní slovní kombinace tak lze řadit především idiomatické výrazy, které v následujícím seznamu převažují (seřazeny od nejfrekventovanějšího):

to je v pořádku Různé obměny této fráze mají v překladech výrazně vyšší výskyt: *To je v pořádku* zde najdeme pětkrát častěji (9,35 ipm ku 1,98 ipm, $\chi^2 = 126{,}40$, $p < 0{,}001$). Možným vysvětlením je zde zdrojový jazyk – v naprosté většině jde o angličtinu. Nejedná se však o interferenci v pravém slova smyslu, ale spíše o odlišný

komunikační úzus, který se odráží především v dialozích beletristických textů.

ze všech sil (se) Slovní kombinace s tímto frazémem (např. *ze všech sil se*, *ze všech sil snažila*) se v překladových textech vyskytují více než dvakrát častěji (21,12 ipm oproti 9,01 ipm, $\chi^2 = 130{,}25$, $p < 0{,}001$).

pro všechno na světě Stojí-li tento frazém na začátku věty, objevuje se takřka výlučně pouze v překladech (24 výskytů v 18 různých textech oproti jedinému výskytu v nepřekladové beletrii). Počítáme-li všechny výskyty (bez ohledu na velikost písmen), dojdeme k sedmkrát větší frekvenci v překladech oproti nepřekladové beletrii (2,32 ipm oproti 0,31 ipm, $\chi^2 = 40{,}26$, $p < 0{,}001$).

držet jazyk za zuby Tento frazém (v infinitivním tvaru) je v překladech více než desetkrát častější (1,73 ipm oproti 0,15 ipm, $\chi^2 = 33{,}72$, $p < 0{,}001$) než v původních česky psaných textech. V aktivním tvaru slovesa je trend obdobný (1,95 ipm vůči 0,38, $\chi^2 = 27{,}21$, $p < 0{,}001$). Zdrojovým jazykem je opět zejména angličtina, přičemž bez zdrojových textů můžeme pouze předpokládat, že jde o překlad fráze *keep one's mouth shut*.

Ve skupině slovních kombinací, které se naopak často vyskytují v **nepřekladové beletrii** a v překladu výrazně méně, najdeme následující výrazy (opět řazeny od nejfrekventovanějšího):

v té době (se) Příslovečná určení tohoto typu se substantivem *doba* se vyskytují dvakrát více v nepřekladové beletrii než v překladech (3-gram *v té době*: 83,35 ipm oproti 39,64 ipm, $\chi^2 = 406{,}78$, $p < 0{,}001$), ovšem při bližším pohledu se zde projeví vliv složení korpusu – za častý výskyt může opět větší počet děl literatury faktu FAC (viz s. 73).

v neposlední řadě (i) Tento frazém je také typičtější pro nepřeklady (3,31 ipm oproti 1,88 ipm, $\chi^2 = 9{,}83$, $p < 0{,}01$; v kombinaci se slovem *i* 1,05 oproti 0,34, $\chi^2 = 8{,}71$, $p < 0{,}01$).

od rána do noci Výskyt této slovní kombinace se také omezuje především na nepřekladové texty (1,73 ipm oproti 0,56 ipm, $\chi^2 = 14{,}68$, $p < 0{,}001$).

mezi nebem a zemí Ač nepatří mezi nejfrekventovanější, na tento frazém také narazíme přibližně třikrát častěji v nepřekladových textech (1,62 ipm oproti 0,53, $\chi^2 = 13{,}69$, $p < 0{,}001$). Nejedná se přitom o frázi omezenou na několik málo textů, nýbrž 43 výskytů v nepřekladech najdeme rovnoměrně v celkem 32 textech.

(do) roka a do dne Tento výraz je specifický především pro žánr pohádek a dětskou literaturu (JUN). V překladových textech se s ním nesetkáme takřka vůbec (jediný výskyt), kdežto v nepřekladové beletrii se vyskytl třicetkrát (1,13 ipm) v 16 různých textech.

stále nové a nové Tento typ fráze založené na repetici, jež nese význam zdůraznění, také vykazuje jiné chování v překladech a nepřekladových textech. Tato konkrétní slovní kombinace se v překladu

vyskytla pouze dvakrát, kdežto v nepřekladových textech ji najdeme šestnáctkrát. Vyhledáme-li všechny repetitivní kombinace s adjektivy, tendence vyššího výskytu v nepřekladových textech se potvrdí (26,13 ipm oproti 16,84 ipm, $\chi^2 = 52{,}53$, $p < 0{,}001$). U stejných konstrukcí s adverbii je však výskyt vyrovnaný, proto patrně nelze mluvit o tendenci překladatelů vyhýbat se repetitivnosti (Baker 1993: 244).

V **odborné literatuře** bez ohledu na status najdeme v první stovce slovních kombinací především kolokace a termíny úzce spjaté s tématem textu. Jedním z žánrů s nejvíce ustáleným vyjadřováním (ve smyslu opakujících se n-gramů) jsou kuchařky a atlasy rostlin a hub (např. *posypeme strouhaným sýrem*, *a ve vyhřáté troubě*, *(na) slunci i v polostínu*). I v tomto žánru bychom jistě mohli provést srovnání překladových a nepřekladových textů: známým faktem je např. odlišný úzus, co se týče osoby a čísla v pokynech při vaření v českých a anglických kuchařkách, který by bylo možné porovnat z hlediska překladu. Na takovou analýzu zde však bohužel není prostor, proto se omezím pouze na výrazy obecné a žánrově neomezené, kam spadají především nejrůznější uvozovací věty strukturující text.

V **odborných překladech** více než v původních textech tak narazíme na následující výrazy (seřazeno opět od nejfrekventovanějších):

mít co do činění Kolokace *mít co do činění* se ve všech tvarech slovesa vyskytla v překladech více než dvakrát častěji (3,08 ipm oproti 1,20 ipm, $\chi^2 = 12{,}36$, $p < 0{,}001$), a to celkem v 16 různých textech.

má se za to Tato slovní kombinace se v překladu objevila přibližně osmkrát častěji (2,51 ipm oproti 0,32 ipm, $\chi^2 = 25{,}68$, $p < 0{,}001$), přičemž se nejedná o idiolekt několika málo překladatelů, neboť se tato fráze vyskytla v 21 textech (u 21 různých překladatelů). Nejčastěji se vyskytuje na začátku věty.

jak jsme již viděli Tuto slovní kombinaci, opět v polovině případů na začátku věty, najdeme výrazně častěji v překladech, celkem v 19 různých textech (2,01 ipm oproti 0,43 ipm, $\chi^2 = 11{,}02$, $p < 0{,}001$).

už jsme se zmínili Totéž platí i pro tuto kombinaci (1,32 ipm oproti 0,25 ipm, $\chi^2 = 10{,}24$, $p < 0{,}01$), přičemž převažuje výskyt v podobě *(Jak) už jsme se zmínili.*

V **odborných nepřekladových textech** se ukázaly jako prominentní následující dvě slovní kombinace:

jedná se (především) o Tato slovní kombinace v mnoha variacích (např. *jedná se zejména o*, *jedná se například o*) vykazuje v nepřekladových textech trojnásobnou frekvenci než v překladech (pro frázi *jedná se* činí rozdíl 74,48 ipm oproti 25,27 ipm, $\chi^2 = 386{,}53$, $p < 0{,}001$). Tři čtvrtiny výskytů tvoří v obou skupinách pozice na začátku věty.

s tím souvisí i Tato slovní kombinace má také prominentní výskyt v původních textech, v překladech na ni prakticky nenarazíme (2,07 ipm oproti 0,19 ipm, $\chi^2 = 23{,}37$, $p < 0{,}001$). Takřka vždy stojí na začátku věty.

4.4.3 Shrnutí výzkumu (ne)typických slovních kombinací

Cílem této analýzy bylo odhalit případné odlišnosti mezi překlady a nepřeklady na základě srovnání 4-gramů. Ačkoli metoda analýzy n-gramů se snaží vycházet ze všech dat a poskytnout tak co možná nejobjektivnější pohled, výběr konkrétních slovních kombinací k analýze je i přes předem stanovená kritéria nevyhnutelně subjektivní. S tímto vědomím je možné opatrně formulovat následující tendence, které lze v datech na základě zvoleného postupu vypozorovat.

Nejprve je třeba konstatovat, že analýza n-gramů skutečně poukázala na dílčí rozdíly v použití konkrétních slovních kombinací, nelze však říci, že by se 4-gramy v překladovém jazyce na první pohled výrazně vymykaly. Na základě uvedených slovních kombinací, které tvoří poměrně různorodý celek, je poměrně obtížné formulovat nějaký obecný trend pro překladové texty. Zjevný je zde opět rozdíl mezi žánry (ve smyslu beletrie a odborné literatury), neboť v beletrii nalezneme na základě kritérií zejména slovní kombinace idiomatického charakteru, kdežto v odborné literatuře jde převážně o metatextové formulace a pasivní konstrukce. Na základě analýzy nelze ani konstatovat, že by překladový jazyk vykazoval více či naopak méně idiomatických výrazů, spíše můžeme říci, že preferované frazémy se liší, což však může souviset i se složením korpusu a zaměřením textů. V překladu se prosazují spojení *pro všechno na světě*, *držet jazyk za zuby* či *ze všech sil*, zatímco v nepřekladové beletrii častěji narazíme na výrazy *do roka a do dne*, *od rána do noci* či *mezi nebem a zemí*.

Tendence vyhýbat se repetitivnosti se objevuje u struktur po sobě jdoucích adjektiv (např. *(stále) lepší a lepší, nové a nové*), ovšem v případě adverbií (*víc a víc*, *dál a dál* apod.) k odlišnému trendu nedochází, výskyt je zcela srovnatelný. Jistou tendenci rozlišovat mezi přibližnými synonymy (Mauranen 2000) můžeme pozorovat u výrazů *souviset s* a *mít co do činění s*, jež mají podobný význam, avšak každý ze zkoumaných souborů preferuje jiný z nich. Bez zdrojových textů můžeme opět pouze odhadovat, že vliv na preferenci *mít co do činění* v překladech může mít anglická konstrukce *have something to do with*, která překladatele patrně inspiruje k použití překladového ekvivalentu se stejným základovým slovesem.

Z hlediska jedinečných jednotek nelze říci, že by se na seznamu nacházely typické české jazykové prostředky, které jsou v překladu výrazně méně užívány a zároveň nemají přímočarý ekvivalent ve zdrojovém jazyce (tedy v angličtině jakožto nejvíce zastoupeném jazyce). Za možného kandidáta bychom mohli považovat pouze vazbu *jednat se o*, která nemá v angličtině jednoznačný a převažující ekvivalent a může tedy při překladu z angličtiny zůstat častěji opomenuta, což dokazuje její nižší výskyt v překladových odborných textech.

Některé slovní kombinace také poukázaly na možné lexikální vzorce, které se v překladu odlišují: jako příklad může sloužit spojení se substantivem *doba*, jež

je velmi frekventované v nepřekladové beletrii. Při dodatečném pohledu na jeho kolokace se ukázalo, že v překladu častěji narazíme na spojení *během doby*, zatímco v nepřekladových textech se objevují častěji *v průběhu doby* či *postupem doby*. Může jít o doslovný překlad anglického *during the time*, ovšem pro ověření této teorie bychom potřebovali analyzovat i zdrojové texty a zde použité fráze.

Tím se dostáváme k jednomu z klíčových závěrů. V okamžiku analýzy nižší jazykové roviny, jako jsou konkrétní lexikální vzorce či preference synonymických kombinací, se více než kdy jindy ukazuje zásadní nedostatek jednojazyčného srovnatelného korpusu, kterým je absence originálů. Bez zdrojových textů je velmi obtížné dostat se za hranici deskripce a spolehlivě vysvětlit, proč k výše zmíněným tendencím dochází, zda jde o vliv konkrétního jazykového jevu v konkrétním zdrojovém jazyce či o obecnou překladovou tendenci. Jedním z možných budoucích kroků by tak mělo být ověření zjištěných tendencí na reprezentativním paralelním korpusu, který by byl s to odhalit příčiny těchto jevů. Pozorované rozdíly tak prozatím slouží jako určité vodítko na dosud nepříliš probádaném území víceslovných jednotek v překladu, jako jakési ukazatele na mapě, které naznačují, kde se z hlediska překladu odehrává něco potenciálně zajímavého.

Kapitola 5

Závěr

Cílem této disertace bylo zmapovat a charakterizovat češtinu v překladech na základě rozsáhlých korpusových dat a s využitím kvantitativních metod, tedy tak, jak v českém prostředí doposud systematicky zkoumána nebyla. K tématu překladového jazyka lze jistě přistupovat z mnoha různých úhlů a na základě četných teorií, jak bylo naznačeno v kapitole 2, jež shrnovala dosavadní výzkum v oblasti translatologie. Přístup využitý v této práci v sobě kombinuje možnosti a metody korpusové lingvistiky s nejnovějšími poznatky deskriptivní translatologie, především pak s ohledem na teorii překladových univerzálií.

Výhody kvantitativního přístupu jsou v případě zvoleného tématu nesporné: patří mezi ně především větší míra spolehlivosti a autentičnosti rozsáhlých dat, uplatnění statistických metod a snížení rizika subjektivního faktoru při výzkumu. Kombinace těchto rysů umožňuje badateli alespoň částečnou generalizaci výsledných zjištění, v tomto případě ve vztahu k překladové češtině jako celku, nikoli jen k několika málo konkrétním překladovým textům.

Ve všech výzkumech kvantitativní povahy hrají přitom klíčovou roli především dvě věci – **zvolená data**, především co do velikosti vzorku, a operacionalizace výzkumné otázky. Co se týče prvního bodu, jako datová základna pro tento výzkum posloužil korpus Jerome, rozsáhlý jednojazyčný srovnatelný korpus, který byl sestaven přímo za účelem srovnávání překladové a nepřekladové češtiny. Ačkoli jeho budování předcházela pečlivá příprava a volba kritérií, podle nichž byly texty do korpusu vybírány, v průběhu výzkumu se ukázaly i jeho limity, zvláště s ohledem na vliv určitých žánrů či textových typů (např. literatury faktu zařazené mezi beletrii). Snaha vykompenzovat převažující zdrojový jazyk textů (angličtinu) v podobě malého vyváženého subkorpusu se také ukázala být v určitých aspektech nedostatečná – malá velikost subkorpusu v řádu jednotek milionů totiž znemožňuje podrobnější statistické testování a srovnání výsledků s rozsáhlejšími daty. I přes tyto nedostatky však korpus Jerome poskytl dostatečný zdroj informací na to, abychom na základě testovaných hypotéz mohli konstatovat, jaká čeština v překladech je a co ji charakterizuje.

Druhým klíčovým bodem kvantitativního výzkumu je **operacionalizace výzkumných otázek**. Inspirací k výběru výzkumných hypotéz byly v případě této práce překladové univerzálie (Baker 1993) a jejich často velmi obecné definice.

Bylo proto nutné je operacionalizovat, tedy zvolit konkrétní jazykové indikátory a formální metody jejich testování. Tento úkol s sebou kromě nevyhnutelného redukcionismu vždy nese jistý subjektivní faktor, neboť není prakticky možné předem určit, zda byla operacionalizace přiměřená či nikoli. Hlavním cílem při výběru jazykových indikátorů jednotlivých testovaných rysů proto bylo zahrnout co možná nejvíce různých ukazatelů a testů a tím eliminovat riziko nevhodně zvolené metody. I přes tato opatření je však třeba brát v potaz, že zvolené operacionalizace jsou zcela jistě jen jedny z možných, a s tímto vědomím přistupovat k interpretaci výsledků.

Shrneme-li tedy nejpodstatnější závěry, které z výzkumu vyplynuly, dojdeme k následujícím charakteristickým rysům překladové češtiny, jak je zachycena v korpusu Jerome. Z pohledu **simplifikace**, za jejíž možné indikátory zde byly považovány menší bohatost a pestrost lexikonu, menší hutnost vyjadřování a větší srozumitelnost textu, se překladová čeština jeví na základě pěti ze šesti testů skutečně nepatrně jednodušší, např. obsahuje kratší věty, častěji používá velmi frekventovaná slova a zahrnuje méně autosémantik než čeština nepřekladová. Všechny rozdíly se ukázaly být statisticky signifikantní, ovšem z hlediska věcné významnosti nikterak dramatické. Obdobný výsledek pozorujeme i v případě porovnání frekvenční distribuce slovních druhů a jejich kombinací, u nichž lze rovněž vysledovat odlišné tendence (např. překlady obsahují v beletrii i odborné literatuře přibližně o 2 % více substantiv a o 2 % méně verb), ovšem tento rozdíl nebude pro čtenáře patrně vůbec postřehnutelný.

Ve shodě s hypotézou další zkoumané univerzálie **konvergence** lze říci, že beletristické překladové texty jsou si na základě provedené metody PCA podobnější než texty nepřekladové, jinými slovy mají tendenci dosahovat v testech obdobných hodnot. V odborné literatuře na základě téže analýzy ke konvergenci u překladových textů nedochází. Stejně jako v případě simplifikace lze výraznější rozdíl pozorovat mezi beletrií a odbornou literaturou než mezi překlady a nepřeklady. Závěrečná **analýza n-gramů** pak poukázala na dílčí odlišnosti v překladové češtině, především v preferenci idiomatických výrazů a přibližných synonym, a znovu upozornila na zásadní vliv složení korpusu.

Zbývá tedy celkově vyhodnotit **platnost výchozí hypotézy**. S vědomím všech výše zmíněných faktorů je možné konstatovat, že překladová čeština se od nepřekladové češtiny skutečně liší, ale hned vzápětí je třeba dodat, že odhalené rozdíly zdaleka nejsou tak výrazné a zásadní, jak by se na základě formulovaných hypotéz a předchozích translatologických prací mohlo zdát. Vzhledem k tomu, že se u obou zkoumaných souborů (překladů i nepřekladů) jedná stále o tentýž jazyk, o češtinu, nelze ani očekávat, že rozdíly budou na první pohled zarážející, je tedy poměrně obtížné vyjádřit, jak výrazné ony odlišnosti v překladové češtině jsou.

Vhodným referenčním rámcem zde však může být **rozdíl mezi textovými typy**, beletrií a odbornou literaturou, které v rámci tohoto výzkumu prošly stejným testováním, a je tedy možné výsledky navzájem porovnávat. V takřka všech provedených testech se rozdíly mezi beletrií a odbornou literaturou ukázaly být jako výraznější a prominentnější než v případě srovnání překladů a nepřekladů v rámci žánru. Nejlepším názorným příkladem v tomto ohledu byla metoda PCA, která na základě zvolených rysů seskupila dohromady nikoli překlady a nepřeklady, ale

spíše beletrii a odbornou literaturu (viz graf 4.67). Obdobné výsledky pozorujeme u všech provedených testů. Je tedy pravděpodobné, že čtenář na základě zvolených jazykových indikátorů mnohem lépe postřehne rozdíl mezi beletrií a odbornou literaturou než mezi překlady a nepřeklady.

To samozřejmě neznamená, že by překladová čeština nedisponovala i takovými rysy, které ji mohou do jisté míry identifikovat – nikoli však na vyšších rovinách zobecnění (jaké zde byly především zkoumány), ale spíše na úrovni konkrétních slovních kombinací a lexikálních vzorců v rámci textu. Dá se přitom konstatovat, že čím níže badatel při výzkumu sestupuje, tím větší roli zde hraje složení korpusu (především z hlediska žánru, zdrojového jazyka apod.). Především při výzkumu typických slovních kombinací na základě n-gramů se tak ukázalo, že k lepší interpretaci zjištěných rozdílů by bylo zapotřebí analyzovat kromě referenčních nepřekladových textů také texty zdrojové, tedy originály zahrnutých překladů. Těmi však jednojazyčný srovnatelný korpus nedisponuje, proto tato oblast zůstává otevřena budoucím badatelům.

S posledním bodem úzce souvisejí **možnosti budoucího výzkumu** v oblasti překladové češtiny a rysů překladového jazyka. Je nesporné, že využití reprezentativního paralelního či ještě lépe recipročního korpusu by přineslo nové poznatky (především v oblasti S-univerzálií) a zároveň umožnilo přesnější interpretaci zde zjištěných fakt. Jednojazyčný srovnatelný korpus Jerome má z hlediska svého složení přes veškerou snahu také své limity, což platí i pro jeho subkorpus vyvážený z hlediska zdrojového jazyka. Zvláště v případě subkorpusu by bylo vhodné shromáždit větší množství překladů z více jazyků, aby jeho výsledná velikost umožnila pokročilejší statistické testování a ověřování možného vlivu interference.

Z hlediska témat, jež se ukázala jako potenciální zdroj dalších informací o překladové češtině, ale vzhledem k rozsahu práce zde nemohla být dále rozvíjena, stojí za zmínku především otázka víceslovných jednotek v překladu. Zde provedená analýza naznačila, že na úrovni konkrétních lexikálních jednotek a jejich kombinací může docházet k rozdílům, jež by si zasloužily další výzkum. Jednou z možností, jak víceslovné jednotky v překladu zkoumat, přitom může být kvantitativní analýza variability kontextu (viz Cvrček 2013) nebo dílčí kvalitativní studie věnované konkrétním kolokacím v překladové a nepřekladové češtině.

Závěrem je možné konstatovat, že čeština v překladech představuje (nejen) pro korpusovětranslatologický výzkum bohatý inspirační zdroj, který touto disertační prací zdaleka nebyl vyčerpán. Ačkoli výsledky naznačily, že překladová čeština není zcela odlišná od češtiny v původně českých dílech, současně s tím ukázaly, že i přesto má lingvistům a translatologům z výzkumného hlediska mnoho co nabídnout.

Literatura

Aijmer, K., Altenberg, B. & Johansson, S. (Eds.) (1996). *Languages in Contrast: Papers from a Symposium on text-based Cross-linguistics Studies*. Lund: Lund University Press.

Anderman, G. & Rogers, M. (2008). *Incorporating Corpora: The Linguist and the Translator*. Clevendon: Multilingual Matters.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 233–250). Amsterdam-Philadelphia: John Benjamins.

Baker, M. (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*, 7(2), 223–243.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and Translation: Studies in language engineering, in honour of Juan C. Sager* (pp. 175–86). Amsterdam: John Benjamins.

Baker, M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics*, 9(2), 167–193.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Baroni, M. & Bernardini, S. (2006). A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3), 259–274.

Becher, V. (2010). Abandoning the Notion of "Translation-Inherent"Explicitation. Against a Dogma of Translation Studies. *Across Languages and Cultures*, 11(1), 1–28.

Bernardini, S. (2007). Collocations in Translated Language. Combining parallel, comparable and reference corpora. Příspěvek přednesený na konferenci *Corpus Linguistics 2007*, Lancaster, UK. Dostupný z: `http://ucrel.lancs.ac.uk/publications/CL2007/paper/15_Paper.pdf`

Bernardini, S. & Zanettin, F. (2004). When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals. In A. Mauranen & P. Kujamäki (Eds.), *Translation Universals – Do They Exist?* (pp. 53–62), Amsterdam-Philadelphia: John Benjamins.

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

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

Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-kulka (Eds.), *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies* (pp. 17–35). Tübingen: Gunter Narr.

Cantos Goméz, P. (2013). *Statistical Methods in Language and Linguistic Research*. Sheffield: Equinox.

Catford, J. C. (1965). *A Linguistics Theory of Translation*. London: Oxford University Press.

Corpas Pastor, G., Mitkov, R., Afzal, N. & Pekar, V. (2008). Translation universals: Do they exist? A corpus-based NLP study of convergence and simplification. *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-08)*. Waikiki Hawaii.

Cvrček, V. (2013). *Kvantitativní analýza kontextu*. Praha: Nakladatelství Lidové noviny.

Cvrček, V. et al. (2010). *Mluvnice současné češtiny*. Praha: Nakladatelství Karolinum.

Cvrček V., Čermák, F. & Křen M. (2007). Statistické aspekty jazyka Karla Čapka, zvláště jeho lexikonu. In F. Čermák (Ed.), *Slovník Karla Čapka* (pp. 673–690). Praha: Nakladatelství Lidové noviny.

Cvrček, V. & Fidler, M. (v tisku). A data-driven analysis of reader viewpoints: Reconstructing the historical reader using keyword analysis. *Journal of Slavic Linguistics*.

Cvrček, V. & Chlumská, L. (v tisku). Simplification in translated Czech: a new approach to type-token ratio. *Russian Linguistics*, 39(3).

Cvrček, V. & Kováříková, D. (2011). Možnosti a meze korpusové lingvistiky. *Naše řeč*, 94(3), 113–133.

Cvrček, V. & Václavík, J. (v tisku). Jednoznačnost a kontext. Kvantitativní studie. *Korpus – gramatika – axiologie*.

Čermák, F., Hronek J. & Machač, J. (Eds.) (2007). *Slovník české frazeologie a idiomatiky 1–4 (SČFI)*. Praha: LEDA.

Dayrell, C. (2007). A quantitative approach to compare collocational patterns in translated and non-translated texts. *International Journal of Corpus Linguistics*, 12(3), 375–414.

De Sutter, G., Cappelle, B. & Loock, R. (2013). Competing Motivations in Dutch/French Legal Translation: a Quantitative Corpus-Based Study of the Interaction between Interference and Normalisation. Příspěvek přednesený na konferenci *ICLC 7 – UCCTS 3*, 11–13. 7. 2013, Gent, Belgie.

De Sutter, G., Goethals, P., Leuschner, T. & Vandepitte, S. (2012). Towards methodologically more rigorous corpus-based translation studies. *Across Languages and Cultures*, 13(2), 137–143.

Duguid, A. (2010). Newspapers discourse informalisation: a diachronic comparison from keywords. *Corpora*, 5(2), 109–138.

Dušková, L. (1988). *Mluvnice současné angličtiny na pozadí češtiny*. Praha: Academia.

Fernandes, L. (2006). Corpora in Translation Studies: revisiting Baker's typology. *Fragmentos*, 30, 87–95.

Frawley, W. (1984). *Prolegomenon to a Theory of Translation Translation: literary, linguistic and philosophical perspectives*. Newark: University od Delaware Press.

Friedbichler, I. & Friedbichler, M. (1997). The potential of domainspecific target-language corpora for the translator?s workbench. Příspěvek přednesený na konferenci *Conference on Corpus Use and Learning to Translate*, Bertinoro.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindquist (Eds.), *Translation Studies in Scandinavia* (pp. 88–95). Lund: CWK Gleerup.

Gentzler, E. (1993). *Contemporary Translation Theories*. London and New York: Routledge.

Goethals, P. (2007). Corpus-driven Hypothesis Generation in Translation Studies, Contrastive Linguistics and Text Linguistics: A Case Study of Demonstratives in Spanish and Dutch Parallel Texts. *Belgian Journal of Linguistics*, 21, 87–103.

Grabowski, Ł. (2012). On translation universals in selected contemporary Polish literary translations. *Studies in Polish Linguistics*, 7(1), 165–183.

Granger, S. (1996). From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (Eds.), *Languages in Contrast: Papers from a Symposium on text-based Cross-linguistics Studies* (1994) (pp. 38–51). Lund: Lund University Press.

Granger, S. (2013). Tracking the third code: A crosslinguistic corpus-driven approach to discourse markers. Příspěvek přednesený na konferenci *ICLC 7 - UCCTS 3*, 11-13. 7. 2013, Gent, Belgie.

Halverson, S. (2003). The cognitive basis of translation universals. *Target*, 15(2), 197–241.

Hareide, L. & Hofland, K. (2012). Compiling a Norwegian-Spanish Parallel Corpus: methods and challenges. In Michael P. Oakes & Meng Ji (Eds.), *Quantitative Methods in Corpus-Based Translation Studies* (pp. 75–113). Amsterdam: John Benjamins.

Hendl, J. (2009). *Přehled statistických metod*. Praha: Portál.

Hermans, T. (1999). *Translation in Systems*. Manchester: St. Jerome.

Holmes, J. S. (1972). The name and nature of translation studies. In L. Venuti (Ed.), *The Translation Studies Reader* (pp. 172–185). London and New York: Routledge.

House, J. (1997). *Translation Quality Assessment: A Model Revisited*. Tübingen: Gunter Narr.

House, J. (2008). Beyond Intervention: Universals in Translation? *trans-kom: Zeitschrift für Translationswissenschaft und Fachkommunikation*, 1(1), 6–19.

Hrala, M. et al. (2002). *Kapitoly z dějin českého překladu*. Praha: Karolinum.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Chesterman, A. (2004a). Hypotheses about translation universals. In G. Hanse, K. Malmkjær & D. Gile (Eds.), *Claims, Changes and Challenges in Translation*

Studies. Selected Contributions from the EST Congress Copenhagen 2001. (pp. 1–14). Amsterdam-Philadelphia: John Benjamins.

Chesterman, A. (2004b). Beyond the Particular. In A. Mauranen & P. Kujamäki (Eds.), *Translation Universals – Do They Exist?* (pp. 33–49). Amsterdam-Philadelphia: John Benjamins.

Chlumská, L. (2013). JEROME: jednojazyčný srovnatelný korpus pro výzkum překladové češtiny. Ústav Českého národního korpusu FF UK, Praha 2013. Dostupný z `http://www.korpus.cz`.

Chlumská, L. (2014). Není *korpus* jako *korpus*. Korpusy v kontrastivní lingvistice a translatologii. Časopis pro moderní filologii, 96(2), 221–232.

Chlumská, L. & Kováříková, D. (2009). The reflection of linguistics tradition in translation. In F. Čermák, P. Corness & A. Klégr (Eds.), *InterCorp: Exploring a Multilingual Corpus* (pp. 146–158). Praha: Nakladatelství Lidové noviny.

Chlumská, L. & Richterová, O. (2014). Jak zkoumat překladovou češtinu. Výzkum simplifikace na korpusu Jerome. *Korpus – gramatika – axiologie*, 9, 16–29.

Ilisei, I., Inkpen, D., Corpas Pastor, G. & Mitkov, R. (2010). Identification of Translationese: A Machine Learning Approach. Příspěvek přednesený na konferenci Cicling 2010.

Jakobson, R. (1959/2000). On linguistic aspects of translation. In L. Venuti (Ed.), *The Translation Studies Reader* (pp. 113–118). London and New York: Routledge.

Johansson, S. & Oksenfjell, S. (1998). *Corpora and Cross-linguistic Research*. Amsterdam-Atlanta: Rodopi.

Johansson, S. (1998). On the role of corpora in cross-linguistic research. In S. Johansson & S. Oksefjell (Eds.), *Corpora and Cross-linguistics Research* (pp. 3–24). Amsterdam-Atlanta: Rodopi.

Johansson, S. (2004). Why change the subject? On changes in subject selection in translation from English into Norwegian. *Target*, 16(1), 29–52.

Kenny, D. (1998). Creatures of Habit? What Translators Usually Do with Words. *Meta*, 43(4), 515–523.

Kenny, D. (2001). *Lexis and Creativity in Translation*. A Corpus-based Study. Manchester: St. Jerome.

Kočová, I. (2009). *Legis Speak and Creative Translation Solutions in EU Texts: Strictly Legislative Versus Related Texts* (Diplomová práce). Vedoucí DP: Mgr. Renata Kamenická, Ph.D. Brno: FF MU.

Kruger, A. (2002). Corpus-based translation research: Its development and implications for general, literary and Bible translation. *Acta Theologica Supplementum*, 2, 70–106.

Kubáčková, J. (2008). *Generalizace a specifikace lexikálního významu v soudobém uměleckém překladu* (Diplomová práce). Vedoucí DP: PhDr. Zuzana Jettmarová, M.Sc. Praha: FF UK.

Lapshinova-Koltunski, E. (2015). Variation in translation: evidence from corpora. In C. Fantionuoli & F. Zanettin (Eds.), *New directions in corpus-based translation studies* (pp. 93–114). Berlin: Language Science Press.

Laviosa, S. (1998a). The English Comparable Corpus. In L. Bowker, M. Cronin, D. Kenny & J. Pearson (Eds.), *Unity in Diversity?*. Manchester: St. Jerome Publishing.

Laviosa, S. (1998b). Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta: Translator's Journal*, 43(4), 557–571.

Laviosa, S. (2002). *Corpus-based Translation Studies. Theory, findings, applications*. Amsterdam-New York: Rodopi.

Laviosa-Braithwaite, S. (1996). *Investigating Simplification in English Comparable Corpus of Newspaper Articles*. Szombathely: Daniel Berzsenyi College Printing Press.

Levý, J. (1971). Bude teorie překladu užitečná překladatelům? *Bude literární věda exaktní vědou?* (pp. 147–157). Praha: Československý spisovatel. Levý, J. (1983). *Umění překladu*. Praha: Panorama.

Malmkjær, K. (1997). Punctuation in Hans Christian Andersen's Stories and in their Translations into English. In F. Poyatos (Ed.), *Nonverbal Communication and Translation: New Perspectives and Changes in Literature, Interpretation and the Media* (pp. 151–162). Amsterdam: John Benjamins.

Malmkjær, K. (1998). Love thy Neighbour: Will Parallel Corpora Endear Linguists to Translators? *Meta*, 43(4), 534–541.

Mathesius, V. (1947). Přívlastkové ten, ta, to v hovorové češtině. In *Čeština a obecný jazykozpyt* (pp. 185–189). Praha: Melantric.

Mauranen, A. (2000). Strange Strings in Translated Language. A Study on Corpora. In M. Olohan (Ed.), *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects* (pp. 119–141). Manchester: St. Jerome Publishing.

Mauranen, A. & Kujamäki, P. (Eds.) (2004). *Translation Universals - Do They Exist?* Amsterdam-Philadelphia: John Benjamins.

May, R. (1997). Sensible Elocution: How Translation Works in & upon Punctuation. *The Translator*, 3(1), 1–20.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics*. Cambridge: Cambridge University Press.

McEnery, T., Xiao R. & Tono, Y. (2006). *Corpus-based Language Studie: An Advanced Resource Book*. London and New York: Routledge.

McEnery, T.& Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Munday, J. (2008). *Introducing Translation Studies. Theories and applications*. London: Routledge.

Neumann, S. (2014). Beyond translation properties: The contribution of corpus studies to empirical translation theory. Plenární přednáška přednesená na konferenci *UCCTS 4*, Lancaster, UK, 25.7.2014.

Newmark, P. (1981). *Approaches to Translation*. Oxford and New York: Pergamon.

Newmark, P. (1988). *A Textbook of Translation*. New York and London: Prentice-Hall.

Nida, E. (1964). *Towards a science of translating*. Leiden: E. J. Brill.

Nida, E. & Taber, C. R. (1969). *The Theory and Practise of Translation*. Leiden: E. J. Brill.

Nord, Ch. (1988). *Textanalyze und Übersetzen: Theuretische Grundlagen, Methode und diduktische Anwendung einer übersetzungsrelevanten Textanalyze*. Heidelberg: J. Groos.

Oakes, Michael P. & Ji, M. (2012). *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam-Philadelphia: John Benjamins.

Olohan, M. (2004). *Introducing Corpora in Translation Studies*. London: Routledge.

Øverås, L. (1998). In Search of the Third Code: An Investigation of Norms in Literary Translation. *Meta: Translator's Journal*, 43(4), 557–570.

Paloposki, O. (2001). Enriching translations, simplified language? An alternative viewpoint to lexical simplification. *Target*, 13(2), 265–288.

Partington, A. (2010). Modern Diachronic Corpus-Assisted Discourse Studies (MDCADS) on UK newspapers – an overview of the project. *Corpora*, 5(2), 83–108.

Polišenská, M. (2010). *Translation Universals in the English and Spanish Translations of Saturnin by Zdeněk Jirotka* (Diplomová práce). Vedoucí DP: Mgr. Renata Kamenická, Ph.D. Brno: FF MU.

Popovič, A. (1975). *Teória umeleckého prekladu: aspekty textu a literárnej metakomunikácie*. Bratislava: Tatran.

Puurtinen, T. (2003). Genre-specific Features of Translationese? Linguistic Differences between Translated and Non-translated Finnish Children's Literature. *Literary and Linguistic Computing*, 18(4), 389–406.

Pym, A. (2008). On Toury's laws of how translators translate. In A. Pym, M. Shlesinger & D. Simeoni (Eds)., *Beyond Descriptive Translation Studies*. Amsterdam-Philadelphia: John Benjamins.

Reiss, K. (1971). *Möglichkeiten und Grenzen der Übersetzungskritik*. München: Max Hueber.

Rodríguez-Castro, M. (2011). Translationese and punctuation. *Translation and Interpreting Studies*, 6(1), 40–61.

Savický, P. & Hlaváčová, J. (2003). Measures of word commonness. *Journal of Quantitative Linguistics*, 9(3), 215–231.

Shamaa, N. (1978). *A Linguistic Analysis of Some Problems of Arabic to English Translation* (Dizertační práce). Oxford: Oxford University.

Shlesinger, M. (1989). *Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts on the Oral-Literate Continuum* (Diplomová práce). Tel Aviv: Tel Aviv University.

Shlesinger, M. (1991). Interpreter Latitude vs. Due Process. Simultaneous and Consecutive Interpretation in Multilingual Trials. In S. Tirkkonen-Condit (Ed.), *Empirical Research in Translation and Intercultural Studies* (pp. 147–155). Tübingen: Gunter Narr.

Smith, E. A. & Senter, R. J. (1967). *Automated Readability Index*. Ohio: Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Paterson Air Force Base.

Soukup, P. (2013). Věcná významnost výsledků a její možnosti měření. *Data a výzkum – SDA Info 2013*, 7(2), 125–148.

Středová, A. (2009). *Explicitation and Implicitation in Non-literary Translations* (Diplomová práce). Vedoucí DP: Mgr. Renata Kamenická, Ph.D. Brno: FF MU.

Stubbs, M. (1986). *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.

Teich, E. (2003). *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Mouton de Gruyter.

Thompson, P. & Sealey, A. (2007). Through children's eyes? Corpus evidence of the features of children's literature Vol. 12 No. 1 1–12

Tirkkonen-Condit, S. (2000). In Search of Translation Universals: Non-equivalence or Unique Items in a Corpus test. Příspěvek přednesený na konferenci *UMIST/UCL Research Models in Translation Studies*, Manchester, 28.–30. 4. 2000.

Tirkkonen-Condit, S. (2002). Translationese – a myth or an empirical fact? A study into the linguistic identifiability of translated language. *Target*, 14(2), 207–220.

Tirkkonen-Condit, S. (2004). Unique Items? Over- or Under-Represented in Translated Language? In A. Mauranen & P. Kujamäki (Eds.), *Translation Universals – Do They Exist?* (pp. 177-185). Amsterdam-Philadelphia: John Benjamins.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam-Philadelphia: John Benjamins.

Toury, G. (1980). *In Search of a Theory of Translation*. Tel Aviv: The Porter Institute for Poetics and Semiotics.

Toury, G. (1995). *Descriptive Translation Studies ?- and Beyond*. Amsterdam-Philadelphia: John Benjamins.

Toury, G. (2004a). Probabilistic Explanations in Translation Studies: Universals – or a Challenge to the Very Concept? In G. Hanse, K. Malmkj(æ)r & D. Gile (Eds.), *Claims, Changes and Challenges in Translation Studies. Selected contributions from the EST Congress Copenhagen 2001* (pp. 15–25). Amsterdam-Philadelphia: John Benjamins.

Toury, G. (2004b). Probabilistic Explanations in Translation Studies: Welcome as they are, would they qualify as universal? In A. Mauranen & P. Kujamäki (Eds.), Translation Universals – Do They Exist? (pp. 15–32). Amsterdam-Philadelphia: John Benjamins.

Tymoczko, M. (1998). Computerized Corpora and the Future of Translation Studies. *Meta: Translator's Journal*, 43(4), 652–660.

Vanderauwera, R. (1985). *Dutch Novels Translated into English: The Transformation of a "Minority"Literature*. Amsterdam: Rodopi.

Venuti, L. (Ed.) (2000). *The Translation Studies Reader*. London and New York: Routledge.

Vermeer, H. & Reiss, K. (1984). *Grundlegung einer ullgemeinen Translationstheorie*. Tübingen: Niemeyer.

Vinay, J.-P. & Darbelnet, J. (1958). *Stylistique comparée du français et de l?anglais: méthode de traduction*. Paris: Didier.

Volín, J. (2007). *Statistické metody ve fonetickém výzkumu*. Praha: Epocha.

Wang, K. & Qin, H. (2010). A Parallel Corpus-based Study of Translational Chinese. In R. Xiao (Ed.), *Using Corpora in Contrastive and Translation Studies* (pp. 164–181). Newcastle upon Tyne: Cambridge Scholars Publishing.

Wang, W. (2006). *A corpus-driven study on translation units in an English-Chinese parallel corpus* (Diplomová práce). University of Birmingham.

Xiao, R. (2010). How different is translated Chinese from native Chinese? *International Journal of Corpus Linguistics*, 15(1), 5–35.

Yule, G. U. (1944). *The statistical study of literary vocabulary.* Cambridge: Cambridge University Press.

Zanettin, F. (2011). Translation and corpus design. *SYNAPS – A Journal of professional Communication*, 26, 14–23.

Zanettin, F. (2013). Corpus Methods for Descriptive Translation Studies. *Procedia – Social and Behavioral Sciences*, 95, 20–32.

Zubatý, J. (1917). Ten. *Naše řeč*, 1(10), 289–293.

Zubatý, J. (1920). Ten nejlepší člověk. *Naše řeč*, 4(3), 74–76.

Seznam tabulek

Seznam obrázků