UNIVERZITA KARLOVA

Filozofická fakulta

Ústav východoevropských studií

Bakalářská práce

Praha 2021

UNIVERZITA KARLOVA

FILOZOFICKÁ FAKULTA

Ústav východoevropských studií

Využití dat z frekvenčních slovníků lotyštiny v chystaném lotyšsko-českém překladovém slovníku

Use of data from Latvian frequency dictionaries in the forthcoming Latvian-Czech bilingual dictionary

Bakalářská práce

Autor práce: Anna Sedláčková

Studijní program: Východoevropská studia (lotyština)

Vedoucí práce: Michal Škrabal, Ph.D.

Rok obhajoby: 2021

Prohlášení

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

V Praze dne 10.5. 2021 Anna Sedláčková……

Bibliografický záznam

SEDLÁČKOVÁ, Anna. Využití dat z frekvenčních slovníků lotyštiny v chystaném lotyšsko-českém překladovém slovníku. Praha, 2021. 47s. Bakalářská práce (Bc.), Univerzita Karlova, Filozofická fakulta, Ústav východoevropských studií, vedoucí práce Michal Škrabal, Ph.D.

Rozsah práce: 82 777 znaků/11421 slov

Abstrakt

Cíl této práce je explicitně vyjádřen v jejím titulu: snažíme se o aplikaci lotyšských frekvenčních dat, dostupných ve specializovaných, tj. frekvenčních slovnících, případě nověji přímo z korpusů, v konkrétním lexikografickém projektu: chystaném lotyšsko-českém překladovém slovníku. Práce se skládá z části teoretické a praktické. První jmenovaná bude obsahovat úvod do problematiky korpusové lingvistiky v lotyšském i českém prostředí, popis zdrojů, popis materiálové základny (čtyři odlišné materiály - dva tištěné zdroje, dva zdroje elektronické) a komparaci současné situace na poli korpusové lingvistiky/lexikografie v Lotyšsku a ČR. Druhá, praktická část je zaměřena na aplikaci získaných dat a jejich popis. Neoddělitelnou součástí praktické části je prezentace získaných dat společně s ukázkou slovníkového hesla. Cílem je poskytnout uživateli konkrétnější představu o podobě hesel v lotyšsko-českém překladovém slovníku a nabídnout možnost aplikace v jiných lexikografických projektech.

Klíčová slova

frekvenční slovníky, lotyština, lotyšsko-český překladový slovník, lexikografie, korpus

Abstract

The aim of this work is explicitly stated in its title: our aim is to apply data from the frequency dictionaires of Latvian language, available in specialized, ie frequency dictionaries, or more recently directly from corpora, in a specific lexicographic project: the upcoming Latvian-Czech dictionary. The thesis consists of theoretical and practical parts. The theoretical part will contain an introduction to the issues of corpus linguistics in Latvia and Czech Republic, a description of sources, a description of the material base (four different materials - two printed sources, two electronic sources) and a comparison of the current situation in corpus linguistics / lexicography in Latvia and the Czech Republic. The second, practical part is focused on the application of the obtained data and their description. An inseparable part of the practical part is the presentation of the obtained data together with a sample dictionary entry. The aim is to provide the user with a more specific idea of the form of entries in the Latvian-Czech translation dictionary and to offer the possibility of application in other lexicographic projects.

Keywords

frequency dictionaries, Latvian language, Latvian-Czech dictionary, lexicography, corpora

Poděkování

Na tomto místě bych chtěla poděkovat svému vedoucímu za cenné připomínky a trpělivé vedení práce. Jsem vděčná za možnost podílet se alespoň malým dílem na čemkoliv, co lotyštinu přiblíží dalším zájemcům.

Děkuji také Robertsi Darģisovi a Kristīne-Levāne Petrové za poskytnutí přístupu k materiálům LSRC. Bez tohoto materiálu by tato práce nemohla vzniknout. Chtěla bych poděkovat také Pavlu Štollovi a Kristīne Ante, kteří mi v průběhu mého studia trpělivě předávali mnoho zajímavých postřehů a rad.

Obsah

Úvod

Frekvenční slovníky se společně s jazykovými korpusy čím dál častěji stávají nepostradatelnou pomůckou nejen pro studenty filologických oborů, ale i pro laické zájemce o příslušný jazyk. Svému potenciálnímu uživateli frekvenční slovník nabízí celou řadu výhod, jako například efektivní osvojování nového lexika nebo praktický přehled reálné aplikovatelnosti osvojené slovní zásoby. Užívání frekvenčního slovníku, případně ad hoc vytvořených frekvenčních soupisů získaných z jazykových korpusů v elektronické podobě (např. Lists[1]) může proces učení zásadně urychlit a zároveň studenta vybavit skutečně adekvátní slovní zásobou jak v počátcích, tak i v pokročilejších fázích studia.

Význam jazykového korpusu ovšem nespočívá pouze v potenciálním využití při osvojování cizího jazyka. Jeho primární funkcí je systematický přehled o užívání slovní zásoby z různých oborů, odvětví, komunikačních situací, období apod. Korpus může nabývat různých podob, které úzce souvisí s povahou shromažďovaného a následně analyzovaného materiálu. V dnešní době se tedy nesetkáváme jen s korpusy kompilovanými výhradně na základě textů, tedy opírajícími se o tištěnou podobu jazyka, ale také s těmi zachycujícími jazyk mluvený, případně též dalšími typy, například s korpusy znakového jazyka či korpusy multimodálními (korpus obsahuje zároveň grafický přepis, fonetickou původní i průvodní nahrávku, někdy také videozáznam)[2]. S ohledem na zaměření této práce považujeme za důležité zdůraznit, že původní účel jazykových korpusů byl čistě praktický: měl posloužit lexikografům k věrnějšímu, objektivnějšímu popisu slovní zásoby daného jazyka, k eliminaci přílišného vlivu slovníkářova idiolektu a introspekce.

Cílem této práce je analyzovat získaná data ze čtyř níže uvedených zdrojů z několika různých období. Jde o data z tištěných i elektronických zdrojů, která budou blíže představena v příslušné sekci. Mají posloužit v chystaném lotyšsko-českém překladovém slovníku (dále: LČPS) v podobě frekvenčního modulu. Primárně nás zajímá slovní zásoba, která se objevuje ve všech užitých zdrojích, svou podobu nezměnila ani po několik desetiletí a je skutečně aktuální i v dnešní době. Ta pro nás představuje kontinuální jádro lotyšského lexikonu, které by mělo být explicitně zobrazeno právě v LČPS.

[1] Lists: Prohlížeč frekvenčních seznamů, SYN2015. [online] Cit. 10.5. 2021. <https://www.korpus.cz/lists>
[2] Více o typech korpusů na: https://www.czechency.org/slovnik/TYPY%20KORPUSŮ

Bylo vytyčeno několik výzkumných otázek, jako např.

- Do jaké míry se jazyk v uvedeném období změnil?
- Jaké lexikum se změnilo? Jak je možné si to vysvětlit?
- Jak velké bude procento specifické slovní zásoby jednotlivých zdrojů a jaké faktory to ovlivňují?
- Budou tento specifický lexikon označovat pouze určitou skupinu slov (určitý obor, odvětví, oblast), nebo bude distribuován rovnoměrně napříč materiálem?
- Vykazuje lotyština tzv. diglosii?
- Do jaké míry bude korpus mluveného slova odrážet specifika hovorového jazyka a jak se jednotky v něm obsažené budou lišit od jiných zkoumaných zdrojů?

Metodologie

Cílem této práce je poskytnout nejen teoretický rámec popisující současný stav lotyšského lexikonu, ale také reálná data získaná ze srovnání různých zdrojů využít pro potřeby aktuálně vznikajícího lotyšsko-českého slovníku. V praktické části proto budou srovnány lexémy získané ze čtyř různých frekvenčních slovníků lotyšského jazyka. Každý jednotlivý materiál bude dopodrobna popsán v sekci Materiálová základna. Ke komparaci byly užity dva tištěné zdroje (Kuzina – dále: K, Jakubaite, dále: J) a dva zdroje elektronické (LVK2018 – dále: LVK, LSRC). Před samotným začátkem práce byla potřeba oba tištěné zdroje zdigitalizovat, protože nebyly elektronicky dostupné. Za výchozí dataset jsme zvolili K; původně bylo naším záměrem pracovat přesně s třemi tisíci lexikálními jednotkami, avizovaných ostatně v plném názvu K. Ve skutečnosti však tento frekvenční slovník obsahoval položek 3288, což bylo nutné zohlednit rovněž u zbylých zdrojů, jinak by zjištěné údaje mohly být zkreslené.

Takto docházíme k elementární statistice, jež nám poslouží k následným úvahám o možných rozdílech mezi psaným a mluveným jazykem a také o možných změnách lexikální frekvenční špičky v průběhu času. Konkrétně jde o údaje, kolik procent slovní zásoby je společných pro všechny čtyři zkoumané zdroje, případně jen pro některé z nich, a jaká část lexémů je vlastní právě jednomu původnímu zdroji. Následně byla data zkoumána na základě svých vlastností (zda má daná jednotka potenciál vyskytovat se spíše v psaném, či mluveném projevu, v jakém žánrovém okruhu se vyskytuje nejčastěji, do jaké sémantické kategorie patří apod.).

Ke srovnání dat byly záměrně vybrány zdroje z různých časových období: nejstarší pokrývá poválečnou slovní zásobu, tj. od 50. let minulého století, kdežto LVK2018 a LSRC

zahrnují slovní zásobu doby posledních let, díky čemuž je možné lexikální jednotky vnímat též v diachronním aspektu. Proto je také možné porovnávat vývoj slovní zásoby, tj. zjišťovat, jaká slova se objevovala jen v určitých obdobích a jaké vlivy na tom měly podíl, které lexikální jednotky se již dnes v prvních třech tisících frekvenční špičky neobjevují apod.).

Lexémy ze všech datasetů byly zaneseny do společného dokumentu a opatřeny informací, do jaké kategorie patří. Kategorií bylo celkem osm (průnik ve všech třech původních zdrojích, tj. slovní zásoba společná všem zdrojům, průnik v K+J, K+LVK+LSRC, J+LRSC, J+LVK+LSRC, K+LRSC, J+K+LVK, LRSC+LVK). Každé z těchto kategorií bylo přiřazeno specifické grafické označení, aby byla zachována přehlednost a zároveň bylo možné jednotky zkoumat na jednom společném místě. Získaná data byla následně statisticky vyhodnocena kvůli zjištění, kolik procent lexikálních jednotek v celém sledovaném materiálu zůstalo ve frekvenční špičce a které jsou vlastní pouze jednomu, dvěma nebo třem zdrojům. Výsledná data budou aplikována v LČPS.

Při psaní práce jsme přihlíželi též k vybrané metalexikografické literatuře, především té zabývající se problematikou dvoujazyčné lexikografie. Jak upozorňuje i S. Nikuļceva (2013) [3], množství těchto odborných příruček je však pro kombinaci lotyština-čeština velmi omezené. Jako důvod uvádí, že lexikografové, kteří se kompilací slovníků zabývají, se věnují primárně praktické činnosti a sestavení obdobných teoretických děl je často opomíjeno. Ze světové lexikografie mezi takové příručky můžeme zařadit např. lexikografickou encyklopedii pod redakčním vedením F. J. Hausmanna (Hausmann et al. 1989–1991)[4], z českého prostředí pak autorským kolektivem sestavený Manuál lexikografie (Čermák, Blatná 1995)[5].

[3] NIKUĻCEVA, S., (2013). Ekvivalence lexémů v Česko-lotyšském slovníku: přístup a zpracování. In P. Štoll et al. *Zkušenosti a vztahy. Lotyšská a česká společnost ve 20. století.* Praha: FF UK, s. 213–225.

[4] HAUSMANN, F. J. – REICHMANN, O. – WIEGAND, H. E. – ZGUSTA, L. (eds.). (1989–1991). *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie* [sv. 1–3]. Berlin – New York: Walter de Gruyter

[5] ČERMÁK, F. – BLATNÁ, R. (eds.) (1995). *Manuál lexikografie*. Jinočany: H&H.

Část první, teoretická

1. Korpusová lingvistika v současném Lotyšsku, její možnosti a meze

Korpusová lingvistika se v současném Lotyšsku těší čím dál větší pozornosti a jazykovědci se na korpusové zdroje i nástroje odkazují ve stále větším množství prací. Základy této disciplíny se přitom dají najít již před počátkem užívání elektronických zdrojů a před samotnou digitalizací, v době nezaostávající nikterak za celosvětovým vývojem. Počátky světové korpusové lingvistiky sahají k 60. létům 20. století, přičemž se za tradiční mezník této disciplíny bere výstup *Computational Analysis of Present-Day American English* (Kučera a Francis, 1969)[6] z konce 60. let. Pod tímto korpusem je podepsán americký lingvista českého původu Henry Kučera společně s W. Nelsonem Francisem. Zmíněná publikace se opírala o zcela první korpus svého druhu, *The Brown Corpus* (plným názvem *The Brown University Standard Corpus of Present-Day American English)*[7]. Korpus obsahuje 500 úseků po 2000 slovech, které byly do systému zaneseny na základě analýzy textů zveřejněných do roku 1961. Pro větší tematickou pestrost v něm byla slovní zásoba vybírána z patnácti různých textových okruhů. Korpus dal vzniknout prvnímu počítačově zpracovanému frekvenčnímu slovníku *The American Heritage Dictionary of the English Language* (1969).

Těžiště korpusové lingvistiky se postupně ze Spojených států přesouvalo do Evropy, konkrétně do Velké Británie a dalších sdružených pracovišť, pod jejichž vedením vznikl např. tzv. LOB korpus (*Lancaster-Oslo-Bergen corpus*), který reflektuje britské zdroje, a je tedy ideálním kontrastním protějšek k Brownovu korpusu. Existence obou korpusů umožnila první kvantitativní analýzu britské a americké varianty anglického jazyka. Jedná se již o korpus tzv. druhé generace (korpus byl vybudován v letech 1970–1978). Mezi další korpusy druhé generace můžeme zařadit např. Frantext[8] nebo Corpus of Contemporary American English[9].

---

[6] KUČERA, Henry a W.Nelson FRANCIS, (1969). Computational Analysis of Present-Day American English. *International Journal of American Linguistics*. (35), 71-78.
Recenze korpusu dostupná zde: https://www.journals.uchicago.edu/doi/10.1086/465045

[7] SKETCH ENGINE. BROWN corpus: Corpus of American English. [online] Cit. 10.5. 2021 <www.sketchengine.eu/brown-corpus/>

[8] ATILF, 2021. Frantext, [online] Cit. 10.5. 2021. <https://www.frantext.fr>

[9] Corpus of Contemporary American English, (1990-2020). [online] Cit. 10.5. 2021. <https://www.english-corpora.org/coca/>

Dále jsou v průběhu 60. let započaty práce na mnoha dalších evropských korpusech, mezi něž patří např. DEREKO[10], korpus německého jazyka, který postupně vzniká již od roku 1964 a v současné době obsahuje 49 miliard jednotek (stav k r. 2020). Tento korpus je co do počtu jednotek aktuálně nejobsáhlejším korpusem na světě. Korpusy vzniklé po roce 2000 se následně považují za korpusy tzv. třetí generace. Jejich výhodou je, že již začínají čerpat také data z webu.

Přibližně ve stejnou dobu, tedy v průběhu šedesátých let minulého století, v Lotyšsku začínají vznikat publikace zohledňující kvantitativní aspekty jazyka, např. distribuci grafémů v jednotlivých ukázkách prózy (Lorenc, Nesaule 1963) nebo užití předložek v lotyšských časopisech (Freidenfelds 1967). Rozsahem i významem nejzásadnějším dílem tohoto období je nepochybně *Latviešu valodas biežuma vārdnīca (Frekvenční slovník lotyšského jazyka)*, který vznikal v průběhu let 1966–1976 v týmu pod vedením T. Jakubaite. O tomto díle bude podrobněji pojednáno v sekci Materiálová základna; již zde je však nutno poznamenat, že šlo o první lotyšské dílo svého druhu, přičemž svým vznikem položilo základ dalším publikacím. Na základě již zmíněného frekvenčního slovníku vzniká v roce 1985 další práce *Latviešu valodas pamata un tematisko vārdu krājums (Základní a tematická slovní zásoba lotyšského jazyka)* M. Soikāne-Trapāne, v níž je zaneseno okolo 6000 slov v 19 rozličných okruzích. V mnoha ohledech podobná práce vychází o šest let později v roce 1991 s názvem *1000 vārdu. Latviešu valodas leksikas minimums ar tulkojumiem krievu un angļu valodā (1000 slov. Lexikální minimum lotyšského jazyka s překlady do ruského a anglického jazyka)* (Bušs, Baldunčiks 1991), reflektující tisíc nejfrekventovanějších slov lotyšského jazyka společně s jejich překlady do ruštiny a angličtiny. Dalším dílem velmi obdobného obsahu, ovšem s trojnásobně širším heslářem, je poté *3000 latviešu sarunvalodas biežāk lietotie vārdi ar tulkojumu krievu, vācu un angļu valodā* (Kuzina 1998). Tato publikace bude rovněž blíže popsána v následující kapitole. Souběžně se od konce 80. let přistupuje k digitalizaci zdrojů: digitalizují se fragmenty Šmitsovy folkloristické antologie *Latviešu tautas ticējumi* a části Gliksova překladu Bible ze 17. století.

Poptávka po tvorbě tištěných frekvenčních slovníků však postupem času ustupovala přípravám digitalizovaného korpusu, shodně s trendy v ostatních zemích. Dnes je tvorba jazykových korpusů a nástrojů k jejich vytěžování spjatá s Institutem matematiky a informatiky

[10] Leibniz-Institut für Deutsche Sprache, (2021). DeReKo, [online] Cit. 10.5. 2021. <https://www1.ids-mannheim.de/kl/projekte/korpora.html>

Lotyšské univerzity, konkrétně se specializovaným pracovištěm Laboratoř umělé inteligence (*Mākslīgā intelekta laboratorija, AiLab*). Spolupráce Filologické fakulty Lotyšské univerzity a Lotyšské národní knihovny roku 2003 vyústila ve vznik Korpusu starých lotyšských textů, což je zároveň také první veřejně přístupný specializovaný korpus v Lotyšsku. Práce na všeobecném korpusu lotyšského jazyka začíná v roce 2005 vyhotovením koncepce (*Latviešu valodas korpusa koncepcija* – LU MII 2005), kde je nastíněna problematika jazykových korpusů obecně, přičemž největší akcent je kladen na možnosti a nutnost vytváření jazykových korpusů v lotyšském prostředí. Nepostradatelnou součástí této koncepce je také pečlivá analýza korpusů zahraničních. Pilotní verze korpusu vyšla pod názvem *Līdzsvarots mūsdienu latviešu valodas tekstu korpuss (Vyvážený korpus současných lotyšských textů)* a byla vyhotovena mezi lety 2007–2009, obsahovala 3,5 milionů jednotek.

Aktuálně nejnovějším jazykovým korpusem v Lotyšsku je korpus LVK2018, který bude opět podrobněji popsán později. Jde aktuálně o rozsahem největší všeobecný korpus lotyšského jazyka, obsahuje okolo 10 miliónů jednotek. Dlužno podotknout, že se nejedná o standard dnešní doby, korpusy o takovém objemu jsou typické spíše pro osmdesátá a devadesátá léta minulého století, novým etalonem co do velikosti se stal *British National Corpus* z roku 1994, o velikosti 100 milionů slov. Té by měl docílit i aktuálně chystaný LVK2022.

LVK2018 však není korpusem jediným, ani posledním, protože paralelně s ním probíhaly práce na dalších korpusech, např. syntaktický korpus lotyštiny LVTB[11] (*Latviešu valodas sintaktiski marķētais korpuss*) nebo korpus FullStack[12] k víceúrovňové anotaci dat (mimo jiné pomocí nástrojů Universal Dependencies, FramNet, PropBank, pojmenovávacích entit aj.), které však svým rozsahem nejsou s LVK2018 srovnatelné (řádově desetitisíců vs. miliony slov). V poslední dekádě vznikl také korpus mluvené lotyštiny LaRKo, i ten je relevantní pro účely této práce. Přestože na našem seznamu datasetů nefigurují, považujeme za vhodné stručně zmínit i jiné počiny lotyšské korpusové lingvistiky. Kromě lotyšských specifik, jako například korpus díla slavného básníka Rainise[13] (zveřejněn roku 2018) nebo korpus textů

[11] LINDAT, Digital Research Infrastructure for Language Technologies, Arts and Humanities, LVTB Latvian dependency constituency treebank. [online] Cit. 4.5. 2021. <https://lindat.mff.cuni.cz/services/pmltq/#!/treebank/lvtb25/query/>

[12] FullStack [online] Cit. 4.5. 2021. <https://github.com/LUMII-AILab/FullStack/pulls>

[13] SKETCH ENGINE. Raiņa darbu korpuss. [online] Cit. 4.5. 2021. <http://nosketch.korpuss.lv/#dashboard?corpname=rainis>

v latgalském nářečí MuLa (2013)[14] lze najít například korpus *Emuāri*[15], který mezi lety 2014–2015 zkoumal lexikum internetových blogů. Mezi lety 2002–2018 vznikal již zmíněný korpus starých textů (*Senie*)[16].

Právě probíhajícím projektem je například žákovský korpus LaVA (*Latviešu valodas apguvēju korpuss*)[17], který mapuje práce studentů lotyšského jazyka, pro něž lotyština není mateřským jazykem. Projekt byl spuštěn od roku 2018 a trvat má do roku 2021. Ve stejném časovém období je budován jiný žákovský korpus *Skolēnu pārspriedumu korpuss*[18], který mapuje práce žáků 12. třídy. Jedná se však vesměs o korpusy doplňkové a úzce specializované, přičemž největší důraz do budoucna je kladen na postupné rozšiřování již stávajících všeobecných jazykových korpusů řady LVK. Přínos výše zmíněných korpusů je nezanedbatelný zejména v pedagogickém prostředí, výuce nebo procesu učení jazyka.

2. Materiálová základna – popis zdrojů

Popisovaná materiálová základna se sestává ze dvou tištěných a dvou elektronických zdrojů. Cílem této práce je srovnání lexika v průřezu několika časových období, je tedy potřeba zohlednit i materiály tištěné, a to nejen pro dřívější absenci zdrojů elektronických. Konkrétně jde o tyto prameny: z tištěných čtyřdílný frekvenční slovník lotyšského jazyka (*Latviešu valodas biežuma vārdnīca*), který vznikal v letech 1966–1976, spolu s publikací *3000 latviešu sarunvalodas biežāk lietotie vārdi ar tulkojumu krievu, vācu un angļu valodā (3000 nejběžnějších slov z hovorové lotyštiny s překlady do ruského, německého a anglického jazyka)*; z elektronických korpusů LVK2018 a LSRC, posledně jmenovaný jakožto zástupce korpusů mluveného jazyka.

[14] SKETCH ENGINE. MuLa, Mūsdienu latgaliešu tekstu korpuss. [online] Cit. 4.5. 2021. <http://nosketch.korpuss.lv/#dashboard?corpname=mula>

[15] SKETCH ENGINE. Emuāri, Latviešu valodas emuāru korpuss. [online] Cit. 4.5. 2021.<http://nosketch.korpuss.lv/#dashboard?corpname=emuari>

[16] Senie. Latviešu valodas seno tekstu korpuss. [online] Cit. 4.5. 2021. <http://senie.korpuss.lv/toc.jsp>

[17] LaVa, 2018-2021. *Latviešu valodas apguvēju korpusa izveide:metodes, rīki un izmantojums*. [online] Cit. 6.5. 2021.: <http://lava.korpuss.lv>

[18] Skolēnu pārspriedumu korpuss (2021), [online] Cit. 10.5. 2021 <https://nosketch.korpuss.lv/#dashboard?corpname=vidusskolu_diktati >

Naopak upouštíme od původního záměru do analýzy materiálů zapojit i korpus mluvené lotyštiny LaRKo, a to kvůli malému objemu dat. Byl proto nahrazen korpusem LSRC, který se pro potřeby této práce svým rozsahem jevil jako adekvátnější. V této sekci nicméně o korpusu LaRKo stručně poreferujeme, v kontextu současné lotyšské korpusové lingvistiky jde totiž o významný počin a jeho využití je samozřejmě možné i nad rámec této práce. Podrobněji bude popsán také korpus LSRC.

2.1. Jakubaite et al. 1973

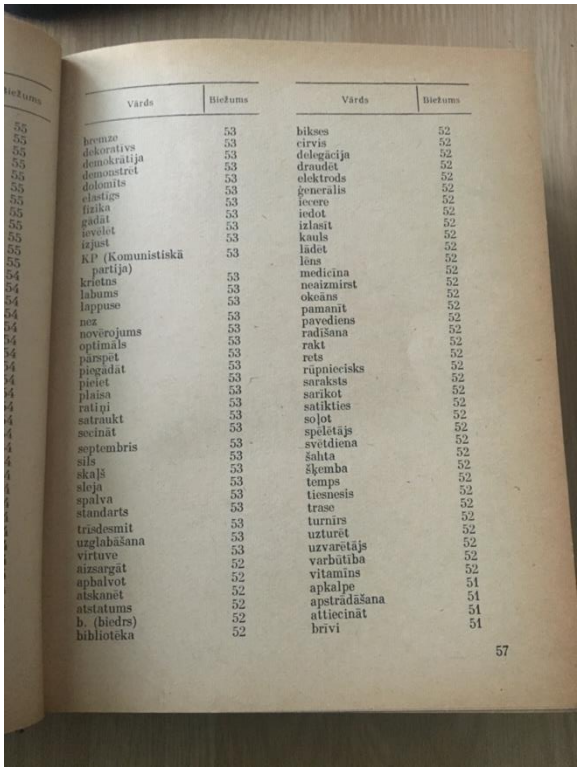
Prvním zkoumaným tištěným materiálem je dílo *Latviešu valodas biežuma vārdnīca*, tedy *Frekvenční slovník lotyšského jazyka*. Na jeho vzniku se podílel šestičlenný tým lotyšských jazykovědkyň pod vedením T. Jakubaite. Přestože se může zdát, že z dnešního pohledu je tento zdroj již dávno překonán a jeho význam je pro dnešní lexikografii mizivý, jedná se o vůbec první lotyšský frekvenční slovník. Toto čtyřsvazkové dílo je pionýrským počinem a představuje první ucelený výsledek kvantitativního přístupu k jazyku v Lotyšsku. Dobře odráží dřevní doby dané disciplíny: zatímco první a druhý díl frekvenčního slovníku byl zpracován čistě ručně, při sestavování třetího svazku bylo použito již elektronické zpracování pomocí samočinného počítače ESM Minsk23.

Výrazným rysem slovníku je jeho praktičnost především v tematickém rozdělení jednotlivých svazků. V prvním svazku nacházíme podrobnější popis zpracování dobového lotyšského lexikonu jakožto celku, další svazky se zaměřují na konkrétní žánrové okruhy. Lexikum bylo vybráno na základě analýzy celkového počtu 892 ukázek (1. svazek – 292, 2. svazek – 300, 3. svazek – 300) po tisícovce slov o velikosti 31 039 lemmat; celková velikost zdrojových dat tak činila 892 000 textových slov.[19]

Každý z uvedených svazků obsahuje úvod, kde je teoreticky i za grafického doprovodu znázorněna metodologie užitá při jeho tvorbě. Každý z daných svazků obsahuje kromě seznamu nejfrekventovanějších slov také seznam abecedně řazených slov s příslušným údajem o počtu výskytů a jeho ranku.

[19] JAKUBAITE et al. (1973): *Latviešu valodas biežuma vārdnīca. [Apvienotais (1.–3.) sējums]*. Rīga: Zinātne, s. 5-6.

V přibližně stejném období, konkrétně v roce 1961, v českém prostředí vychází publikace *Frekvence slov, slovních druhů a tvarů v českém jazyce* (Jelínek, Bečka, Těšitelová 1961)[20]. Hlavní rozdílem obou děl je povaha analyzovaných textů a zpracování získaných dat. Česká publikace vznikala s přestávkami od roku 1930, množství shromážděných textů a jednotek je větší (srov. 1 623 527 tokenů vs. 892 000 tokenů v LVBV[21]) a zahrnuje širší spektrum textů: osm tematických okruhů v 75 ukázkách. Jmenovitě se jednalo o tyto okruhy: beletrie (A), poezie (B), literatura pro mládež (C), dramata (D), odborná literatura (E), žurnalistika (F), vědecká literatura (G), mluvené projevy uveřejněné tiskem (H). Lotyšská publikace oproti tomu zpracovávala omezenější počet okruhů: technika a průmysl, krásná literatura, věda.



Vārds	Biežums	Vārds	Biežums
bronza	53	bikses	52
dekoratīvs	53	cirvis	52
demokrātija	53	delegācija	52
demonstrēt	53	draudēt	52
dolomīts	53	elektrods	52
elastīgs	53	ģenerālis	52
fizika	53	iecere	52
gadāt	53	iedot	52
ievērot	53	izlasīt	52
izjust	53	kauls	52
KP (Komunistiskā partija)	53	lādēt	52
krietns	53	lēns	52
labums	53	medicīna	52
lappuse	53	neaizmirst	52
nez	53	okeāns	52
noverojums	53	pamanīt	52
optimāls	53	pavedziņa	52
pārspēt	53	radīšana	52
piegādāt	53	rakt	52
pieiet	53	rets	52
plaisa	53	rūpniecisks	52
ratiņi	53	saraksts	52
satraukt	53	sarikot	52
secināt	53	satikties	52
septembris	53	sojot	52
sils	53	spēlētājs	52
skalš	53	svētdiena	52
sleja	53	šahta	52
spalva	53	šķemba	52
standarts	53	temps	52
trisdesmit	53	tiesnesis	52
uzglabāšana	53	trase	52
virtuve	52	turnīrs	52
aizsargāt	52	uzturēt	52
apbalvot	52	uzvarētājs	52
atskanēt	52	varbūtība	52
atstatums	52	vitamīns	52
b. (biedrs)	52	apkalpe	51
bibliotēka	52	apstrādāšana	51
		attiecināt	51
		brīvi	51

57

Obrázek 1. Ukázka z Jakubaite et al. 1973

[20]JELÍNEK, BEČKA, TĚŠITELOVÁ, *Frekvence slov, slovních druhů a tvarů v českém jazyce* [online] Cit. 4.5. 2021. < http://www.ujc.cas.cz/miranda2/export/sitesavcr/data.avcr.cz/humansci/ujc/infoz/Frekvence-slov-1961/Frekvence-slov-1961.pdf >

[21] JAKUBAITE et al. (1973): *Latviešu valodas biežuma vārdnīca. [Apvienotais (1.–3.) sējums]*. Rīga: Zinātne, s. 5-6.

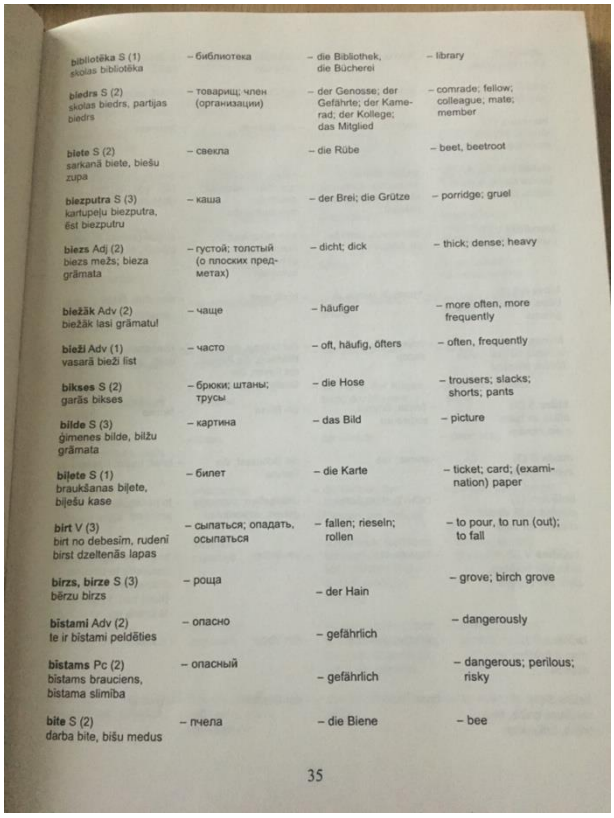
2.2. Kuzina 1998

Druhým tištěným zdrojem, o který se opírá naše analýza dat, je dílo *3000 latviešu sarunvalodas biežāk lietotie vārdi ar tulkojumu krievu, vācu un angļu valodā (3000 nejběžnějších hovorových slov lotyšského jazyka s překlady do ruského, německého a anglického jazyka).* Jedná se o dílo jazykovědkyně a pedagožky V. Kuziny z roku 1998. Již v úvodu je akcentována didaktická úloha díla: potenciálnímu uživateli je příručka doporučována pro efektivnější a systematičtější osvojování slovní zásoby. V předmluvě se můžeme dočíst, že na základě výzkumů pokrývá 100 nejběžněji užívaných slov až 70 % slovní zásoby v tištěných zdrojích, tři tisíce pak údajně pokryjí 90 % (viz dále kap. 4). Jak je již z názvu patrné, dílo obsahuje tři tisíce slov nejběžněji užívaných v mluvené lotyštině spolu s jejich ekvivalenty ve třech cizích jazycích (ruština, němčina, angličtina), což dále zvyšuje praktičnost příručky.

Slovní zásoba byla vybrána na základě rozšířenosti daného slova v textech shromážděných z poměrně širokého období, let 1976–1997. Z uvedené analýzy zdrojů byl vytvořen korpus o velikosti cca 300 000 jednotek, počet lemmat dosahoval 15 875. Následně byl vygenerován soupis 3288 nejfrekventovanějších slov, přestože samotný název knihy avizuje nepatrně nižší lemmat, rovných 3000.[22]

Řazení lexika je abecední. Uspořádání publikace je následovné: v prvním sloupci je uvedeno slovo v lotyšském jazyce, pod ním se objevuje kolokace, případně jiný frekventovaný způsob užití dané jednotky. Vedle každého slova je rovněž uvedeno, do jaké frekvenční úrovně patří (k 1., 2. či 3. tisícovce). Vše doplňuje překlad do ruského, německého a anglického jazyka ve zbývajících sloupcích.

[22] Tuto skutečnost, která nepříjemně komplikuje srovnání jednotlivých datasetů, jsme zjistili díky digitalizaci díla. Slov patřících do první tisícovky je 990, do druhé 1046, ve třetí jich figuruje 1252; celkově se tedy jedná o 3288 lexémů. O důvodech těchto disproporcí lze jen spekulovat – jednotlivé tisícovky hesel lze přesně odpočítat i ručně, natožpak s pomocí počítače.



bibliotēka S (1) skolas bibliotēka	– библиотека	– die Bibliothek, die Bücherei	– library
biedrs S (2) skolas biedrs, partijas biedrs	– товарищ; член (организации)	– der Genosse; der Gefährte; der Kamerad; der Kollege; das Mitglied	– comrade; fellow; colleague; mate; member
biete S (2) sarkanā biete, biešu zupa	– свекла	– die Rübe	– beet, beetroot
biezputra S (3) kartupeļu biezputra, ēst biezputru	– каша	– der Brei; die Grütze	– porridge; gruel
biezs Adj (2) biezs mežs; bieza grāmata	– густой; толстый (о плоских предметах)	– dicht; dick	– thick; dense; heavy
biežāk Adv (2) biežāk lasi grāmatu!	– чаще	– häufiger	– more often, more frequently
bieži Adv (1) vasarā bieži līst	– часто	– oft, häufig, öfters	– often, frequently
bikses S (2) garās bikses	– брюки; штаны; трусы	– die Hose	– trousers; slacks; shorts; pants
bilde S (3) ģimenes bilde, bilžu grāmata	– картина	– das Bild	– picture
biļete S (1) braukšanas biļete, biļešu kase	– билет	– die Karte	– ticket; card; (examination) paper
birt V (3) birt no debesīm, rudenī birst dzeltenās lapas	– сыпаться; опадать, осыпаться	– fallen; rieseln; rollen	– to pour, to run (out); to fall
birzs, birze S (3) bērzu birzs	– роща	– der Hain	– grove; birch grove
bīstami Adv (2) te ir bīstami peldēties	– опасно	– gefährlich	– dangerously
bīstams Pc (2) bīstams brauciens, bīstama slimība	– опасный	– gefährlich	– dangerous; perilous; risky
bite S (2) darba bite, bišu medus	– пчела	– die Biene	– bee

35

Obrázek 2. Ukázka z Kuzina 1998.

2.3. LVK2018[23]

Jak bylo nastíněno již v úvodu, kromě tištěných materiálů se tato práce zabývá popisem také zdrojů elektronických. Korpus LVK2018 je aktuálně největším reprezentativním korpusem lotyšského jazyka. Práce na vývoji jazykových korpusů v Lotyšsku byly započaty teprve na počátku tohoto století, přičemž impuls k jejich tvorbě byl vydán Lotyšskou jazykovou agenturou (*Latviešu valodas aģentura*), která plán k vypracování zveřejnila již v roce 2005; na jeho základě vznikla i předchozí verze korpusu (LVK2013). Při návrhu koncepce byly zohledněny různé faktory jako např. povaha textů, které měly být do korpusu zařazeny, jejich tematická rozmanitost nebo proporce jazykových zdrojů. Jak uvádí K. Levāne-Petrova (2012)[24],

[23] SKETCH ENGINE. LVK2018. [online] Cit. 6.5. 2021. <http://nosketch.korpuss.lv/#dashboard?corpname=LVK2018>

[24] LEVĀNE-PETROVA, Kristīne, (2012). Līdzsvarots mūsdienu latviešu valodas tekstu korpuss un tā tekstu atlases kritēriji. *Baltistica*. VIII, 89-98
[online] Cit. 4.5. 2021. < http://www.baltistica.lt/index.php/baltistica/article/view/2113 >

která se na koncepci podílela, byly při přípravách zohledněny zkušenosti ze zpracování cizích korpusů, jmenovitě britského, litevského či českého.

Veškerá dostupná data z korpusu LVK2013 byla aktualizována a plně použita v LVK2018. Mírnou změnou oproti původní verzi je například typ okruhů, ze kterých byly texty čerpány. Verze LVK2018 má pět kategorií: beletrie, periodika, vědecké texty, regulační předpisy a stenografické přepisy ze Saeimy (lotyšský parlament). Do budoucna se počítá s případným omezením počtu sekcí, přičemž by zůstala zachována sekce beletrie, periodika a sekce nebeletristických textů, kam budou zařazeny texty, které nemohly být zařazeny ani do jedné z předchozích kategorií.[25] V celém korpusu se objevují pouze texty, které byly publikovány po roce 1991, zároveň se muselo jednat o texty původní, tedy nepřekladové. Texty rovněž nesměly obsahovat tabulky nebo matematické symboly. Zvláštní důraz byl poté kladen jak na tematickou pestrost vybíraných ukázek, ať už se jednalo o texty z periodik (kdy nemělo docházet k ukládání textů se shodným tématem, jen z jiného zdroje), tak na pestrost témat v jiných okruzích (například pestrost témat prezentovaných v Saeimě nebo různorodost mluvčích). Jedním ze společných kritérií pro zajištění rozmanitosti textů je to, že jedna textová položka by neměla přesáhnout 5 % části korpusu.

Pro další zajištění kvality textů byly také stanoveny horní a dolní limity pro počet jednotek u všech položek, přičemž jako položky byla hodnocena slova, čísla, interpunkční znaménka a další symboly. Tyto limity byly stanoveny pro každou z pěti zkoumaných sekcí. Například pro sekci periodika byl limit stanoven počet užitých textů min. 30, max. 2500, pro sekci beletrie max. 55 000, regulační předpisy max. 22 000, Saeima min. 30. Překročení horní hranice tohoto limitu nebylo překážkou k zařazení dokumentu, byl však zařazen jen jeho fragment. Je na místě poznamenat, že stanovená velikost korpusu (10 miliónů jednotek) umožnila sice zahrnutí i delších textů, pokud však nejdůležitějším kritériem byla tematická rozmanitost, bylo adekvátní vybírat raději větší počet textů kratší povahy.

Vědecké texty byly vybírány z disertačních prací dostupných na internetu nebo z jejich abstraktů, kvůli kýžené rozmanitosti z různých oborů, např. fyziky, lingvistiky, geografie atd. V případě regulačních předpisů byl zachován pouze základní text zákona. Kromě toho byla shromažďována metadata: datum přijetí zákona, datum vstupu v platnost a číslo oficiální

[25] LEVĀNE-PETROVA, Kristīne (2012) Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos [online] Cit. 10.5. 2021. <https://www.apgads.lu.lv/fileadmin/user_upload/lu_portal/apgads/PDF/Valoda-nozime-forma/VNF-10/vnf_10-12_Levane_Petrova.pdf >

publikace. V sekci Saeima byli řečníci rozděleni do šesti kategorií, přičemž jako vhodné přepisy k zařazení do LVK2018 byly vyhodnoceny pouze projevy poslanců a projevy standardní.

Další změna oproti původní verzi je například zpracování doprovodných metadat. Některé z hodnoty byly viditelné pouze pro interní užití. V této verzi měla být metadata sjednocena, některá byla úplně vyškrtnuta (jako např. pohlaví autora).

LVK2018 nabízí nejen různé možnosti vyhledávání, ale též prohlížení a třídění dat. Je například možné upravit délku požadovaného kontextu, počet příkladů na stránce a rozsah pravého a levého okolí. Lze také vytvořit seznamy frekvencí pro nalezená data a následně data uložit.

Problematická může být morfologická anotace. Vzhledem k vysoké míře flektivnosti lotyštiny se v korpusových datech často vyskytuje tvarová homonymie, s jejíž desambiguací mohou mít automatické anotační nástroje – lemmatizátory a taggery – občas potíže.

2.4. LaRKo[26]

Jak bylo avizováno v úvodu této kapitoly, korpus LaRKo nakonec do analýzy zařazen nebyl, jeho malý objem neumožňoval nezkreslené určení frekvenční špičky mluvené lotyštiny. Korpus byl vybudován v Institutu matematiky a informatiky Lotyšské univerzity v roce 2014. K jeho vytvoření posloužily ukázky z rozličných televizních a rozhlasových kanálů a také přepisy projevů poslanců v Saeimě. Korpus obsahuje nahrávky od 300 mluvčích a celkově pokrývá 8 hodin[27]. Ke každému zvukovému souboru jsou připojena metadata: podrobnosti o místě, kde byl záznam pořízen (např. ve studiu, ve studiu se šumem v pozadí, mimo studio v prostorech bez šumu v pozadí, mimo studio s hlukem pozadí, v automobilu, na ulici), či trvání zvukového fragmentu; dále zda se jedná o předem připravený, spontánní nebo čtený text a také je postižen rozdíl mezi projevem veřejným a soukromým. Nechybí ani popis jednotlivých mluvčích: např. jejich pohlaví, věk, jazykové zázemí (zda mluvčí hovořil ve svém rodném jazyce, zda jeho výpověď‘ obsahovala prvky dialektu, nebo zda je mluvčí bilingvní). Na

[26] Latvijas Universitātes Matemātikas un informātikas institūts. Latviešu valodas runas korpusā, (2014). [online] Cit. 6.5. 2021. <http://larko.ailab.lv/index.php/info>

[27] Porovnejme LaRKo s českými korpusy mluveného jazyka. Například korpus ORATOR ve své první verzi obsahoval přes 72 hodin nahrávek, jeho druhá verze pak obsahovala více než dvojnásobek původní délky, tj. 148 hodin. (https://wiki.korpus.cz/doku.php/cnk:orator). Ortofon obsahoval přes 102 hodin nahrávek (https://wiki.korpus.cz/doku.php/cnk:ortofon). Nejobsáhlejším korpusem je poté korpus ORAL, který ve verzi ORAL2013 a ORAL-Z dosahuje délky přes 354 hodin. (https://wiki.korpus.cz/doku.php/cnk:oral)

základě těchto parametrů je v korpusu rovněž možno vyhledávat. Tato skutečnost nemá zanedbatelný význam vzhledem k sociolingvistické situaci v současném Lotyšsku. Na základě dat z roku 2015[28] se v Lotyšsku 61,6 % obyvatelstva označuje za etnické Lotyše. S tím samozřejmě úzce souvisí jazyková otázka. V sociolingvistickém průzkumu Lotyšské jazykové agentury z roku 2014 (*Valodas situācijas sociolingvistiskā izpēte*)[29] se uvádí, že k roku 2014 lotyštinu ovládalo 90 % představitelů menšin[30]. Obdobné výsledky zaznamenalo i šetření z roku 2009. Pro srovnání, v roce 1989 to bylo pouze 23 %, v roce 2000 již 53 %[31]. Předmětem této práce není popis jazykových znalostí lotyšského obyvatelstva, relevantní je však na specifickou sociolingvistickou situaci a pravděpodobné charakteristiky v mluvě této skupiny mluvčích poukázat. Promluva takového mluvčího může kromě fonetických zvláštností být charakteristická například výběrem použitého lexika, což by se následně na cíli této práce mohlo projevit.

K nahlédnutí je uživateli také statistický přehled pořízených záznamů. Největší část pochází z úst mluvčích ve věkové kategorii 26–50 let (68,2 %). Co se týče jazykové stránky, největší počet nahrávek neobsahuje jazykové odchylky a je nahrán rodilým mluvčím jazyka (83,4 %). Co se týče spontánnosti, respektive připravenosti projevu, nepozorujeme větší rozdíly, procento spontánních projevů činí 47,4 % oproti připraveným (20,9 %) a čteným (31,7 %). Dostupná data se rovněž vztahují na délku promluvy a její kvalitu[32]. Každý použitý zvukový záznam byl rozdělen na fragmenty k přepsání vhodné a nevhodné. Za fragmenty vhodné k transkripci jsou považovány ty, které obsahují ničím nepřerušený projev mluvčí. Jako nevhodné jsou poté hodnoceny ty, kde je řeč přerušena či určitým způsobem rušena, například paralelní projev jiného mluvčího aj. Fragmenty, které byly zhodnoceny jako adekvátní, byly následně transkribovány. Průměrná délka takové výpovědi je 2–3 sekundy. Fráze jsou

[28] VALODAS SITUĀCIJA LATVIJĀ 2010–2015. [online] Cit. 7.5. 2021. <https://valoda.lv/wp-content/uploads/aktual/Val_sit_informat_lapa_3.pdf>

[29] *Valodas situācijas sociolingvistiskā izpēte.* Kvantitatīvā pētījuma rezultātu ziņojums. Pasūtītājs: LVA. Īstenotājs: SIA „Excolo Latvia", (2014)[online] Cit. 7.5. 2021. <https://valoda.lv/wp-content/uploads/docs/Petijumi/Sociolingvistika/VSL_2015_web.pdf>

[30] *Valodas situācijas sociolingvistiskā izpēte*. Kvantitatīvā pētījuma rezultātu ziņojums. Pasūtītājs: LVA. Īstenotājs: SIA „Aptauju aģentūra", (2012) [online] Cit. 7.5. 2021.<https://valoda.lv/wp-content/uploads/docs/Petijumi/Sociolingvistika/VSL_2015_web.pdf>

[31] Tamtéž

[32] Accessible Speech Recognition, Signal to Noise Ratio Estimation; [online] Cit. 4.5. 2021.<https://www.isip.piconepress.com/projects/speech/software/legacy/signal_to_noise/index.html >

rozděleny podle intonační struktury řeči. Nádechy, výdechy a pauzy přesahující délku 0,3 sekundy a pauzy, které oddělují jednu intonační jednotku od druhé, jsou ve zvukovém souboru odděleny jako samostatné fragmenty. Zvukové materiály byly přepsány v ortografické transkripci, což je doslovná mluvená reprezentace v písemné formě, při níž jsou dodrženy všechny zásady jazykového přepisu. V takovém přepisu je zapsáno vše, včetně čísel a zkratek, navíc jsou podchyceny i neverbální prvky, pauzy a cizojazyčné prvky. Rovněž jsou vyznačeny odchylky od norem lotyšské ortoepie a pravopisu.

Zde se zároveň nabízí možnost tento korpus stručně porovnat s českým protějškem. Korpus ORATOR nabízí uživateli přehled monologických projevů, které byly pořízeny u různých příležitostí. Podmínkou však bylo, aby byl takový projev přednesen rodilým mluvčím češtiny a aby mluvčí na svůj projev byl předem připraven. Původním kritériem ovšem bylo, aby takový projev nebyl pouze čten. Při shromažďování dat ovšem opakovaně docházelo k situacím, kde byl takový projev čten alespoň částečně. Při postupném zpracování se od tohoto kritéria pomalu upouštělo, a v korpusu jsou tedy zaneseny i projevy čtené nebo částečně čtené[33]. Mluvčí měl přesně na pronesení svého projevu přesně vymezený prostor. Možnou nevýhodou tohoto kritéria je následná formálnost takového projevu. ORATOR se skládá z 318 nahrávek, které byly pořízeny mezi lety 2005–2019 od 332 mluvčích. Délka jednotlivých ukázek je individuální, záleží na typu promluvy. Dalším mluveným korpusem je např. ORTOFON, korpus neformální mluvené češtiny s víceúrovňovým přepisem. Shromažďovány byly spontánní projevy mluvčích, kteří se již znají. V rámci sběru dat byly zachycovány také některé informace o mluvčím, jako např. pohlaví, věk, nejvyšší dosažené vzdělání a nářeční oblast, v níž mluvčí strávil většinu života do svých 15 let.[34]

[33] KOPŘIVOVÁ, M. – LAUBEOVÁ, Z. – LUKEŠ, D. – POUKAROVÁ, P.: *ORATOR: Korpus monologů*. Ústav Českého národního korpusu FF UK, Praha (2019) [online] Cit. 4.5. 2021.<https://www.korpus.cz>

[34] KOPŘIVOVÁ, M. – KOMRSKOVÁ, Z. – LUKEŠ, D. – POUKAROVÁ, P. – ŠKARPOVÁ, M.: *ORTOFON: Korpus neformální mluvené češtiny s víceúrovňovým přepisem*. Ústav Českého národního korpusu FF UK, Praha 2017. [online] Cit. 4.5. 2021.< http://www.korpus.cz>

2.5. Latvian Speech Recognition Corpus[35]

V úvodní části této kapitoly jsme uvedli, že místo dat z korpusu LaRKo byla v práci využita data z korpusu LSRC. Ten byl vytvořen pro účely rozpoznávání řeči a její následné syntézy (Pinnis, Auziņa a Goba 2014)[36]. Data v něm jsou anotována dvojím způsobem: většina (o celkové délce 100 hodin) – ortograficky, malá část (celkem 4 hodiny) foneticky. Soubory metadat ve formátu XML poskytují dodatečné informace o úrovních hluku, stylech řeči atd. Korpus LSRC je foneticky vyvážený a bohatý: zahrnuje promluvy celkem 1851 mluvčích (1016 mužů a 835 žen) a celková délka činí 6001 minut. Přinejmenším 60 % dat pochází od mluvčích bez specifického dialektu či přízvuku, zastoupeni jsou i mluvčí hovořící různými dialekty (až 15 %) nebo akcentem (až 25 %; např. běloruský, anglický, ruský, ukrajinský). Korpus se skládá z připravených promluv (40 %) a spontánní řeči (60 %). Připravené projevy zahrnují televizní a rozhlasové zprávy, audioknihy, veřejně čtené projevy, čtené prezentace atd. Do spontánních projevů spadají televizní a rozhlasové diskuse, rozhovory, nahrané rozhovory, projevy podle připraveného plánu (ne však čtené projevy), např. prezentace, přednášky apod. Korpus pokrývá různé styly řeči, a to ze dvou důvodů: 1) kvůli větší rozmanitosti řečových dat, pokud jde o pokrytí doplňujících slov, rozdíly v rychlosti vyslovovaných slov, větší rozmanitost intonací, a 2) kvůli pozdější možnosti vylepšení nástrojů pro rozpoznávání řeči pro konkrétní úkoly (např. přepis diktátu, přepis vysílaných zpráv, přepis přednášek atd.) Oficiálně udávané statistiky (Pinnis, Auziņa a Goba 2014: 1551)[37] uvádějí velikost cca 837 tisíc tokenů a cca 72,5 tisíc unikátních slov (typů); dataset, jenž jsme dostali k dispozici a s nímž pracujeme v této práci, má nicméně jinou kvantitativní charakteristiku: přes 847 tisíc tokenů a více než 31,3 tisíc různých lemmat.

[35] Statistika. [online] Cit. 10.5. 2021. <http://runa.korpuss.lv>

[36] PINNIS, Mārcis – AUZIŅA, Ilze – Goba, Kārlis (2014). Designing the Latvian Speech Recognition Corpus - presentation.

[37] tamtéž

Část druhá, praktická

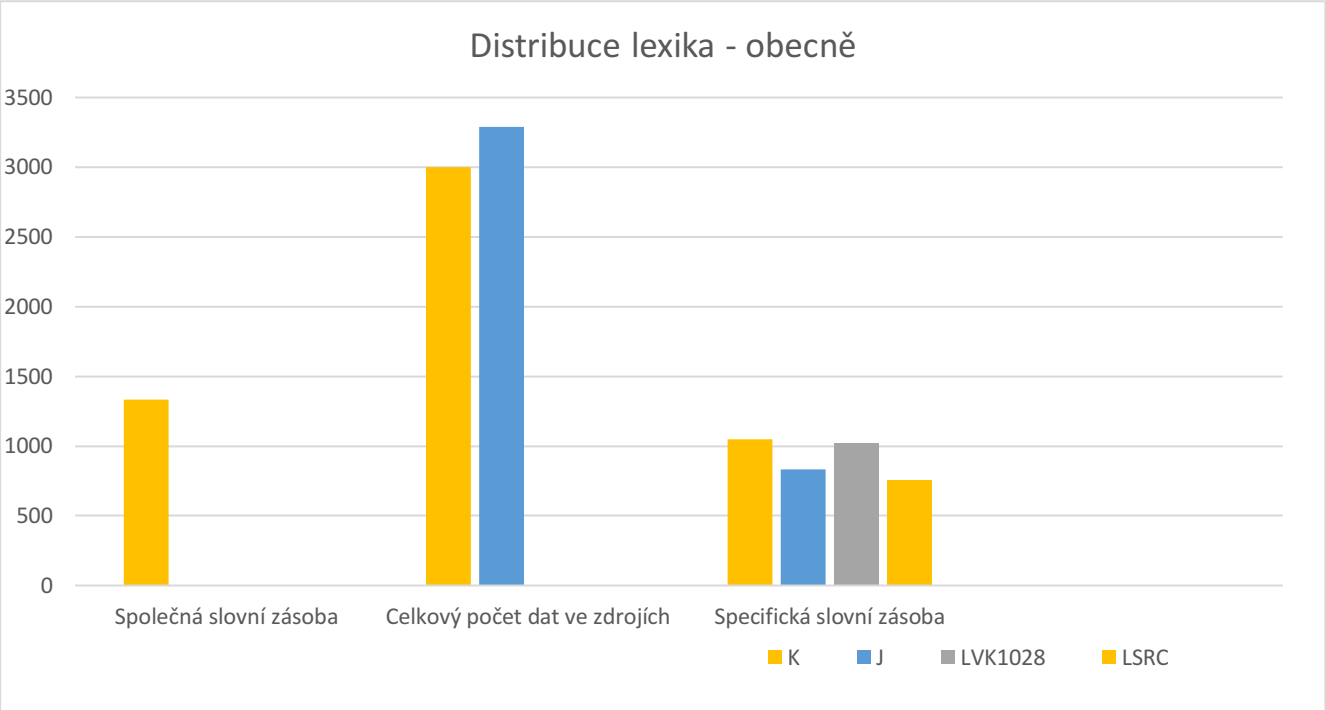
3. Srovnání frekvenčních špiček z použitých datasetů

3.1. Společné jádro

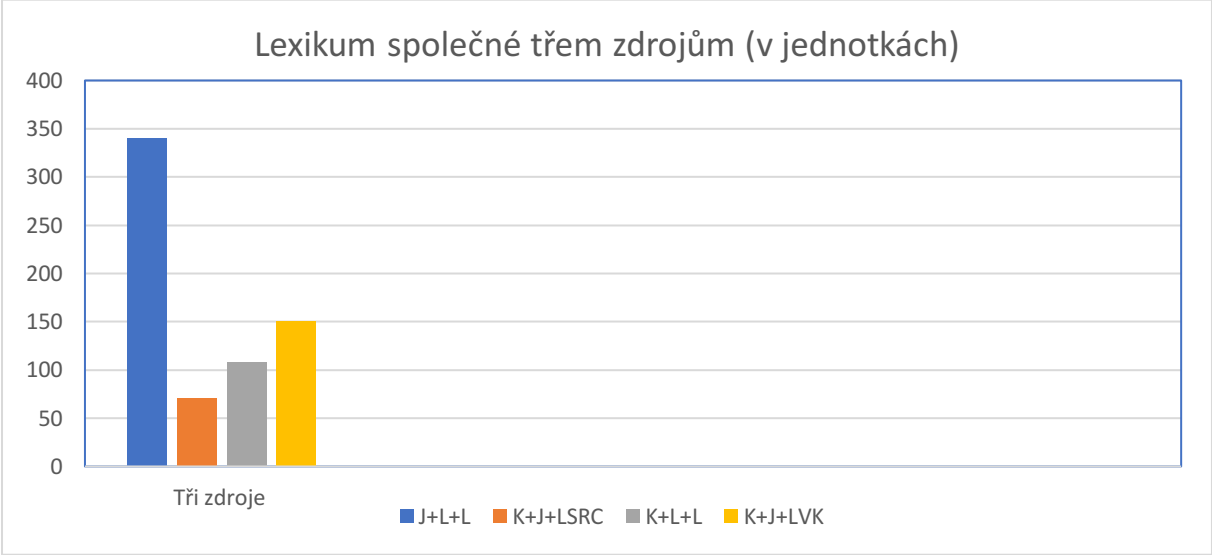
Nejdůležitějším sledovaným údajem je přítomnost lexikální jednotky v průniku frekvenčních špiček, představovaných našimi výchozími datasety. Tímto průnikem lze vymezit stabilní lexikální jádro lotyštiny za posledních nejméně 70 let. Frekvenční špička reflektuje jednotky s největším výskytem ve zkoumaném materiálu, a to doložené napříč ve všech zkoumaných materiálech.

Do společné frekvenční špičky se dostalo přesně 1337 **j**ednotek; tyto jednotky se objevovaly ve všech zkoumaných zdrojích. To je důležité především proto, že dokazuje aktuálnost těchto jednotek v různých zkoumaných obdobích. Připomeňme, že data nejstaršího zkoumaného zdroje byla sbírána v průběhu 50.–60. let; ta nejnovější pocházejí z počátku 21. století. Skutečnost, že se daná jednotka objevuje ve frekvenční špičce zohledňující různá období, podtrhuje dlouhodobou univerzálnost a aktuálnost dané jednotky a její živost v reálném úzu. Důležitý je tu aspekt diachronní, bez něj by výše popsaná vlastnost těchto jednotek nebyla názorně doložena.

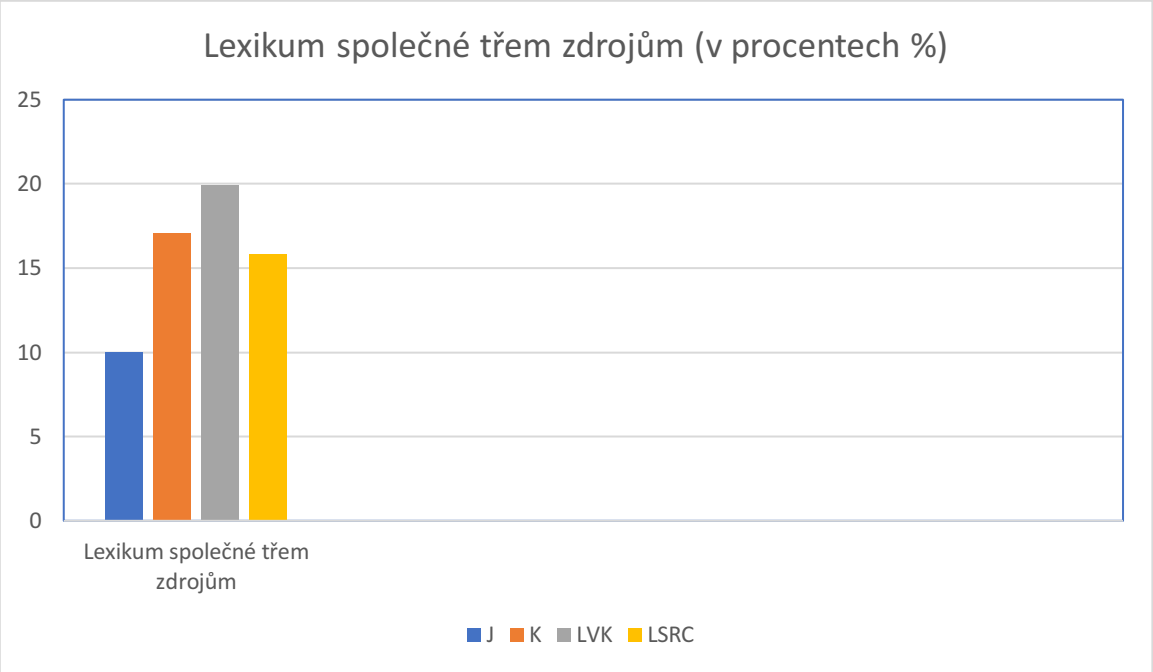
Rozložení je následovné: u K počet společných jednotek pro všechny zdroje dosáhl přesně 40,8 %. Kromě lexika společného pro všechny zdroje jsme rovněž chtěli zjistit počet (respektive procento) lexémů, které se objeví jen v určité kombinaci zdrojů, tři zdroje. Jako první uvádíme přehled slov společných právě třem zdrojům. Čísla vyšla následovně (viz graf č.1: J+LVK+LSRC = 340 slov, K+J+L = 71 slov, K+L+L = 108 slov, J+L+L = 339 slov a u K+L+L = 108 slov. To znamená, že slovní zásoba společná právě třem zdrojům se v K vyskytuje z 10 %, u J ze 17,1 %, u LVK z 19,9 % a u LSRC z 15,8 %.



Graf č. 1 Distribuce lexika ve zkoumaných datasetech

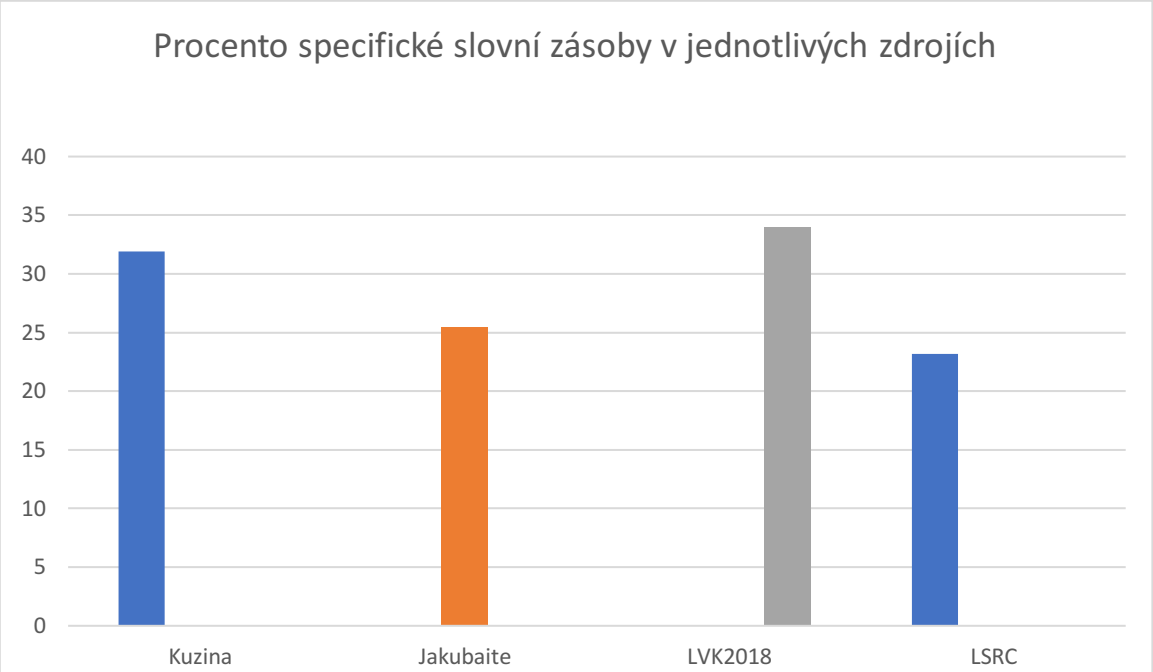


Graf č. 2 Lexikum společné třem zdrojům (v počtech lemmat)



Graf č. 3 Lexikum společné třem zdrojům (v procentech)

Po započtení těchto procent s procenty zohledňujícími společnou slovní zásobu docházíme ke zjištění, kolik procent unikátních jednotek každý zdroj obsahoval. Tyto jednotky jsou unikátní každému zdroji, protože se v žádném dalším zkoumaném zdroji již neobjevily. Slovník K obsahoval právě 31,9 % jednotek vlastních pouze danému zdroji. U J se objevilo přesně 25,4 % vlastní slovní zásoby. Korpus LVK2018 obsahoval přesně 34 % vlastní slovní zásoby. Korpus mluveného jazyka LSRC 23,2 %. Vše je pro větší přehlednost uvedeno v následující tabulce.



Graf č. 4 Procentuální rozložení specifické slovní zásoby pro jeden zdroj (v procentech)

Některé jednotky objevující se v rámci frekvenčních špiček mohou být pro laického nepoučeného uživatele matoucí. Nejčastěji se jedná o to, jak vytyčit hranice samostatné lexikální jednotky a kdy se stále ještě jedná o identické slovo v nepatrně odlišné formě. Uveďme příklad dvou velmi podobně znějících jednotek, nicméně s odlišnými gramatickými funkcemi: *augša* × *augšā*. Zatímco první je substantivum (‚vršek, hořejšek'), druhé plní funkci adverbia (‚nahoře'), byť z diachronního hlediska se jedná o ustrnulou pádovou formu (lokativu) daného substantiva. Důležité je proto mít na paměti, že přestože se rozdíl mezi těmito dvěma slovními jednotkami může zdát zanedbatelný, je potřeba ho v analýze dat řádně odlišovat, protože lotyšskému mluvčímu se tyto dvě jednotky jeví jako samostatné. Pro větší autenticitu zjištěné společné slovní zásoby tyto jednotky byly považovány za dva odlišné příklady, a byly proto započítávány zvlášť. Podobně jsme naložili také s adjektivy a adverbii. Byť je rozdíl mezi těmito slovy v lotyštině minimální (např. *augsts* × *augsti, atsevišķs* × *atsevišķi),* jedná se o jiný slovní druh. Dále bylo zajímavé zjistit, zda mají adjektiva a adverbia tendenci se vyskytovat ve všech zdrojích ve stejné míře. Nejdříve jsme zjišťovali, jak to vypadá napříč všemi zdroji. Úplná shoda, tedy výskyt párové dvojice adjektiva a adverbia (např. *grūti+grūts/slikti+slikts*) ve všech zdrojích, byl zaznamenán 21krát. V osmi případech se vyskytovalo pouze adverbium (*pēkšņi*, *patīkami*) bez odpovídajícího adjektiva. Nejčastějším případem ovšem byl výskyt samotného adjektiva; to se bez odpovídajícího adverbia objevilo padesátkrát. Tentýž jev jsme také

srovnávali pouze ve dvou mluvených zdrojích, tj. K a LSRC. Samotný počet společného lexika právě těchto dvou zdrojů není nikterak velký (jde o pouhých 80 jednotek). Oproti předchozímu případu, kdy se nejčastěji objevovala adjektiva bez svých párových adverbií, v tomto zkoumaném okruhu se nejčastěji objevovala adverbia bez odpovídajících adjektiv (celkem 8, např. *laipni*, *neapšaubāmi*, *savādi*). Objevily se pouze dva případy samostatných adjektiv (*prāvs*, *sportisks*). Obě varianty, tedy adjektivum a odpovídající adverbium, se objevily rovněž ve dvou případech (*godīgi*+*godīgs, kārtīgi*+*kārtīgs*).

3.2. Specifika psané vs. mluvené lexikální špičky, potenciální vývojové tendence

Jedním z cílů této práce je vypozorovat vývojové tendence slovní zásoby lotyštiny. Jak se dalo předpokládat, velká část zkoumaných jednotek spisovného jazyka zůstala i přes velké časové rozpětí stále stejná, není totiž možné, aby se jazyk změnil natolik, že bychom v rámci časového rozpětí asi sedmi desetiletí nebyli schopni rozeznat ani ty nejzákladnější výrazy. Zajímavější ovšem bylo pozorovat potenciální vývojové tendence v rámci specifické slovní zásoby. Zajímala nás především přítomnost slov poplatných době, které jsme hledali nejen u starších zkoumaných zdrojů (J), ale i u těch nejnovějších (LSRC, LVK2018).

I přestože jsme v práci v rámci korpusů zkoumali pro větší reprezentaci různých rovin jazyka také mluvená data, abychom mohli lépe zachytit rozdíly mezi jazykem mluveným a psaným, nedosahovaly zjištěné údaje původní míry. Původním předpokladem bylo, že se ve výsledné analýze objeví statisticky prominentnější přítomnost jednotek typických pro hovorovou mluvu (především částice, citoslovce, ale též hovorové podoby plnovýznamových slov). Například v korpusu LSRC se v námi omezeném okruhu slov objevilo pouze několik jednotek zcela příznakových pro mluvenou lotyštinu, jmenovitě se jednalo o tyto: *bišķiņ, bišķs, drusciņ, forš, mazlietiņ, nešpetns, skaistule, vot.*

Pro větší přehlednost uvádíme tato slova také v následující tabulce, v druhém řádku vždy připojujeme ekvivalenty těchto slov ve spisovném jazyce.

Bišķiņ{bišķs	**Drusciņ**	**Forš**	**Mazlietiņ**
Mazliet	*Mazliet*	*Labs*	*Mazliet*
Nešpetns	**Skaistule**	**Vot**	**pļāpāt**
niķīgs	*= skaista meitene*	*Nu*	runāt

Tabulka 1. Hovorová slova a jejich spisovné varianty

Pozoruhodná je např. přítomnost částice *vot*, běžně užívaná v ruštině; může být překvapivé, že se umístila v druhé tisícovce. Podle oficiálních informací o korpusu (viz 2.5.) by měly nahrávky shromážděné v korpusu pocházet v 60 % případů od mluvčích bez příznaků dialektu. Až 25 % všech zaznamenaných projevů však má údajně pocházet od mluvčích, pro které pravděpodobně lotyština není prvním jazykem a jejichž mluva může vykazovat nestandardní prvky. Je tedy pravděpodobné, že tato částice mohla pocházet převážně z improvizovaných projevů mluvčích, které nějakým způsobem pojí znalost ruštiny, případně ruský původ. Přesná distribuce ovšem v korpusu detailněji popsána nebyla, můžeme tedy pouze předpokládat, že v tomto případě jde o mluvčí bilingvní, případně znalé obou jazyků do té míry, že jejich projev byl považován za dostatečně autentický pro potřeby korpusu. Tuto hypotézu nebylo možné potvrdit, neboť jsme neměli k dispozici samotný korpus, kde by se dané výskyty výrazu daly blíže analyzovat, i s ohledem na příslušná metadata; jediným dostupným zdrojem nám byl pouze frekvenční slovník samotný.

Počet jednotek příznačných pro hovorový jazyk byl skutečně nízký, což znamená, že také zachycení rozdílů mezi psanou a mluvenou lexikální špičkou není dostatečně průkazné. Zvláštním jevem ovšem byla přítomnost slov příznakových pro mluvený jazyk i v tištěných zdrojích. Tato skutečnost se může jevit jako překvapující, zejména protože od tohoto typu textů bychom takovou míru hovorového jazyka nečekali. Důležité je si však uvědomit, že i v českém (a samozřejmě nejenom v českém) mediálním prostředí stále častěji dochází k pronikání prvků hovorového jazyka i do útvarů, kde by jeho přítomnost ještě před nějakou dobou byla považována za nevhodnou. To samozřejmě ani lotyštině není jev neznámý. Více o lotyštině v médiích např. Martišūne (2004)[38].

Příkladem slov příznačných pro hovorový jazyk byly pozdravy typu *labdien, labrīt* či *labvakar*. I přestože bychom jejich výskyt ve frekvenčních slovnících pro svou povahu nepředpokládali, tato tři slova se objevila v K, J a LSRC. Naopak u LVK2018, které obsahuje malé procento (2 %) mluvených projevů, resp. stenogramů z jednání parlamentu (Levāne-Petrova 2019: 133)[39], ani jedno z těchto slov nenacházíme. Možným vysvětlením těchto tendencí je míra spontánnosti a povaha zvukových nahrávek v korpusu. V případě

[38] MARTIŠŪNE, Signe, (2004). *Valodas lietujoms Latvijas elektroniskajos medijos:likumdošana un prakse*. Riga: Nordik.

[39] LEVĀNE-PETROVA, Kristīne, 2019. Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. *LU Humanitāro zinātņu fakultātes rakstu krājums*. Riga: LU Akadēmiskais apgāds.

improvizovaných hovorových projevů si lze spíše obtížně představit užití formálních oslovení. Naproti tomu u předem připravených projevů, např. v parlamentu, je již možné takovou jazykovou situaci velmi jasně rozeznat. Dalšími slovy typickými pro mluvený jazyk, které bychom ve frekvenční špičce psaných jazyků možná nehledali, jsou částice *paldies* a *lūdzu*. První zmiňovaná se umístila ve společném lexiku, objevovala se ve všech zdrojích; druhá se kromě J také vyskytovala ve všech zdrojích.

Další vlastností korpusů, která je odlišovala od frekvenčních slovníků, je přítomnost vyššího procenta vlastních jmen a toponym. Ve frekvenčních slovnících žádné takové jednotky nenacházíme (a to třeba ani lotyšské hlavní město *Rīga, z*ato obyvatelské jméno *rīdzinieks* ano). Tato skupina zahrnuje vlastní jména a názvy měst, a to jak lotyšských (*Valmiera*, *Daugavpils*), tak cizích (*Maskava*, *Londona*). Jejich procentuální zastoupení v korpusech je ovšem v rámci celých korpusů nepříliš významné**.** Oproti tomu však určitou část specifické slovní zásoby u J tvoří jednotky, které bychom opět mohly označit za jednotky poplatné době. Jednalo se například o názvy již zaniklých institucí nebo fenoménů spojených s dřívějším režimem, např. *PSKP* [*Padomju Savienības Komunistiskā partija*], *PSRS* [*Padomju Sociālistisko Republiku Savienība*], *CK* [*Centrālā Komitēja*], KP [*Komunistiskā partija*], *komjaunieties, komjaunatne, Latvijas PSR, ļeņinisks, ļeņinisms, sovhozs, spartakiāde, TASS* [*Padomju Savienības telegrāfa aģentūra*], *VDR* [*Vācijas Demokrātiska Republika*], *VFR* [*Vācijas Federatīvā Republika*]. Přítomnost těchto jednotek byla specifická pouze pro slovník J.

Zcela záměrně bylo usilováno o co největší rozmanitost zkoumaných zdrojů, aby se tyto tendence mohly následně projevit ve frekvenční špičce. Z tohoto důvodu byl do zkoumaných zdrojů zařazen také korpus mluveného jazyka LSRC. Cílem bylo zjistit, jaká specifika jsou tomuto korpusu vlastní a jaké prvky jsou naopak identické se zbylými zdroji.

Do LSRC byly zpracovány ukázky z rozličných televizních a rozhlasových kanálů a také přepisy projevů poslanců v lotyšské Saeimě. I přes poměrně široký záběr materiálu se dá předpokládat, že v korpusu nalezneme slova, která pravděpodobně v jiných datasetech nenajdeme. Jedná se např. o pozdravy, citoslovce**,** hovorové výrazy. Očekávali jsme, že tyto předpoklady budeme moci uplatnit do jisté míry i u korpusu LVK.

Vzhledem k celkově dosti širokému zkoumanému období, které je reflektováno stářím materiálů, se dalo předpokládat, že minimálně nejstarší materiály budou odrážet slovní zásobu poplatnou své době, respektive dobovému režimu. Tyto tendence byly nejlépe patrné v J, kde se objevovala nejčastěji v podstatě výlučně. Tento slovník obsahuje lexikum související se sovětským obdobím, především tehdejší reálie (viz výše). Výskyt těchto jednotek však není

v rozporu s výskytem jednotek patřící do frekvenční špičky – tyto jednotky tedy rozhodně netvořily nijak velké procento lexika, pro svou specifičnost a výskyt pouze v tomto slovníku je však na místě je uvést.

Vynecháme-li korpus LSRC, pro mluvený jazyk příznakového lexika je v dalších zkoumaných zdrojích jen několik málo jednotek. Jediným dalším datasetem, který je založen primárně na mluveném jazyce je frekvenční slovník K, naopak např. u korpusu LVK je procento zastoupeného mluveného jazyka velmi nízké, tj. 2%. Naopak několik jednotek příznakových spíše pro mluvený jazyk bylo pozorováno u K a J. Jednalo se o tyto jednotky: *nu* (K+J+LVK+LSRC), *pag!* (K), *paldies!* (JKL), *pietiek!* (K), *sveiki!* (K), *labrīt!* (K, J), *lūdzu!* (K, J) *lūk!* (K). Statisticky se však jedná o zanedbatelný počet lemmat a neprokazuje jasným způsobem významnou reprezentaci mluveného jazyka v těchto třech zdrojích (JKL). Jak vyplývá z tohoto nevelkého počtu příkladu, tyto jednotky příznakové pro mluvený jazyk se nejčastěji objevovaly v K, což odpovídá tomu, že je slovník založen na mluvených datech.

3.3. Potenciální vývojové tendence

Jak bylo zjištěno při srovnání dat, jádro slovní zásoby se významně nemění ani v průběhu několika desetiletí. Zajímavější a v oblasti lexikografie nutností je ovšem sledovat potenciální vývojové tendence a faktory, které na její utváření mohou mít vliv. Užití dat z korpusů poskytuje možnost skutečně reflektovat skutečný užívaný jazyk, ať už se jedná o korpusy složené z textů nebo zvukových nahrávek. I přestože se ale na první pohled může zdát, že v případě kompilace slovníků je užití dat z korpusů jediné a samospásné řešení, jejich užití není v lexikografii vždy zcela bezproblémové. Problémem často bývá velikost korpusů, která ne vždy odpovídá standardům nebo zamýšleným rozsahům připravovaných děl. Jak uvádí M. Škrabal (2016: 34): „Užívání lotyšských korpusů však má svá úskalí, která negativně ovlivňují především analytickou fázi lexikografické práce. Hlavním záporem je malá velikost LVK: na takto omezeném objemu dat lze zkoumat jevy pouze centrální, periferní úkazy se často vyskytují v několika málo výskytech (včetně nízké variantnosti flektivních forem daného lexému), případně se neobjevují vůbec. LVK by tak mohl posloužit jen jako materiálová základna pro mnohem menší překladový slovník (slovník výkladový lze pak vyloučit úplně)“[40]. Pro běžného uživatele však užití korpusu může znamenat lepší orientaci v aktuálně užívané

[40] ŠKRABAL, Michal, 2016. *Srovnávací aspekty lotyšského a českého lexikonu: Materiály k sestavení lotyšsko-českého slovníku*. Praha. Dizertační práce. Univerzita Karlova.

slovní zásobě, už jen z toho důvodu, že data z korpusu bývají označeny doplňujícími informacemi, které je velmi obtížné, nebo úplně nemožné ze standardního slovníku získat (jedná se právě o indikátory frekvence a příklady užití jednotky v autentickém textu aj.).

Na základě analýzy dat jsme zjistili, že ani korpusy, které by primárně měly odrážet jazyk aktuální, obsahují ve frekvenční špičce jen zanedbatelné množství soudobých reálií, klíčových slov dnešní doby. Např. v korpusu LSRC jsme se mezi třemi tisíci nejfrekventovanějších jednotek našli jen trojici slov reflektujících moderní technologie: *e-pasts, Facebook* a *internets*. U starších zdrojů, zvláště v J, najdeme podobě specifickou vrstvu dobově podmíněných lexémů (připomeňme, že se jednalo o jednotky, které souvisely s minulým režimem, jako *sovhozs, komjaunieties* nebo *Padomju Savienības Komunistiskā partija* a jako takové se už v moderních korpusech nevyskytují – protože spolu se změnou režimu zmizela z úzu).

3.4. Faktor diglosie

Vzhledem k povaze dat (frekvenční špičky opírající se o jazyk psaný vs. mluvený) se přímo nabízí zkoumat je i z hlediska potenciální diglosie. Ta je definována jako sociolingvistický jev, při kterém dochází k paralelnímu užívání dvou jazyků nebo jazykových forem, přičemž každá má svou specifickou funkci a zároveň buď vyšší, či nižší společenskou prestiž (Ferguson 1959)[41]. Problematikou diglosie se v české lexikografii zabývá např. M. Škrabal a Z. Komrsková[42]. V rámci svého výzkumu porovnávali míru diglosie v českém a anglickém jazyce, přičemž bylo zjištěno, že čeština – alespoň v rámci jimi zkoumaného lexikálního materiálu, frekvenčních špiček mluveného a psaného jazyka – vykazuje mnohem větší znaky diglosie než angličtina. Bylo zjištěno, kolik procent slovní zásoby se vyskytovalo pouze v psaném jazyku, jaké procento v mluveném jazyku a průnik obou těchto variant. Zatímco u angličtiny byly pro hovorovou mluvu signifikantní lexémy typické pro regionální varianty jazyka a v menší míře také vulgarismy, pro češtinu těchto kategorií slov bylo podstatně více. Byla to například určitá konkrétní substantiva, vulgarismy, hypokoristika, deminutiva a lexémy užívané ve zdvořilostních frázích.

[41] FERGUSON, A. Chatles, 1959. Diglossia, WORD, 15:2, S. 325-340

[42] Škrabal, M. – Komrsková, Z. (2019). *Top frequency words in written/spoken Czech and English – can they be used to measure a rate of diglossia?* Corpus Linguistics 2019, Cardiff.

Tento výzkum nás inspiroval k ověření situace v souvislosti s lotyštinou. Lotyština tradičně nebývá považována za jazyk s vysokou mírou diglosie, jako tomu je právě u češtiny (viz příklady: *dům* × *barák, láhev* × *flaška, ano* × *jo*), několik párových dvojic se nicméně najít i v ní. Pravděpodobně nejlépe je tento jev demonstrován v tabulce č. 1 výše, kde uvádíme přehled původních slov s jejich ekvivalenty ve spisovném jazyce. Všechna tato slova byla nalezena v korpusu LSRC. S výjimkou slov *druscіņ* a *bišķiņ* (1. tis.) se všechna slova nacházela v druhé tisícovce frekvence. Oproti výše uvedenému příkladu srovnání míry diglosie v českém a anglickém jazyce se však v námi zkoumanému vzorku nepodařilo najít žádný obsáhlejší vzorek.

4. Využití frekvenčních údajů v dnešní lexikografii

Při kompilaci slovníků je potřeba stanovit, jakým principem bude probíhat výběr hesel do hesláře. Dřívější tendence kompilovat slovníky na základě již publikovaných dat svých předchůdců se mnohdy prokázaly jako ne zcela vhodný způsob, automaticky se tak přejímala mnohdy slova, která mezitím zmizela z reálného úzu, resp. přesunula se z centra slovní zásoby na její periferii.

LČPS se jasně profiluje jako na korpusu založený (*corpus-based*), a to už od samého počátku, kdy první verze hesláře (z konce roku 2008) vznikla na základě korpusu lotyšských textů o velikosti 11,5 milionů pozic, jenž sloužil jako materiálová základna pro pozdější korpus LVK2013. Z tohoto zdroje bylo extrahováno 25 758 jednotek o minimální frekvenci 3. Tento původní heslář byl následně doplňován o některá chybějící hesla z LVV 2006*, Latviešu valodas vārdnīca*), případně i další zdroje včetně těch elektronických. Aktuálně činí objem hesláře kolem 35 tisíc položek, ale má dále narůstat.

S rostoucím počtem slovníkových hesel úměrně vzrůstá i tzv. lexikální pokrytí textu, tedy poměr slov v textu, které lze (alespoň teoreticky) pomocí slovníku úspěšně dekódovat. F. Čermák (2010a: 245) s oporou v datech z korpusu SYN2000 udává tyto poměry: znalost 1000 nejfrekventovanějších českých slov by měla stačit k porozumění 64,9 % slov v libovolném českém textu, 10 000 slov k porozumění 91,38 % textu, procenta se dále zvyšují (20 000 – 95,53 %; 30 000 – 97,09 %; 40 000 – 97,86 %; 50 000 – 98,27 %). Pro lotyštinu jsou dostupné údaje z LVBV (s. 13): 1000 nejfrekventovanějších slov – 64 %, 2000 – 75%, přičemž nejvyšší pokrytí textu vykazují beletristické texty (69, resp. 78 %), nejnižší texty vědecké (55, resp. 64 %). Srov. též údaje uváděné v K (Kuzina 1998a, 1998b) s maximální hodnotou 5000 slov – cca 93 % textu.

Položky jsou do hesláře LČPS zařazovány primárně na základě frekvenčního kritéria, tj. podle frekvence dané lexikální jednotky v dostupných datasetech. Čistě kvantitativní kritérium v sobě však nese několik úskalí, která je potřeba vyvážit aplikací dalších sekundárních kritérií. Jedná se o tato kritéria: nominativní (pojmenovávací) reprezentativnost[43], výskyt polysémie a její míra (čím vyšší míra, tím větší je oblast, kterou lexikon pokrývá, z čehož následně vzniká potřeba zařadit ho rovněž do hesláře). Dalším kritériem je míra derivačního potenciálu, což prakticky znamená nutnost zařazení lexému sloužícího jako slovotvorný základ pro jiné jednotky. Nezanedbatelný kritériem je také invariantnost – pokud je vyhrazený prostor

[43] Nominativní (pojmenovávací) reprezentativnost, tedy „výběrová úplnost pokrytí zvolení úrovně pojmenovávacích potřeb daného typu slovníku“ (Čermák 1995d.: 234)

limitován a je potřeba mezi vícero jednotkami vybrat jen omezené množství, volí lexikograf zpravidla bezpříznakovou variantu.[44]

Frekvenční kritérium však zůstává prvořadým aspektem chystaného LČPS. Tento aspekt je obzvláště důležitý například tehdy, pokud nastane situace, kdy je třeba vybrat k zařazení specifickou jednotku, která se sice v materiálech objevuje, ovšem ne natolik hojně, aby byla do hesláře skutečně zařazena. Od hlediska frekvenčního se odvíjí značkování diafrekvenční, lišící slova běžně se vyskytující od těch řídkých či vzácných. Tento typ markeru nevyužíváme: frekvenční kritérium je pro nás prvořadé, a není-li pro to pádný důvod, slova nedoložená či doložená jen spoře nejsou do hesláře LČPS zařazena.

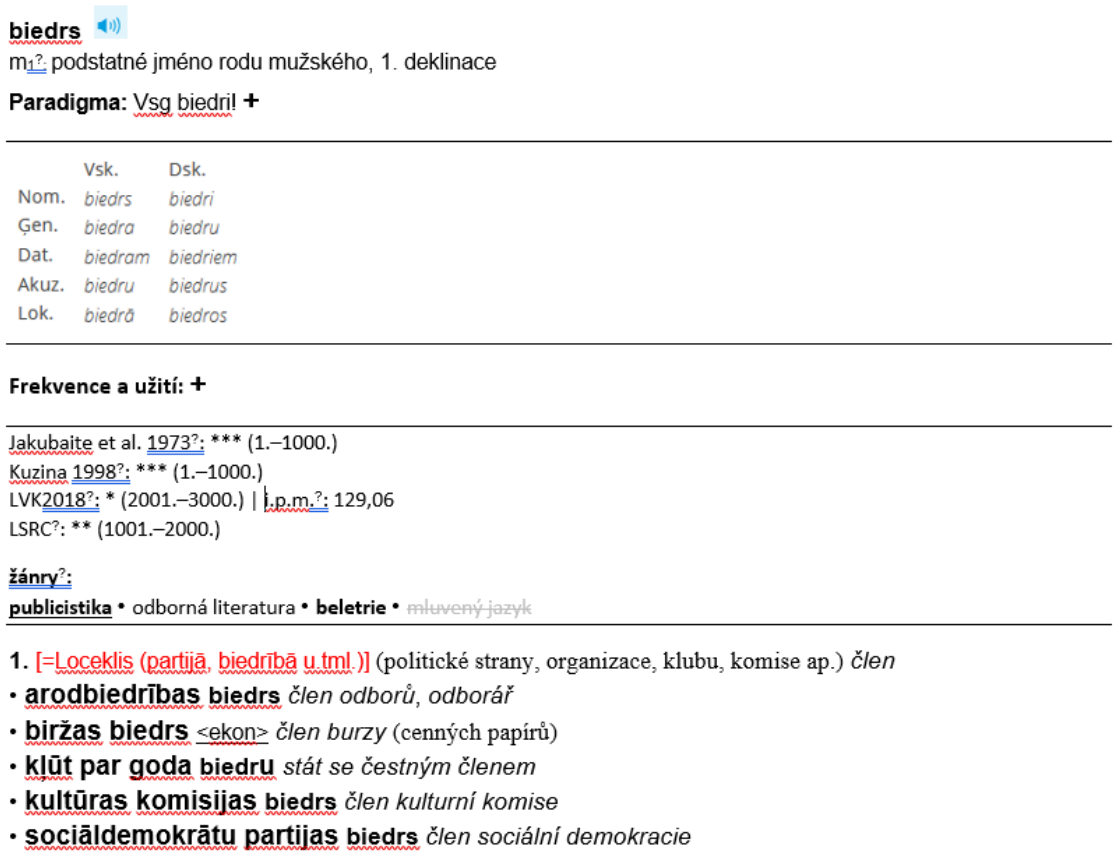
5. Aplikace 1–3 v budoucím LČPS – „frekvenční modul“

V této části popíšeme metodu, jakou by se výše extrahované frekvenční špičky daly použít v chystaném LČPS v podobě jakéhosi „frekvenčního modulu“. Slovník počítá i se zahrnutím mimo jiné i informace o četnosti vybraných jednotek, zvláště těch nejfrekventovanějších. Jde tedy o praktickou aplikaci našich poznatků a dat pro účely konkrétního lexikografického díla.

5.1. Ukázkové heslo *biedrs*

Získaná data demonstrujme na vybraném slovníkovém hesle. Na obrázku níže je zobrazena relevantní část heslové stati *biedrs*[45], tj. ta, jež obsahuje informaci o četnosti daného lexému.

[44] (Čermák 1995d: 235).
[45] Jde o pouhou pracovní verzi, která se může ještě změnit



Obrázek 3. Možná aplikace frekvenčních dat v LČPS

Heslová stať je co do své stavby a posloupnosti jednotlivých složek víceméně tradiční, využívá výhod elektronického média s prakticky neomezeným prostorem, což dosvědčuje kromě vertikální orientace (oproti výchozí horizontální orientaci ve slovnících tištěných) také např. explicitní rozepisování zkratek a značek apod.

Tradičně na začátku hesla je uživatel slovníku seznámen s výslovností (pomocí nahrávky, případně též upozornění na nepravidelnou či potenciálně problematickou výslovnost) a se základními gramatickými informacemi (slovnědruhová příslušnost, nepravidelné či jinak problematické tvary, po rozkliknutí celé deklinační či konjugační paradigma).

Následovat by měla informace o frekvenci a užití; ta je koncipována tak, že bude uživateli k dispozici až po rozkliknutí ikonky +; zároveň je možné, že v pozdější fázi tvorby LČPS se může přemístit na jiné, perifernější místo heslové stati – důvodem pro to je skutečnost, že tento typ informace nepatří k těm, jež by předpokládaného uživatele zajímaly primárně (tou bude patrně především sémantická stránka, zvláště pak české ekvivalenty lotyšských lexémů).

Na druhou stranu nebude-li frekvenční statistika defaultně zobrazena explicitně, ale až na vyžádání, nepůsobí příliš rušivě a může zůstat i na tomto poměrně exponovaném místě stati.

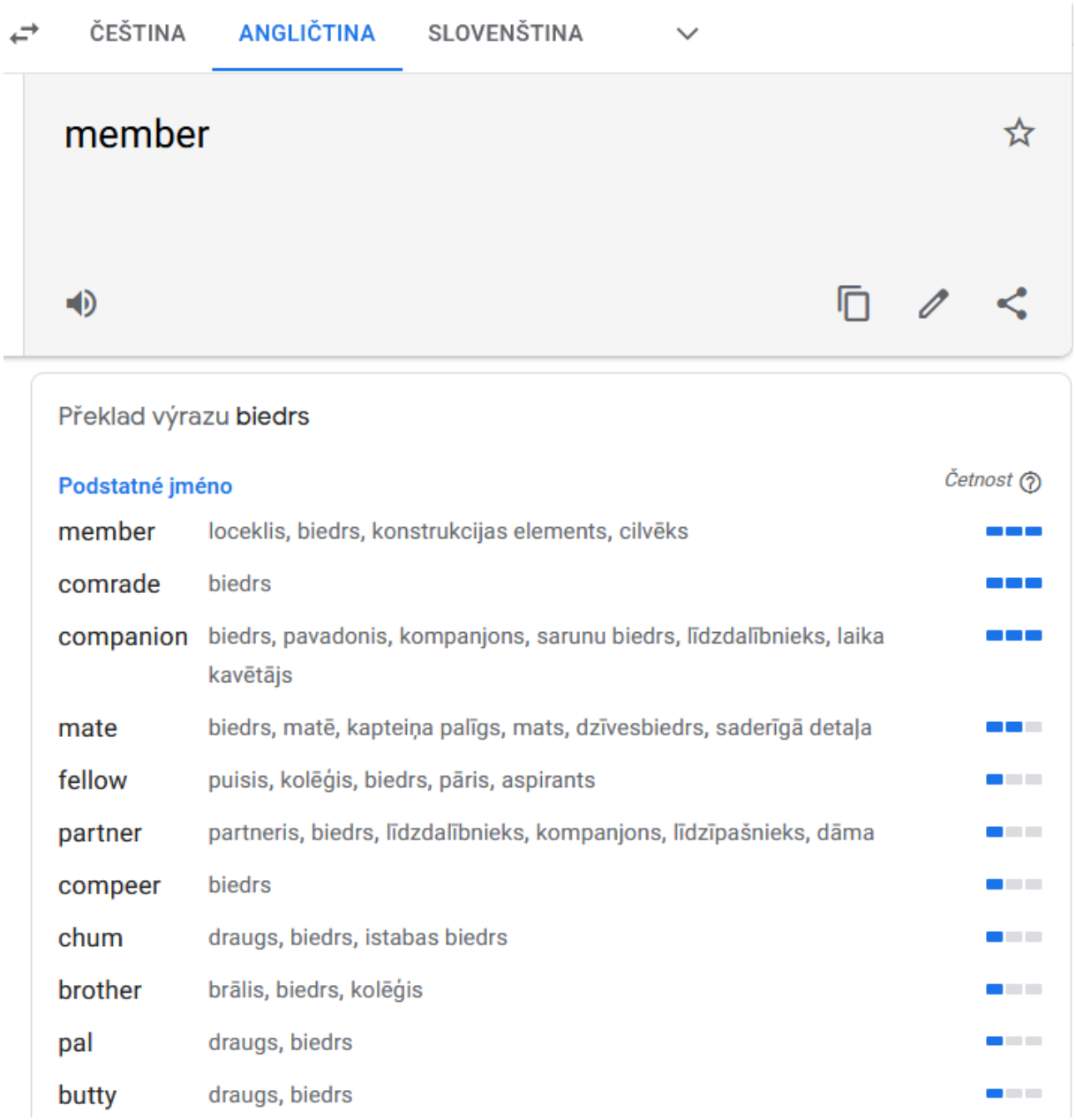
Frekvence je uvažována vzhledem ke čtyřem zkoumaným zdrojům (J, K, LVK, LSRC), vždy je uvedené, do jaké tisícovky frekvence daná jednotka patří. Další užitečnou informací je žánr, ve kterém se daná jednotka nejčastěji objevuje. Pokud je výskyt v daném žánru významněji zastoupen, je graficky zvýrazněn tučným písmem a podtržením. Pokud je výskyt v určitém žánru alespoň částečně relevantní, je daný žánr uveden bez jakýchkoliv grafických úprav. Pokud se však daná jednotka v příslušném žánru neobjevuje vůbec či minimálně, případně její výskyt není ze statistického hlediska významný, je označení žánru přeškrtnuto a vyznačeno šedivým písmem. Tato informace může být užitečná uživateli, který chce získat komplexnější pohled na užití slova v kontextu a typ textu, v němž by se s ním mohl nejpravděpodobněji setkat. Aktuální verze pracuje se čtyřmi makrotypy: první tři – *publicistika*, *odborná literatura* a *beletrie* – odrážejí psanou (a většinou i redigovanou) formu jazyka, kdežto poslední reprezentuje jazyk mluvený, většinou v jeho spontánní podobě. Mezi jednotlivými datasety a těmito makrotypy tak existuje úzká souvislost, např. objeví-li se nějaký lexém ve frekvenční špičce datasetů K a/nebo LSRC, bude označen v příslušném řádku odpovídajícím počtem hvězdiček a makrotyp *mluvený jazyk* bude vytučněn.

Zde je na místě připomenout, z jakých zdrojů byly kompilovány samotné zdroje, které jsou předmětem zkoumání této práce (detailněji viz kap. 2). Připomeňme například, že Jakubaite zkoumala takové okruhy pouze tři (technika a průmysl, beletrie, věda), korpus LVK naproti tomu zkoumal žánrově pestřejší materiály (beletrie, periodika, vědecké texty, regulační předpisy a stenografické přepisy ze Saeimy). Pokud dva odlišné materiály zpracovávají odlišné žánrové okruhy, může to samozřejmě významným způsobem měnit podobu frekvenční špičky a následně také žánry, ve kterých jednotky z frekvenční špičky budou dominovat, tj. pokud budeme uvažovat převážně korpusy zohledňující speciální okruhy a slovní zásobu, výrazně se to projeví na příslušnosti slovních jednotek k žánrovým okruhům. Je nicméně zřejmé, že určitá míra sjednocení různých žánrových klasifikací do případných makrotypů bude patrně nevyhnutelná.

Po frekvenční části následují další typy informace, počínaje sémantikou, ty už jsou však pro účely naší práce nepodstatné, a proto se jimi nebudeme dále zabývat.

5.2. Srovnání LČPS s překladačem Google Translate

Zcela přirozeně vyvstává otázka, nakolik jsou spolehlivé a uživatelsky spolehlivé také volně dostupné zdroje. K prvním zdrojům, po kterých student nebo uživatel jazyka sáhne, patří velmi populární služba Google Translate. Přestože je patrně většině uživatelů známa celá řada úskalí práce s tímto zdrojem, řadí se tento internetový slovník k těm nejužívanějším. Velkou předností (která se však může projevit na kvalitě dat) je velká četnost možných jazykových kombinací a poměrně přívětivé uživatelské rozhraní. Překladač GT podává částečnou informaci o frekvenci lexémů, resp. jejich cizojazyčných ekvivalentů, i když jen pro vybrané jazykové kombinace, viz např. lotyšsko-anglické nastavení. V tomto nastavení se uživateli objeví několik možných překladů, ty jsou řazeny podle frekvence sestupně. Nejméně odpovídající překlad v cílovém jazyce se tedy objeví až pod těmi frekventovanějšími. To je velmi jednoduchá nápověda uživatelům, kteří ještě jazyk neovládají na takové úrovni, aby na jejich základě mohli odhadnout nejvhodnější ekvivalent hledaného slova. Tyto doprovodné informace ovšem nejsou dostupné po všechny jazykové kombinace. Budeme-li to chtít demonstrovat na již užitém slově *biedrs*, zjistíme, že lotyšsko-české rozhraní nabídne pouze jeden český ekvivalent, a to *člen*. Takový překlad je pouhým rychlým hledáním určitého (v mnoha případech ani ne plně odpovídajícího) ekvivalentu.



Obrázek 4. Návrhy překladu slova *biedrs* spolu s frekvenční informací v lotyšsko-anglické části překladače Google Translate

Ve druhém sloupce se vedle tučně vyznačených jednotek vyskytují ještě synonyma, resp. ekvivalenty k anglickým lexémům (kterých bylo celkem 19, srov. s jediným ekvivalentem českým). Lotyšské slovo *biedrs* je tedy podle GT možno přeložit devatenácti různými anglickými lexémy. Zobrazení dalších vhodných synonym, frekvence užití apod. tedy prozatím funguje pouze u velkých a zdrojově bohatých jazyků, u těch ostatních je možné využít pouze jednoduchý překlad. S větším počtem možností a ekvivalentů však roste možnost výběru ne

příliš vhodné jednotky. Jak uživatel s nižší znalostí jazyka bude vědět, že zvolil adekvátní překlad? Takovým nedorozumění by měl zamezit právě frekvenční údaj, který uživateli může pomoci ve výběru vhodné jednotky. Je velmi nepravděpodobné, že se kýžený ekvivalent bude skrývat na posledním místě možných překladů, obecně se tedy dá říci, že čím výše se daná jednotka umístila (tedy čím frekventovanější je), tím pravděpodobněji se bude jednat o vhodný překladový ekvivalent.

Jako druhý příklad můžeme zkusit překlad slova ze sekce 3.2., konkrétně slova *skaistule* (‚kráska'). Jedná se o slovo hovorové, a tak nás právem zajímá, jak dobře si s ním překladač poradí. V lotyšsko-anglickém rozhraní nabízí dva překlady: v prvním případě nabízí abstraktum *beauty* (*krása*, v určitých kontextech také *kráska*) a podporuje ho synonymy jako *skaistums* a *daiļums* (která neodpovídají příliš skutečnému významu lotyšského slova). Tento první příklad byl vyhodnocen jako frekvenčně častější, proto se také objevil hned na prvním místě v možnostech překladu. Námi hledaným ekvivalent se však objevuje spíše na druhém místě, tedy u slova *belle = kráska*. V předchozím odstavci jsem uvedla, že údaje o frekvenci mohou uživateli pomoci se snáze rozhodnout pro jednu z nabízených variant. Problém ovšem nastává v případě, že tento údaj nebude patřičným způsobem reflektovat povahu slova. Pokud by se uživatel řídil čistě na základě údaje o frekvenci, pravděpodobně by vybral překlad první. Zarazit ho pak může (ale nemusí), že k abstraktu *skaistums* v překladu nachází abstraktum netradičního tvaru. Pokud však budeme chtít dané slovo přeložit v lotyšsko-českém rozhraní, dostaneme pouze jeden překlad, a to již zmíněné abstraktum *krása*. To ale původnímu významu slova neodpovídá a uživateli poskytuje nepřesný údaj.

Samostatnou kategorií jsou překlady zcela nevhodné a zavádějící. Jako příklad jsem vybrala opět slovo ze sekce 3.2. *bišķiņ* (‚troš(ič)ku'). Jak v lotyšsko-anglickém, tak v lotyšsko-českém rozhraní dostáváme překlad *bee* (respektive *včela*). Zde se jedná o zcela chybný protějšek, i pokud bychom chtěli brát v úvahu, že překladač vyhodnotil slovo *bišķiņ* jako deminutivum, i jeho tvar by byl chybný, protože správný tvar tohoto slova by zněl *bitīte* (‚včelka'). Zde tedy může dojít k opravdu závažné záměně slova z nepochopení (špatného vyhodnocení překladačem). [46]

Každý uživatel by si měl být těchto úskalí vědom. Vzhledem k povaze chystaného LČPS je možné se takovým nedorozuměním zcela vyhnout právě díky užívání takových údajů,

[46] Srov. lotyšský internetový slovník tezaurs.lv (dostupné z: https://tezaurs.lv/bišķiņ) slovo ve své databázi má společně s údajem, že se jedná o slovo užívané slangově, stejně tak uvádí několik vhodných příkladů z korpusových textů. Ty jsou ovšem samozřejmě vhodné až pro uživatele, který nějaké povědomí o lotyštině má.

jako je třeba výskyt v příslušném žánrovém okruhu. Tak by třeba uživatel jednoduše zjistil, že jím hledané slovo *skaistule* se objevuje v mluveném jazyce. Příklady užití slova v kontextu mohou také významně napomoci hledání správného ekvivalentu.

Závěr

Tato práce teoreticky i prakticky zohlednila poznatky korpusové lingvistiky v lotyšském prostředí. Teoretická část se zaměřila na popis počátků a aktuálních trendů v lotyšské lexikografii a může posloužit jako přehledný výčet lexikografických projektů, které v současné době mohou zájemci využívat. V rámci praktické části jsme vydělili nejfrekventovanější lexémy z dostupných datasetů a reflektovali tak jádro slovní zásoby lotyšského jazyka. Data jsou k práci připojena v příloze, čtenáři tedy mohou rovněž prakticky posloužit; s dalším využitím se počítá v LČPS. Před samotným zpracováním dat bylo vytvořeno několik hypotéz, které jsme se snažili na základě zjištěných dat buď potvrdit, nebo vyvrátit. Zároveň byl zdigitalizován heslář frekvenčního slovníku Kuziny a Jakubaite (3288 jednotek).

Na základě srovnání materiálu reflektující různá období byla zjištěna frekvenční špička, tvořící základ standardní lotyštiny již po několik desetiletí. Těchto jednotek bylo z námi zkoumaných čtyř datasetů celkem 1337. Tyto lexémy tvořily určité procento obsahu, od něhož bylo následně dopočítáno, kolik procent specifické slovní zásoby každý materiál obsahoval. To nám umožnilo udělat si lepší představu o zkoumaném zdroji, protože pouhá charakteristika zdrojů je základem, nedává však přesnou představu o tom, jak extrahované specifické lexikum bude vypadat, zda dataset reflektuje určité okruhy slovní zásoby apod. Na základě těchto charakteristik jsme mohli předpokládat reprezentaci určitých odvětví nebo vrstev jazyka, např. statisticky významnější reprezentaci lexika týkajících se specifických odvětví – technika, zemědělství aj. u Jakubaite nebo právě reprezentaci jednotek mluvené, hovorové lotyštiny (K, LSRC).

Výsledky byly poměrně překvapivé, což potvrdila například distribuce procentuálního rozvrstvení slovní zásoby. Společné lexikum pro všechny zdroje v žádném ze zkoumaných příkladů nedosahovalo více než dvou pětin (40,8 %). To svědčí o poměrně významném procentu specifické slovní zásoby v našich zdrojích, vypočteném na základě součtu procent indikujících společné lexikum pro všechny, tři a pouze dva zdroje. Zbylé procento zachycovalo lexikum specifické právě pro jeden zdroj. Předpokladem bylo, že toto procento bude statisticky významně reprezentovat jednotky, které by měly podle charakteristik zkoumaného lexika být unikátní právě pro každý zdroj a měly by vykazovat příslušné charakteristiky (jako hovorový

jazyk, odborný jazyk určitého odvětví apod.). Toto procento však v mnoha případech reflektovalo slovní zásobu, která sice byla unikátní, nicméně se ji v mnoha případech nepodařilo zařadit do konkrétní skupiny, jež by jasně podtrhovala charakter zkoumaného zdroje. To svědčí o jisté univerzálnosti zdrojů, které i přes svou specifičnost nutně reprezentují široký lexikon, alespoň ve své frekvenční špičce. Projevilo se tedy, že velká část slovní zásoby se ani v průběhu několika desetiletí nezměnila, což je pochopitelné. Největší rozdíly na poli slovní zásoby byly zaznamenány v případě specifických lexémů, které označovaly soudobé reálie, v dnešní době známé pouze jako historický fenomén (*CK, PSKP, PSRS* atd.). Stejně tak tomu bylo i v opačném případě – v novějších datasetech nacházíme lexémy reflektující skutečnosti, které jsou naopak příznačné pro modernější dobu (*e-pasts, Facebook* atd.).

Rovněž jsme se pokusili odhalit možné stopy diglosie v lotyštině. Přestože se nepodařilo shromáždit větší množství případů, nalezené dvojice jsme evidovali. Záhodno by nicméně bylo udělat na toto téma samostatný, hlubší výzkum.

Na závěr práce jsme demonstrovali aplikaci získaných poznatků na ukázkové slovníkové heslo. Doplňkově jsme porovnali chystaný LČPS s výsledky z překladače Google Translate. Úskalí užívání elektronického překladače byla demonstrována na základě tří lexémů, které byly zmíněny v různých částech této práce.

Součástí práce je rovněž seznam slovní zásoby všech datasetů s vyznačeným společným jádrem (viz příloha).

Poznámka k příloze: Seznam obsahuje i chybně lemmatizované příklady (např. komisijā, žurnālistiem). Tyto tvary jsou opatřeny hvězdičkou. Je důležité odlišovat tyto špatně lemmatizované tvary od párových dvojic, u kterých se v jednom případě původní koncovka lexikalizovala a vznikl nový lexém (např. augša x augšā).

Bibliografie

1. ATKINS, B. T. S. – Rundell, M., 2008. *The Oxford Guide to Practical Lexicography.* New York: Oxford University Press Inc.
2. ČERMÁK, F. – Blatná, R. (eds.), 1995. *Manuál lexikografie*. Jinočany: H&H.
3. FERGUSON, A. Chatles, 1959. Diglossia, WORD, 15:2, S. 325-340
4. HAUSMANN, F. REICHMANN, J. WIEGAND, O. ZGUSTA, L. (eds.). (1989–1991).
   Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie [sv. 1–3]. Berlin − New York: Walter de Gruyter
5. JAKUBAITE et al., 1973. *Latviešu valodas biežuma vārdnīca. [Apvienotais (1.–3.) sējums].* Rīga:Zinātne.
6. KUČERA, Henry a W.Nelson FRANCIS, 1969. Computational Analysis of Present-Day American English. *International Journal of American Linguistics*. (35), 71-78.
7. KUZINA, V.,1998. *3000 latviešu sarunvalodas biežak lietotie vardi ar tulkojumu krievu, vācu un angļu valodā*. Rīga: Valsts valodas centrs.
8. LEVĀNE-PETROVA, Kristīne, 2012. Līdzsvarots mūsdienu latviešu valodas tekstu korpuss un tā tekstu atlases kritēriji. Baltistica., *VIII priedas*, Vilnius: Vilniaus Universitetas,
9. MARTIŠŪNE, Signe, 2004. *Valodas lietujoms Latvijas elektroniskajos medijos:likumdošana un prakse*. Riga: Nordik.
10. NIKUĻCEVA, S, 2013. Ekvivalence lexémů v Česko-lotyšském slovníku: přístup a zpracování. In P. Štoll et al. *Zkušenosti a vztahy. Lotyšská a česká společnost ve 20. století.* Praha: FF UK, s. 213–225.
11. PINNIS, Mārcis – AUZIŅA, Ilze – Goba, Kārlis (2014). Designing the Latvian Speech Recognition Corpus - presentation.
12. ŠKRABAL, M., 2013. Moderně či tradičně? Chystaný lotyšsko-český slovník v konfrontaci s dosavadními plody česko-baltské lexikografie. In P. Štoll et al. *Zkušenosti a vztahy. Lotyšská a česká společnost ve 20. století.* Praha: FF UK, s. 227–233.
13. ŠKRABAL, M., 2016. *Srovnávací aspekty lotyšského a českého lexikonu (Materiály k sestavení lotyšsko-českého slovníku).* Disertační práce. Praha: FF UK.

14. ŠKRABAL, M. – Komrsková, Z., 2019. *Top frequency words in written/spoken Czech and English – can they be used to measure a rate of diglossia?* Corpus Linguistics 2019, Cardiff

Korpusy

1. ATILF, 2021. Frantext, Ddostupné z: https://www.frantext.fr
2. Corpus of Contemporary American English, 1990-2020. Ddostupné z: https://www.english-corpora.org/coca/
3. Český národní korpus. *Korpus monologů: ORATOR* [online]. [cit. 2021-5-14]. Dostupné z: https://wiki.korpus.cz/doku.php/cnk:orator
4. Český národní korpus. *Korpus neformální mluvené češtiny s víceúrovňovým přepisem: ORTOFON* [online]. [cit. 2021-5-14]. Dostupné z: https://wiki.korpus.cz/doku.php/cnk:ortofon#korpus_neformalni_mluvene_cestiny_s_viceurovnovym_prepisemortofon
5. Český národní korpus. *Korpus ORAL* [online]. [cit. 2021-5-14]. Dostupné z: https://wiki.korpus.cz/doku.php/cnk:oral
6. FullStack., Ddostupné z: https://github.com/LUMII-AILab/FullStack/pulls
7. LaRKo, 2014. Ddostupné z: http://www.korpuss.lv/id/LaRKo
8. Latvijas Universitātes Matemātikas un informātikas institūts. Latviešu valodas runas korpusā, 2014. http://larko.ailab.lv/index.php/info
9. LaVa, 2018-2021. *Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums*Latviešu valodas apguvēju korpuss., Ddostupné z: http://lava.korpuss.lv
10. Leibniz-Institut für Deutsche Sprache, 2021. DeEReEKoO, Ddostupné z: https://www1.ids-mannheim.de/kl/projekte/korpora.html
11. LINDAT, Digital Research Infrastructure for Language Technologies, Arts and Humanities, LVTB Latvian dependency constituency treebank., Ddostupné z: https://lindat.mff.cuni.cz/services/pmltq/#!/treebank/lvtb25/query/
12. Lists: Prohlížeč frekvenčních seznamů, SYN2015. Ddostupné z: https://www.korpus.cz/lists
13. LVK2018, 2016-2018., Ddostupné z: http://www.korpuss.lv/id/LVK2018
14. Senie, Latviešu valodas seno tekstu korpuss., Ddostupné z: http://senie.korpuss.lv/toc.jsp

15. SKETCH ENGINE. BROWN corpus: Corpus of American English. Dostupné z: https://www.sketchengine.eu/brown-corpus/
16. SKETCH ENGINE. Emuāri, Latviešu valodas emuāru korpuss., Ddostupné z: http://nosketch.korpuss.lv/#dashboard?corpname=emuari
17. SKETCH ENGINE. LVK2018. http://nosketch.korpuss.lv/#dashboard?corpname=LVK2018
18. SKETCH ENGINE. MuLa, Mūsdienu latgaliešu tekstu korpuss., Ddostupné z: http://nosketch.korpuss.lv/#dashboard?corpname=mula
19. SKETCH ENGINE. Raiņa darbu korpuss., Dostupné z: http://nosketch.korpuss.lv/#dashboard?corpname=rainis
20. Statistika. http://runa.korpuss.lv

Články

1. Accessible Speech Recognition, Signal to Noise Ratio Estimation. https://www.isip.piconepress.com/projects/speech/software/legacy/signal_to_noise/index.html
2. BUŠS O., BALDUNČIKS J. 1000 vārdu**:** Latviešu valodas leksikas minimums ar tulkojumu krievu un angļu valodā/ LZA VLI. Atb. red. O. Bušs. - Rīgā: Zinātne, (1991) . - 48 lpp
3. FREIDENFELDS 1967a – Freidenfelds, Ilmārs. Par prievārdu un partikulu homonīmiju. Latviešu valodas teorijas un prakses jautājumi. Rīga : Zvaigzne, (1967), 34.–44. lpp.
4. FREIDENFELDS 1967b – Freidenfelds, Ilmārs. Pusprievārdi latviešu valodā. Latviešu valodas teorijas un prakses jautājumi. Rīga : Zvaigzne, (1967), 45.–54. lpp.
5. LEVĀNE-PETROVA, Kristīne, 2019. Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. *LU Humanitāro zinātņu fakultātes rakstu krājums*. Riga: LU Akadēmiskais apgāds.
6. LORENCS A., NESAULE Z. (Лоренц А.А., Несауле З.Э.) (1967). Статистические свойства латышского языка. АН Латв СССР, Кибернетика, 10 (195), 1963, с.41-48.

7. SOIKANE-TRAPĀNE, M. (1987). Latviešu valodas pamata un tematisko vārdu krājums, ALA
8. ŠKRABAL, M. – KOMRSKOVÁ, Z. (2019). *Top frequency words in written/spoken Czech and English – can they be used to measure a rate of diglossia?* Corpus Linguistics 2019, Cardiff.
9. VALODAS SITUĀCIJA LATVIJĀ 2010-2015. https://valoda.lv/wp-content/uploads/aktual/Val_sit_informat_lapa_3.pdf
10. *Valodas situācijas sociolingvistiskā izpēte.* Kvantitatīvā pētījuma rezultātu ziņojums. Pasūtītājs: LVA. Īstenotājs: SIA „Excolo Latvia", (2014) dostupné z: https://valoda.lv/wp-content/uploads/docs/Petijumi/Sociolingvistika/VSL_2015_web.pdf
11. *Valodas situācijas sociolingvistiskā izpēte*. Kvantitatīvā pētījuma rezultātu ziņojums. Pasūtītājs: LVA. Īstenotājs: SIA „Aptauju aģentūra", (2012) Dostupné z: https://valoda.lv/wp-content/uploads/docs/Petijumi/Sociolingvistika/VSL_2015_web.pdf