Univerzita Karlova

Filozofická fakulta

Ústav českého jazyka a teorie komunikace

Diplomová práce

Bc. Ondřej Schmid

Idiolekt Václava Klause: korpusová analýza

Václav Klaus' Idiolect: A Corpus-based Analysis

Praha 2024

Vedoucí práce: Mgr. Michal Škrabal, Ph.D.

Poděkování

Srdečně bych chtěl poděkovat vedoucímu diplomové práce doktoru Michalovi Škrabalovi za jeho velkou trpělivost a velmi podnětné konzultace. Dále bych chtěl poděkovat zaměstnancům Ústavu Českého národního korpusu za praktickou pomoc s budováním korpusu a s . Moc děkuji svojí rodině za podporu. Nejvíc děkuji svojí přítelkyni, za všechno.

Prohlášení

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

V Praze, dne 19. května 2024 Ondřej Schmid

Abstrakt

Tato diplomová práce zkoumá idiolekt Václava Klause s využitím nástrojů a metod korpusové lingvistiky. Hlavním cílem výzkumu je komplexní analýza Klausova idiolektu a jeho potenciálních proměn v čase a v závislosti na veřejných funkcích, které Klaus vykonával, přičemž největší důraz je kladen na lexikální rovinu idiolektu, která nejvíce vypovídá o významech, jež jsou sdělovány. Aby mohlo být dosaženo hlavního cíle, byl pro účely výzkumu vybudován autorský korpus textů Václava Klause, sestavený z jeho textů veřejně dostupných na jeho oficiálních webových stránkách. Hlavní metodou analýzy dat je v tomto výzkumu korpusová metoda analýzy klíčových slov, která funguje na principu srovnávání relativních frekvencí jednotek (většinou slovních tvarů či lemmat) ve zkoumaném textu nebo korpusu s relativní frekvencí týchž jednotek v referenčním korpusu, který obvykle slouží jako standard jazykového úzu. Na základě tohoto srovnávání s vhodnými referenčními korpusy jsou v korpusu Klausových textů identifikovány klíčové jednotky (ve většině provedených analýz lemmata), které se v něm vyskytují statisticky signifikantně častěji, než by se dalo očekávat na základě jejich výskytu v referenčním korpusu. Kombinací dalších kvantitativních i kvalitativních metod jsou následně určeny ty klíčové lexikální jednotky, které lze považovat za součást Klausova idiolektu. Kromě hlavních analýz lexikální roviny idiolektu jsou v rámci vedlejších cílů výzkumu dále analyzovány klíčové slovesné tvary a Klausovy charakteristické slovotvorné afixy *ne-* a *-ismus*. Také je navržen způsob, jak lze využít analýzu klíčových slov při diskurzní analýze Klausových projevů. Jedním z účelů předkládaného výzkumu je v praxi otestovat nový korpusový nástroj *Analýza klíčových slov* vyvinutý v rámci Českého národního korpusu, přispět k jeho vylepšení a ukázat možnosti, k čemu a jakým způsobem se dá využívat.

Klíčová slova

Václav Klaus, idiolekt, autorský korpus, klíčová slova, lexikologie, analýza diskurzu, klausismy

Abstract

This thesis explores the idiolect of Václav Klaus using the tools and methods of corpus linguistics. The main aim of the research is a comprehensive analysis of Klaus' idiolect and its potential changes over time and depending on the public offices that Klaus held, with particular emphasis on the lexical level of the idiolect, which is most telling about the meanings that are conveyed. In order to achieve the main research goal, a corpus of Václav Klaus' texts was built for research purposes, compiled from his texts publicly available on his official website. The main method of data analysis in this research is the corpus method of keyword analysis, which works on the principle of comparing the relative frequencies of units (usually word forms or lemmas) in the text or corpus under study with the relative frequency of the same units in a reference corpus, which usually serves as a standard of language usage. On the basis of this comparison with appropriate reference corpora, key units (lemmas in most of the analyses conducted) are identified in the Klaus corpus that occur statistically significantly more frequently than would be expected on the basis of their occurrence in the reference corpus. A combination of other quantitative and qualitative methods is then used to identify those key lexical units that can be considered part of Klaus' idiolect. In addition to the main analyses of the lexical level of the idiolect, the key verb forms, and Klaus' characteristic word-formation affixes *ne-* and *-ismus* are further analysed as secondary research objectives. It is also suggested how keyword analysis can be used in the discourse analysis of Klaus' speeches. One of the purposes of the present research is to test in practice the new corpus tool *Analýza klíčových slov* (*Keyword Analysis*) developed within the Czech National Corpus, to contribute to its improvement, to show the possibilities of what it can be used for and how it can be used.

Keywords

Václav Klaus, idiolect, author corpus, keywords, lexicology, discourse analysis, klausisms

Obsah

Seznam zkratek

CADS – corpus assisted discourse studies [studium diskurzu založené na korpusu]

CEP – Centrum pro ekonomiku a politiku

ČNK – Český národní korpus

DIN – difference index [rozdílový index]

EBB – *Euro Business Breakfast*

IVK – Institut Václava Klause

i.p.m. – instances per million [výskyty na milion]

KWIC – key word in context [klíčové slovo v kontextu]

MBA – market basket analysis [analýza nákupního košíku]

MLDPA – multi-level discourse prominence analysis [víceúrovňová analýza diskurzní prominence]

NESČ – *Nový encyklopedický slovník češtiny*

NHM – projekt *The Needle-in-a-Haystack Method*

PSP – Poslanecká sněmovna Parlamentu České republiky

ÚČNK – Ústav Českého národního korpusu

VK – *Korpus textů Václava Klause*

Úvod

„Klausismus je neologismus odvozený od osobního jména, označující ideologii blízkou myšlenkovému světu Václava Klause,“[1]

Druhý prezident České republiky, první předseda vlády ČR, třetí předseda Poslanecké sněmovny, zakladatel a první předseda ODS, profesor ekonomie, basketbalista. Nejen tím vším byl a je Václav Klaus. Jak si totiž již více než 10 let všímají čeští publicisté, Klaus je také velmi produktivním jazykovým inovátorem, který obohatil češtinu o mnohé neologismy. V březnu 2013 vytvořil datový analytik Josef Šlerka aplikaci *Vývoj slovníku Václava Klause*[2], kterou doprovodil také stručnou analýzou v článku *Všechny Klausovy -ismy a další slovíčka odstupujícího prezidenta*,[3] a na nějž navázali Petr Honzejk, Jindřich Šídlo a Ondřej Tuček svým *Příručním slovníkem klausismů*.[4]

Právě pojem *klausismus* funguje polysémně, jednak jako označení ideologie,[5] jednak jako označení pro neologismy a okazionalismy, s nimiž Václav Klaus přišel a které jsou pro něj charakteristické.[6] Tato diplomová práce se zájmem publicistů o Klausovo jazykové chování inspiruje. V předkládaném výzkumu se však snažím k problematice přistupovat co nejempiričtěji a pokouším se o komplexní analýzu Klausova idiolektu založenou na primárně kvantitativním přístupu korpusové lingvistiky. Kromě Klausovy jazykové kreativity je tak dalším důvodem, proč byl pro analýzu idiolektu v tomto výzkumu zvolen právě Václav Klaus, skutečnost, že již několik desetiletí průběžně zveřejňuje velké množství vlastních textů.

Práce má následující strukturu. Po tomto úvodu následuje kapitola Cíle výzkumu, v níž je zformulován jeden hlavní a několik vedlejších cílů a výzkumných otázek. V kapitole

[1] Heslo *Klausismus*, in: *Wikipedie: Otevřená encyklopedie* [online]. URL: https://cs.wikipedia.org/w/index.php?title=Klausismus&oldid=23361856. Cit. 7. 5. 2024.

[2] ŠLERKA, Josef. Vývoj slovníku Václava Klause. *Hospodářské noviny* [online]. 11. 3. 2013. URL: https://hn.cz/c1-59445690-klaus-slovnik. Cit. 18. 5. 2024. V současnosti se zdá být aplikace již neaktivní.

[3] ŠLERKA, Josef. Všechny Klausovy -ismy a další slovíčka odstupujícího prezidenta. *Hospodářské noviny* [online]. 7. 3. 2013. URL: https://nazory.hn.cz/c1-59452240-vsechny-klausovy-ismy. Cit. 18. 5. 2024.

[4] HONZEJK, Petr & ŠÍDLO, Jindřich & TUČEK, Ondřej. Příruční slovník klausismů. *Hospodářské noviny* [online]. 13. 3. 2013. URL https://hn.cz/c1-59489080-prirucni-slovnik-klausismu. Cit. 18. 5. 2024.

[5] PEHE, Jiří. Homosexualismus a klausismus. *Aktuálně.cz* [online]. 6. 8. 2011. URL: https://blog.aktualne.cz/blogy/jiri-pehe.php?itemid=13895. Cit. 18. 5. 2024.

[6] HRUBEŠ, Karel. Nový klausismus - spokojování. Podle expertky je prezident kreativní. *Lidovky.cz* [online]. 19. 11. 2012. URL: https://www.lidovky.cz/domov/klausismy-podle-jazykovedcu-je-prezident-kreativni.A121118_132740_ln_domov_khu. Cit. 18. 5. 2024.

Teoretická část jsou rozebírány dva hlavní teoretické koncepty, z nichž předkládaný výzkum vychází – idiolekt a klíčová slova. Poté v kapitole Metodologie podrobně popisuji kompletní metodologický postup výzkumu, tedy sběr dat a jejich analýzu. Výsledky všech analýz jsou dále představeny a interpretovány v kapitole Praktická část. V závěru jsou předloženy vlastní závěry výzkumu, jeho limity a celou práci zakončuje stručná diskuse.

Všechny doslovné citace z cizojazyčné literatury jsou ve vlastním textu práce uvedeny v mém vlastním neodborném překladu, původní znění je vždy uvedeno v poznámce pod čarou.

1 Cíle výzkumu

1.1 Hlavní cíl výzkumu

Hlavním cílem předkládaného výzkumu je komplexní analýza idiolektu Václava Klause provedená s využitím korpusových metod a nástrojů. Tento cíl zahrnuje sledování potenciálních proměn Klausova idiolektu v různých typech jeho textů a v čase, respektive v závislosti na veřejných funkcích, které Klaus vykonával. Největší pozornost je věnována lexikální rovině jeho idiolektu, cílem tedy je zjistit, jaká ze slov, která Klaus (veřejně) užívá, jsou charakteristická právě pro jeho slovní zásobu. Na základě hlavního cíle formuluji následující hlavní výzkumnou otázku:

Jaké lexikální prostředky jsou charakteristické pro idiolekt Václava Klause?

1.2 Vedlejší cíle výzkumu

Dosažení výše uvedeného hlavního cíle je v tomto výzkumu podmíněno vytvořením korpusu textů Václava Klause. Byť shromáždění jeho textů, jejich technické zpracování včetně opatření metadaty a sestavení do elektronického korpusu je spíše prostředkem než účelem, jedná se o zásadní součást předkládaného výzkumu a zároveň o práci, jejíž produkt bude moci být potenciálně využit také v dalších výzkumech. Z těchto důvodů lze považovat vytvoření korpusu textů za první a nejdůležitější z vedlejších cílů tohoto výzkumu.

Aby byla analýza Klausova idiolektu opravdu komplexní, je potřeba věnovat alespoň částečnou pozornost také jeho jiným součástem než pouze lexikonu. Druhým vedlejším cílem výzkumu je tedy analýza určitého gramatického rysu charakteristického pro Klausův idiolekt. Na základě článku Josefa Šlerky (viz Úvod) jsem se rozhodl konkrétně pro analýzu slovesných tvarů.

Pro přidání další úrovně analýzy idiolektu je možné kromě lexikonu a gramatiky věnovat pozornost také slovotvorbě. Jako třetí vedlejší cíl výzkumu jsem tedy vytyčil analýzu určitého slovotvorného rysu charakteristického pro Klausův idiolekt. Na základě publicistických textů zabývajících se Klausovým jazykovým chováním (viz Úvod) jsem se rozhodl zaměřit se na analýzu jeho afixace, konkrétně na lexémy s prefixem *ne-* a lexémy se sufixem *-ismus*.

Vzhledem k převážně politickému zaměření Klausových textů se nabízí je podrobit analýze diskurzu. Konečně čtvrtým vedlejším cílem výzkumu tedy je korpusově založená diskurzní analýza vybraných Klausových textů. Při sběru dat jsem se rozhodl zvolit k tomuto účelu texty vzniklé pro jednu pravidelně se opakující událost (projevy na setkáních *Euro Business Breakfast*), aby bylo možné sledovat, jak se v průběhu let proměňovala Klausova slovní zásoba (přesněji klíčová slova) v konkrétní komunikační situaci. Pokusím se zjistit, zda identifikovaná klíčová slova souvisejí vždy pouze s tématem projevu, anebo vypovídají také o Klausově diskurzní praxi.

Na základě vedlejších cílů formuluji následující vedlejší výzkumné otázky:

1) **Jaké slovesné tvary jsou charakteristické pro idiolekt Václava Klause?**
2) **V jaké míře jsou pro idiolekt Václava Klause charakteristické lexémy s prefixem *ne-* a lexémy se sufixem *-ismus*?**
3) **Souvisejí klíčová slova v projevech Václava Klause na setkáních *Euro Business Breakfast* pouze s jejich tématy, anebo vypovídají také o Klausově diskurzní praxi?**

2 Teoretická část

V této kapitole prezentuji poznatky sloužící jako základní teoretická východiska předkládaného výzkumu. Tyto poznatky jsou založeny na studiu odborné literatury týkající se především (ale nejen) korpusové lingvistiky, lexikologie a analýzy diskurzu. Teoretická část je členěna na dvě podkapitoly, z nichž každá se soustředí na jeden koncept stěžejní pro korpusovou analýzu idiolektu, tedy pro dosažení hlavního cíle výzkumu. Nejsou zde tudíž podrobně rozebírána všechna teoretická východiska týkající se také vedlejších cílů výzkumu, s nimi související základní poznatky jsou však uváděny na patřičných místech kapitol *Metodologie výzkumu* a *Praktická část*.

Nejprve se zabývám idiolektem, konkrétně původem pojmu, jeho pozicí v některých lingvistických disciplínách, vývojem chápání konceptu individuálního jazyka a samozřejmě různými definicemi tohoto pojmu. Podrobněji se věnuji vztahu idiolektu s korpusovou lingvistikou v oddílu 2.1.1, kde popisuji možnosti využití korpusové metodologie pro analýzu individuálního jazyka a představuji několik v tomto směru průkopnických a inspirativních výzkumů.

Ve druhé podkapitole se zabývám klíčovými slovy. Nejdříve zdůvodňuji, proč se v předkládaném výzkumu zaměřuji především na lexikální rovinu jazyka, a poté vysvětluji, proč pro nalezení „individualit“ ve slovní zásobě jedince nestačí pouze informace o frekvenci jednotlivých slov. V návaznosti na to poukazuji na tři odlišné koncepty klíčových slov a jejich definice. V oddílu 2.2.1 se věnuji korpusové metodě analýzy klíčových slov, kde na příkladu vybraných výzkumů ukazuji, k jakým výzkumným cílům se dá metoda využít, a popisuji, které nástroje ČNK tuto analýzu umožňují. Teoretickou část uzavírám oddílem 2.2.2, kde představuji několik výzkumů užívajících metodu analýzy klíčových slov na textech politického diskurzu.

2.1 Idiolekt

Lingvista usilující o obecně platný popis jazyka založený na generalizaci poznatků získaných analýzou autentického jazykového materiálu musí mít při své snaze na paměti jednu nezanedbatelnou skutečnost — za každým přirozeným jazykovým projevem stojí konkrétní člověk se svým vlastním idiolektem, tedy „soubor[em] jazykových vyjadřovacích prostředků vlastních jednotlivci“ (Krčmová 2017). Již v roce 1960 zaznamenal americký

lingvista Einar Haugen v článku *From idiolect to language* následující poznatek: „Idiolekty jsou jediným druhem jazyka, z kterého můžeme shromažďovat data. Jako lingvisty nás ale nezajímá ničí idiolekt jinak než jako materiál pro zobecnění týkající se jazyka. Taková zobecnění se odvozují vhodným vzorkováním idiolektů“[7] (Haugen 1972, s. 415).

Uživatel jazyka sice může míru vlivu idiolektu na své vyjadřování přizpůsobovat komunikační situaci (modu, stylu, žánru atd.), avšak je fakticky nemožné, aby jej zcela potlačil (Cvrček et al. 2020, s. 139n.).[8] Tato všudypřítomnost idiolektů tak na jednu stranu vyvolává množství teoretických otázek i praktických problémů pro obecný popis jazyka, na druhou stranu však idiolekt představuje svébytnou oblast výzkumu, která má své místo v oborech jak teoretické, tak aplikované lingvistiky. Idiolekt někdy může být objektem zkoumání sám o sobě (např. při popisu variet v sociolingvistice), jindy slouží spíše jako instrument potřebný k dosažení specifických cílů (např. při identifikaci mluvčího či pisatele ve forenzní lingvistice). Pro empiricky podloženou analýzu idiolektu konkrétního člověka se pak přímo nabízí korpusová lingvistika, čemuž je věnována pozornost v kap. 2.1.1.

Význam lexému *idiolekt* nastiňuje již jeho etymologie — morfém *idio-* pochází z řeckého *ídios* znamenajícího ‚vlastní, osobní, zvláštní‘ (heslo *idio-*, in: Rejzek 2015, s. 257); morfém *-lekt* pochází z řeckého *légō* znamenajícího ‚mluvím, čtu, počítám‘ a odkazuje na jazykovou varietu (heslo *dialekt*, in: Rejzek 2015, s. 142).[9] Jakožto lingvistický termín byl pojem idiolekt použit poprvé (dle Hazen 2006, s. 513 i Wright 2018, s. [2]) v roce 1948 v článku amerického lingvisty Bernarda Blocha *A set of postulates for phonemic analysis*, kde uvedl následující definici: „Souhrn možných promluv jednoho mluvčího v jednom okamžiku při používání jazyka k interakci s jiným mluvčím je *idiolekt*“[10] (Bloch 1948, s. 7; zvýraznění v originále). Od té doby byl idiolekt zkoumán především v rámci sociolingvistiky (která se jako samostatná disciplína formovala zhruba od poloviny 20. století), v jejímž systému jazykových variet (národní jazyk, dialekt,

[7] „[I]diolects are the only kind of language we can collect data on. But as linguists we are not interested in anybody’s idiolect except as the material for generalization concerning language. Such generalizations are derived by a suitable sampling of idiolects.“

[8] Specifickými případy jsou plně formalizované a institucionalizované jazykové projevy užívané např. při obřadech, ve formulářích, v zákonech apod. Takové jazykové projevy jsou však ze své podstaty odosobněné, a tudíž by měly být vůči působení idiolektu imunní.

[9] Někdy se v souvislosti s osobitostí jazyka jedince používá též přívlastek *idiosynkratický*, a to zejména ve forenzní lingvistice (Coulthard & Johnson & Wright 2017, s. 15).

[10] „The totality of the possible utterances of one speaker at one time in using language to interact with one other speaker is an *idiolect*.“

sociolekt atd.) představuje co do počtu mluvčích nejmenší jednotku (Čermák 2011, s. 17, 45; Nekvapil 2017).

Koncept individuálního jazyka (či alespoň výsostně individuálních rysů jazyka jednotlivce) je však v dějinách myšlení ještě starší a zabývali se jím nejen lingvisté. Kupříkladu anglický spisovatel Samuel Taylor Coleridge v autobiografii *Biographia Literaria* původně vydané roku 1817 zaznamenal následující myšlenku: „Jazyk každého člověka má za prvé, své individuality; za druhé, společné vlastnosti třídy, do které patří; a za třetí, slova a fráze univerzálního užití“[11] (Coleridge 1884, s. 170). Nejenže si tedy Coleridge uvědomoval existenci individualit (vlastností a prvků) jazyka jednotlivce, ale dokonce pro něj představovaly jeho primární složku.

Ideu existence odlišných „jazyků“, z nichž každý je vlastní vždy pouze jednomu konkrétnímu člověku, zastával např. francouzský filosof Jacques Derrida, který v interview z roku 1983 titulovaném *Derrida l'insoumis*[12] vyjádřil myšlenku, že všichni jsme tlumočníky a překladateli, protože neexistuje jazyk všech (Derrida 1995, s. 116). Jinými slovy: žádní dva lidé na světě neužívají zcela totožný jazyk, byť by oba mluvili např. francouzsky, a proto si při jakékoliv komunikaci musí každý člověk „jazyky“ druhých vždy tlumočit nebo překládat do svého vlastního.

Na základě každodenních zkušeností s komunikací tak mohou skutečnost, že jazyk každého člověka je jedinečný, vnímat i laičtí uživatelé jazyka. Zároveň však tento zdánlivě všední jev implikuje vědecké otázky, jež si kladou jak filosofové jazyka, tak lingvisté. Jednou z prvních odborných prací, v níž byla problematika jazykové individuálnosti reflektována, je kniha *Prinzipien der Sprachgeschichte*[13] německého lingvisty Hermanna Paula, původně vydaná roku 1880. Paul si uvědomoval, že každý jazykový projev je vždy dílem jediného člověka (že vzniká v mysli jedince), a proto prosazoval, aby středem lingvistického zkoumání byl právě jedinec a jeho mysl (Paul 1891, s. xliii; Wright 2018, s. [2n.]).

Od konce 19. století věnovali vědci stále větší pozornost otázkám týkajícím se původu a povahy jazyka — především zda je jazyk fenomén vrozený, či osvojený, případně

[11] „Every man's language has, first, its individualities; secondly, the common properties of the class to which he belongs; and thirdly, words and phrases of universal use.“

[12] V citovaném anglickém vydání nazváno *Unsealing ("the old new language")*.

[13] V citovaném anglickém překladu nazvaná *Principles of the History of Language*.

zda se jedná o kombinaci obojího. Do této debaty, která se od 50. let 20. století stala jedním z nejvýraznějších témat obecné lingvistiky, přispěli v uplynulých desetiletích představitelé různých vědeckých tradic a směrů, kteří své pohledy na problematiku formulovali v rámci odlišných disciplín i metod. Postupně se tak utvořilo mnoho rozličných pojmů týkajících se výše zmíněné opozice (vrozený versus osvojený), mezi nimiž se také vyskytuje právě protiklad idiolektální a neidiolektální perspektivy. Proponenti idiolektální perspektivy zastávají názor, že jazyk jedince je ontologicky primární, tedy že předchází tzv. sociálním jazykům, jež vznikají až překryvem jednotlivých idiolektů; nejvýraznějším představitelem idiolektální perspektivy je americký lingvista Noam Chomsky se svým konceptem I-jazyka (I = interní). Naproti tomu proponenti neidiolektální perspektivy tvrdí, že ontologicky primární jsou sociální jazyky (v Chomského koncepci označovány jako E-jazyky; E = externí), a teprve na jejich základě si jednotliví mluvčí formují vlastní jazykové chování; jedním z nejvýznamnějších představitelů neidiolektální perspektivy byl americký filosof David Kellogg Lewis se svým konceptem jazyka jako společenské konvence (Barber & Garcia Ramirez 2021).

Od publikování výše zmíněného Blochova článku se v odborné literatuře objevovaly různé definice idiolektu, které lze rozdělit na dvě skupiny analogické Saussurově dichotomii *langue* a *parole* — jedna skupina idiolekt pojímá jako jazykový systém dostupný jedinci, zatímco druhá skupina idiolektem označuje realizovanou jazykovou produkci jedince (Wright 2018, s. [2]). V tomto duchu uvádí dvě kategorie významu idiolektu též Hazen: „[1.] Souhrn jazykových projevů jedné osoby, včetně všech možných promluv. [2.] Jazyková produkce jedné osoby (tj. pouze to, co tato osoba řekne, nikoliv vnitřní znalost v mysli). Někteří badatelé zde kladou důraz na konstelaci vzorců jazykových variací, které odlišují jedince od ostatních mluvčích téhož dialektu“[14] (Hazen 2006, s. 512). V korpusové lingvistice je stěžejní vždy druhé z těchto pojetí idiolektu, neboť pouze takto vymezený idiolekt lze zkoumat empiricky.

Jak je termín idiolekt definován v současné bohemistické lingvistice? František Čermák jej definuje jako „jazykový systém jednotlivce představující vždy individuální výběr z (národního) jazyka a směs rysů lokálních, sociálních aj.“ (Čermák 2011, s. 280). Marie Krčmová v *NESČ* uvádí, že idiolekt označuje „soubor jazykových vyjadřovacích prostředků

[14] „[1.] The sum total of language of one person, including all possible utterances. [2.] The linguistic output of one person (i.e., only what that person says and not the internal knowledge in the mind). For some scholars, emphasis here is on the constellation of language variation patterns which distinguish an individual from other speakers of the same dialect.“

vlastních jednotlivci,“ že jeho základem je „mateřský jaz. v původním smyslu slova, tj. ta přirozenou cestou osvojená podoba národního jazyka, které se užívá v běžné denní komunikaci v rodinném prostředí,“ na niž se „navrstvují další variety nár. jaz., včetně jaz. spis., ale také sociolekty n. další (cizí) jaz.“ Dále je v hesle uvedeno, že z idiolektu „vyrůstá jak individuální styl jedince, tak i jeho jaz. povědomí, které je základem postojů jazykově hodnotících,“ a že „bývá obvykle chápán jako individuální jaz. kompetence ve smyslu schopnosti aktivního užívání jaz. a jako takový může být i popisován na základě analýzy velkého souboru textů, jež jednotlivec vytváří,“ ale v širším významu k němu „náleží i pasivní znalost jazyka“ (Krčmová 2017). Právě skutečnost, že idiolekt může být „popisován na základě analýzy velkého souboru textů,“ implikuje velký potenciál korpusové lingvistiky pro jeho studium.

Širší komentovaný přehled literatury věnující se idiolektu v různých perspektivách podává britský lingvista David Wright v hesle *Idiolect* v publikaci *Oxford Bibliographies in Linguistics* (Wright 2018).

2.1.1 Idiolekt v korpusové lingvistice

Aby mohl být objektem či instrumentem jakéhokoliv výzkumu idiolekt konkrétního člověka (nikoliv idiolekt jako poněkud abstraktní mentální systém figurující v debatě o původu a povaze jazyka), je v první řadě potřeba shromáždit a zpracovat jazykovou produkci daného jedince, ideálně v co největším možném rozsahu. V důsledku rozmachu korpusové lingvistiky (a také díky digitalizaci čím dál většího množství textů již existujících i vznikajících), která umožňuje shromažďovat a zpracovávat velké soubory jazykových dat výrazně rychleji a jednodušeji než dříve, lze od počátku 21. století sledovat v lingvistické literatuře rostoucí zájem o empirickou analýzu idiolektu využívající právě korpusová data a metody (Wright 2018, s. [13n.]).

Práce, které Wright uvádí, mohou být považovány za průkopnické, neboť jejich autoři často museli koncipovat nové metodologické postupy pro aplikaci korpusových metod na analýzu idiolektu (Wright 2018, s. [14–17]). Stejně jako u jiných druhů korpusových výzkumů však i pro analýzu idiolektu platí následující univerzální poučka: „Obecně je třeba tedy mít napřed data zjištěná a uchopená, změřená **kvantitativně** (mj. už v podobě konkordance), a až pak, jejich detailním studiem, které začíná zobecněním z jednotlivostí, lze dospět k závěrům **kvalitativním**, tj. především sémantickým, funkčním nebo

pragmatickým. I když je výzkum *kvalitativní* možný jen zprostředkovaně přes formu (od které se přirozeně začíná vždy), není přesto nemožný obecně, sám o sobě, ale závisí už na invenci uživatele, jak dokáže získané formální výsledky interpretovat kvalitativně, tj. sémanticky a jinak […], jaké distinkce a kritéria v datech najde apod.“ (Čermák 2017, s. 93; zvýraznění v originále).

V duchu této poučky je tak právě kvantitativní analýza krokem prvním a základním pro zjištění toho, co konstituuje idiolekt jedince, nikoliv však krokem jediným postačujícím. Kvantitativní analýzou sice lze mimo jiné zjistit, jak často se v jazykové produkci jedince vyskytuje konkrétní prvek (lexém, tvar, kolokace atd.), avšak frekvence sama o sobě neposkytne informaci, nakolik je tento prvek součástí idiolektu jedince. Dokonce ani kvantitativní metoda analýzy klíčových slov, která k údaji o frekvenci prvku doplní informaci o jeho prominenci ve srovnání s referenčním korpusem,[15] není v tomto ohledu zcela dostatečná (viz kap. 2.2.1). Pro komplexní analýzu idiolektu musí lingvista na kvantitativní metody navázat metodami a interpretacemi kvalitativními, přičemž využívá nejen poznatky a metody dalších disciplín, ale do určité míry také svou jazykovou intuici. Korpusová analýza idiolektu by tedy měla být především výzkumem typu corpus-based (Čermák 2017, s. 100; Tognini-Bonelli 2001, s. 65). Jako příklady výstupů korpusového zpracování idiolektu v českém prostředí lze uvést *Slovník Karla Čapka* (Čermák 2007) a *Slovník Bohumila Hrabala* (Čermák & Cvrček 2009).

Všechny analýzy, jejichž výsledky jsou prezentovány v praktické části předkládané práce, byly provedeny na základě metodologie, jež je podrobně popsána v kap. 3. Uplatněná metodologie samozřejmě nevznikla ve vzduchoprázdnu, nýbrž byla do jisté míry inspirována některými již existujícími výzkumy idiolektů (či konkrétních idiolektálních rysů) využívajícími korpusová data a metody. V následujících odstavcích je stručně představena jedna diplomová práce a tři odborné články, které posloužily jako zdroje teoreticko-metodologické inspirace pro můj výzkum.

Nejpřínosnější se ukázala být diplomová práce Marka Leška z roku 2012 nazvaná *The individual textual profile: a corpus-based study of idiolect*, která je tematicky velmi

[15] Termín *referenční korpus* se v korpusové lingvistice obvykle užívá ve významu ‚entita sloužící jako standard pro porovnání frekvencí jevů‘. Často jde tedy o korpus reprezentující běžný úzus (heslo *Referenční korpus*, in: *Příručka ČNK*), pro analýzu klíčových slov se však může jednat též o korpus reprezentující úzus žánrově či tematicky blízký zkoumanému korpusu (Culpeper 2009, s. 35). V prostředí ČNK se kromě toho termín *referenční korpus* užívá také ve významu ‚entita, která je zpětně dostupná‘, čímž je myšlen korpus, který se od doby svého publikování nemění (heslo *Referenční korpus*, in: *Příručka ČNK*). Pokud není explicitně uvedeno jinak, užívá se v této práci pojem *referenční korpus* v prvním z těchto významů.

blízká zde předkládané práci, díky čemuž poskytla jak užitečný vzor pro metodologický rámec, tak přehled o základní literatuře (v současnosti již ne zcela nejnovější, ale stále dostatečně aktuální) týkající se idiolektu a jeho korpusové analýzy. Leško ve svém výzkumu analyzoval idiolekt Baracka Obamy, přesněji výsek jeho idiolektu jakožto amerického prezidentského kandidáta ve specifické komunikační situaci oficiálních předvolebních debat v roce 2008 (Leško 2012, s. 9). Primární metodou analýzy idiolektu byla analýza klíčových slov, přičemž Leško srovnával zkoumaný Obamův korpus obsahující všechny jeho řečové projevy ze tří televizních debat s referenčním korpusem obsahujícím všechny řečové projevy ostatních prezidentských kandidátů z dvanácti předvolebních debat v letech 2000, 2004 a 2008 (ibid., s. 22n.). Užití takto úzce vymezených specializovaných korpusů má své výhody i nevýhody. Hlavní výhodou je především to, že texty v obou korpusech vznikly ve stejné komunikační situaci, díky čemuž nemohly být rozdíly mezi jazykovými projevy mluvčích způsobeny rozdíly v registru, žánru apod., ale lze se oprávněně domnívat, že byly především idiolektální povahy (ibid., s. 18). Hlavní nevýhodou je relativně malý rozsah zkoumaného i referenčního korpusu (zde konkrétně 22 013 a 145 266 pozic), což s sebou nese jednak obecně menší reprezentativnost, jednak může způsobovat nežádoucí statistické anomálie (ibid., s. 18, 67). Výsledkem analýz byla identifikace 37 klíčových slov, které byly rozděleny do tří kategorií — vlastní jména, lexikální klíčová slova a gramatická klíčová slova (ibid., s. 12, 25n.). Přestože klíčová slova rozdělená do kategorií poskytla sama o sobě užitečné informace o Obamově idiolektu, Leško je dále zkoumal kvalitativně v jejich kontextu, a s užitím analýz konkordančních řádků, kolokací a clusterů tak dospěl k ještě přínosnějším závěrům (ibid., s. 20–22).

Metodologicky zásadní prací je článek Sandry Mollin z roku 2009 nazvaný *"I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations*. Mollin se ve svém výzkumu idiolektu Tonyho Blaira zaměřila na kolokace, konkrétně na tzv. *maximiser collocations* (adverbium maximální míry + adjektivum / adverbium / verbum) jako na případovou studii, avšak její systematická a důsledná korpusová metodologie může být s úpravami replikovatelná pro libovolnou analýzu idiolektu či jeho dílčích rysů (Mollin 2009, s. 369). Pro účely výzkumu byl sestaven korpus textů Tonyho Blaira obsahující jeho transkribovaná prohlášení, proslovy, interview a vystoupení v parlamentu, to vše z období, kdy byl britským ministerským předsedou (ibid., s. 370). Jelikož jsou kolokace ze své podstaty víceslovné jednotky, nebyla ve výzkumu užívána metoda analýzy klíčových slov, avšak provedené statistické poměřování hodnot

relativní frekvence, MI-score a log-likelihood u týchž kolokací ve zkoumaném a referenčním korpusu (*British National Corpus*) zde fungovalo na stejném principu (ibid., s. 378–382). Mollin taktéž nezůstala u čistě kvantitativních výsledků a navázala na ně více kvalitativními metodami — podrobila „kandidátní" kolokace testům synonymní preference a registrové specifičnosti (Mollin 2009, s. 382–387). Tím získala finálních 16 kolokací, které mohly být jednoznačně označeny za tzv. blairismy, tedy prvky charakteristické pro Blairův idiolekt (ibid., s. 388n.).

Článek Michaela Barlowa z roku 2013 nazvaný *Individual differences and usage-based grammar* sleduje mírně odlišné cíle než komplexní analýzu idiolektu či idiolektálního prvku konkrétního jedince, avšak také Barlow pro svůj výzkum sestavil a zkoumal korpusy obsahující jazykovou produkci vždy pouze jednoho člověka. Zkoumané korpusy obsahovaly transkribované výstupy šesti tiskových mluvčích Bílého domu na tiskových konferencích (Barlow 2013, s. 447). Barlow v korpusech analyzoval nejčastější bigramy a trigramy (všechny gramatické či lexikogramatické povahy) a srovnával jejich relativní frekvence mezi jednotlivými tiskovými mluvčími a následně i s kolektivním korpusem veřejných výstupů jiných zaměstnanců Bílého domu (ibid., s. 449–471). Výsledky ukázaly, že v jazykové produkci jedinců vystupujících ve stejné funkci a stejné komunikační situaci se projevují jasně pozorovatelné rozdíly mezi frekvencemi určitých gramatických vzorců, které jsou zároveň u každého jedince poměrně konzistentní v průběhu času, a tudíž je lze považovat za součásti jejich idiolektů (ibid., s. 474n.).

Poslední zde představená práce je velmi krátká a nevyznačuje se natolik propracovanou a precizní metodologií jako výše uvedené výzkumy, přesto stojí alespoň za zmínku díky své neotřelé aplikaci korpusové analýzy idiolektu na literární text. Vannessa L. Milom ve svém článku z roku 2022 nazvaném *Corpus Linguistic Analysis of the Idiolects of Gollum and Sméagol* prezentuje výzkum, v němž analyzovala všechny promluvy jedné bytosti (avšak dvou postav) Gluma-Sméagola v knize *Hobit* a v trilogii *Pán prstenů* od J. R. R. Tolkiena. Cílem bylo identifikovat, jakými rysy Tolkien rozlišil idiolekty Gluma a Sméagola (Milom 2022, s. 2). První užitou metodou byla kvantitativní analýza klíčových slov, přičemž zkoumané korpusy sestávaly ze všech Glumových a Sméagolových promluv, zatímco referenční korpus sestával z úplného textu všech čtyř knih. Výsledky této analýzy ukázaly, která slova jsou charakteristická pro kterou ze dvou postav a že nejvýrazněji se jejich slovní zásoba liší v užívání pronomin (ibid., s. 2n.). Další užitou

metodou byla kvalitativní analýza konkordančních řádků, při níž se hlavní rozdíly mezi idiolekty obou postav ukázaly být v reduplikaci sykavek a konjugaci verb (ibid., s. 3n.).

2.2 Klíčová slova

Jak bylo nastíněno v kap. 2.1, idiolekt je úplným souborem vyjadřovacích prostředků jedince, a proto se na jeho celkové podobě podílejí prvky všech jazykových rovin. Při analýze idiolektu se tedy lze zaměřit na rysy fonetické, morfologické, lexikální, syntaktické i stylotvorné. Každá z těchto možností vyžaduje jinou metodologii a může posloužit k dosažení odlišných výzkumných cílů.

Předkládaná práce se při analýze idiolektu zaměřuje především na rovinu lexikální, a to z několika důvodů. Za prvé, lexikon je co do počtu jednotek zdaleka největší rovinou každého přirozeného jazyka — tvoří 99,9 % všech formálních jazykových jednotek (Čermák 2010, s. 11, 319) — a logicky by tedy počtem jednotek měl představovat také největší část idiolektu jedince. Za druhé, lexikon je díky jasně diskrétní povaze svých formálních jednotek, *grafických slov*, rovinou nejsnadněji formálně zpracovatelnou a kvantitativně analyzovatelnou, což je pro korpusovou analýzu idiolektu nejen výhodné, ale přímo zásadní (ibid., s. 17–22). Konečně za třetí, funkční jednotky lexikonu, *lexémy*, jsou prvotními nositeli významu, a jak připomíná Čermák, „[v]ýznam a potřeba jeho přenosu je zdaleka nejdůležitějším důvodem, proč jazyk vznikl a proč lidé spolu komunikují: při komunikaci si sdělují především významy“ (ibid., s. 39). Předkládaná analýza idiolektu Václava Klause se proto soustředí primárně na jeho slovní zásobu (tedy kvantitativně míněný úhrn lexémů; ibid., s. 16), neboť právě lexikální rovina poskytuje nejvíce informací o tom, jaké významy Klaus sděluje a jakými prostředky tak činí.

Hlavní cíl výzkumu již byl zformulován v kap. 1.1, pro potřeby této podkapitoly jej zde připomínám v upravené podobě: Hlavním cílem analýz v předkládaném výzkumu je zjistit, jaká slova (lexémy) jsou charakteristická pro zkoumaný idiolekt daného jedince. Tohoto cíle lze s využitím korpusové lingvistiky dosáhnout různými způsoby lišícími se mírou komplexnosti použité metodologie. Intuitivně se jako nejjednodušší řešení nabízí vytvoření frekvenčního seznamu slov, což by v tomto případě mohly být *wordy* (bez ohledu na lemmatizaci zkoumaného korpusu) nebo *lemmata* (pouze pokud je zkoumaný korpus lemmatizován). Takovýto frekvenční seznam obsahuje všechna slova vyskytující se ve zkoumaném korpusu, a navíc poskytuje údaje o jejich absolutních i relativních frekvencích.

Obecně však platí, že seznamy slov jsou užitečné spíše jako „východisko pro různé další operace," protože „statistický výsledek sám o sobě žádným výsledkem ještě není; vyžaduje vhodnou interpretaci" (Čermák 2017, s. 101). Tato skutečnost se potvrzuje již při letmém pohledu na špičku frekvenčního seznamu slov libovolného autorského korpusu, z něhož je jasné, že nepodává příliš relevantní informace o tom, jaká z uvedených slov jsou pro idiolekt daného autora charakteristická. Důvodem je především to, že nejfrekventovanějšími slovy zde jsou ta, která podle výše uvedeného Coleridgeova citátu nelze označit za „individuality", nýbrž za slova „univerzálního užití" (viz kap. 2.1) — perspektivou korpusové lingvistiky tedy taková slova, která patří mezi nejfrekventovanější také v mnoha jiných korpusech (a tím pádem i v jazyce obecně). Tento jev je odrazem obecně platných statistických zákonitostí v lexikonu, především prvního Zipfova zákona (jenž se však uplatňuje nejen v lexikonu, a dokonce nejen v lingvistice), což je empirická pravidelnost (spíše než zákon v užším smyslu) vyjadřující, „že frekvence slova v korpusu je nepřímo úměrná jeho ranku ve frekvenční tabulce" (Čermák 2010, s. 239n.; Zipf 1949, s. 23–26). Frekvenční slovníky budované na základě korpusů dobře ilustrují skutečnost, že „frekvence slova je inherentní vlastností každého lexému," a tyto slovníky dokáží kvantitativně vyjádřit lexikální pokrytí textů konkrétního jazyka — pro češtinu tak platí, že 10 nejfrekventovanějších lexémů (což jsou slova výlučně gramatická) pokrývá cca 20 % všech slov v libovolném textu (Čermák 2010, s. 244n.).

Frekvenční seznam slov, respektive absolutní a relativní frekvence slov ve zkoumaném korpusu tedy nejsou pro určení charakteristických lexémů idiolektu samy o sobě dostatečnými měřítky. Z poznatků uvedených v předchozím odstavci lze vyvodit, že zkoumaný korpus je potřeba nějakým způsobem srovnat s jiným korpusem, případně s vícero korpusy, aby bylo zřejmé, jaké lexémy vyskytující se ve zkoumaném korpusu lze označit za charakteristické pro konkrétní idiolekt. Nejvhodnějším způsobem takového srovnání je kvantitativní metoda analýzy klíčových slov (Scott & Tribble 2006, s. 58).

Termín *klíčové slovo*[16] se napříč různými vědami užívá v rozličných významech, a dokonce i v rámci korpusové lingvistiky se s ním pracuje ve dvou významech (viz níže).

[16] V angličtině se podle oboru či kontextu užívají zápisy *key word* i *keyword*. Také v korpusové lingvistice se užívají oba zápisy, *key word* obvykle ve smyslu KWIC a *keyword* obvykle ve smyslu jednotky, jejíž frekvence je ve zkoumaném textu výrazně vyšší, než lze očekávat na základě její frekvence v referenčním korpusu (viz níže). Jak lze však vidět i na zde citovaných pracích, nejedná se o striktní rozlišení, kterého by se drželi všichni autoři.

Britský lingvista Michael Stubbs ve své studii *Three concepts of keywords* (ve sborníku *Keyness in Texts*) rozlišuje a popisuje tři různé koncepty klíčových slov. První koncept vychází z kulturních studií, především z prací Johna Ruperta Firtha a Raymonda Williamse, a také z kulturně orientované lingvistiky zastoupené např. v pracích Stubbse a Anny Wierzbické, která klíčová slova definovala jako „slova, která jsou v dané kultuře obzvláště důležitá a vypovídající“[17] a označila je za „ústřední body, kolem nichž se organizují celé kulturní oblasti“[18] (Wierzbicka 1997, s. 15n.; Stubbs 2010, s. 23–25). Druhý koncept vychází z korpusové lingvistiky, konkrétně z prací Mikea Scotta a Christophera Tribblea, váže se k výskytům v textech (oproti abstraktnějšímu kulturnímu pojetí tedy vždy závisí na konkrétních realizacích v textech) a má statistickou povahu — v tomto konceptu jsou za klíčová slova považována ta, „která se ve vzorku textu vyskytují výrazně častěji, než by se očekávalo vzhledem k jejich četnosti ve velkém obecném referenčním korpusu“[19] (Stubbs 2010, s. 25–28). Konečně Stubbsem zmiňovaný třetí koncept vychází z prací Gill Francis, která v korpusech analyzovala, jak se pomocí lexikogramatických vzorců v určitých clusterech (frázích a schématech) vyjadřují kulturně signifikantní jednotky významu (ibid., s. 28–32). Pro analýzu idiolektu vycházející ze zkoumání velkého vzorku textů (korpusu) konkrétního jedince je stěžejní druhý z uvedených tří konceptů, a právě s ním pracuji v předkládaném výzkumu.

Jak bylo zmíněno výše, také v korpusové lingvistice se termín *klíčové slovo* užívá ve dvou odlišných významech. Autoři anglického *A Glossary of Corpus Linguistics* definují první význam následovně: „Slovo, které se v textu nebo korpusu objevuje statisticky signifikantně častěji, než by se při srovnání s větším nebo stejně velkým korpusem dalo očekávat náhodně. Obvykle se k porovnání dvou seznamů slov za účelem určení klíčových slov používají testy log-likelihood nebo chí-kvadrát. […] Nejběžnějšími klíčovými slovy bývají (1) propria; (2) gramatická slova, která jsou často indikátory určitého stylistického profilu; (3) lexikální slova, která indikují, o čem text je“[20] (heslo *keyword*, in: Baker &

[17] „‘Key words’ are words which are particularly important and revealing in a given culture.“

[18] „[…] some words can be studied as focal points around which entire cultural domains are organized.“

[19] „Sense 2 is statistical: keywords are words which are significantly more frequent in a sample of text than would be expected, given their frequency in a large general reference corpus.“

[20] „A word which appears in a text or corpus statistically significantly more frequently than would be expected by chance when compared to a corpus which is larger or of equal size. Usually log-likelihood or chi-squared tests are used to compare two word lists in order to derive keywords. […] Commonly found keywords include (1) proper nouns; (2) grammatical words that are often indicators of a particular stylistic profile; (3) lexical words that give an indication of the ‘aboutness’ of a text.“

Hardie & McEnery 2006, s. 97n.). Z českého prostředí lze k tomuto významu uvést např. Čermákovu definici: „slovo, které je, obv. spolu s několika dalšími, v textu frekventovanější než jiná tak, že je to statisticky signifikantní, obv. pro náznak tematiky textu (‚o čem text je')" (Čermák 2017, s. 263n.). Tento význam tedy odpovídá výše uvedenému korpusově-statistickému konceptu klíčových slov podle Scotta a Tribblea. Druhý význam termínu *klíčové slovo* je ve zmiňovaném *A Glossary of Corpus Linguistics* definován takto: „Slovo, které je předmětem konkordance"[21] (heslo *keyword*, in: Baker & Hardie & McEnery 2006, s. 98).[22] Pro české prostředí lze najít definici tohoto významu např. v online *Příručce ČNK*: „Vyhledaný výraz (slovo, tvar, jeho část nebo kombinace slov) v rámci konkordance" (heslo *Klíčové slovo (keyword)*, in: *Příručka ČNK*). Pro tento význam se užívá též označení *KWIC*, což je zkratka pro *key word in context* neboli ‚klíčové slovo v kontextu' (heslo *KWIC*, in: *Příručka ČNK*). V této práci se pojem *klíčové slovo* užívá v prvním z uvedených významů (pokud není explicitně uvedeno jinak).

2.2.1 Analýza klíčových slov

Již samotné definice termínu *klíčové slovo* uvedené v kap. 2.2 zhruba naznačují, jakým způsobem lze klíčová slova ve zkoumaném textu či korpusu identifikovat. Konkrétní metodologické kroky postupu i užívané statistické míry se však od prvního uvedení metody analýzy klíčových slov v průběhu času mírně proměňovaly, a navíc korpusoví lingvisté vyvíjeli (a stále vyvíjejí) softwarové nástroje, které využití této metody usnadnily a pomohly ji více rozšířit mezi badatele. Z těchto důvodů tak lze v literatuře narazit jednak na výzkumy, které jako použitou metodu jmenovitě uvádějí analýzu klíčových slov, ale v konkrétních postupech se více či méně liší (viz níže), a jednak na výzkumy, které sice pojem *klíčové slovo* vůbec neužívají, ale jejichž metodologické postupy analýze klíčových slov prakticky odpovídají (viz např. Mollin 2009 a Barlow 2013 zmiňované v kap. 2.1.1).

V kap. 3.2 je konkrétně popsán vlastní metodologický postup, kterého jsem se ve svém výzkumu držel při analýzách klíčových slov ve zkoumaných korpusech, přičemž tento

[21] „A word which is made the subject of a concordance."

[22] Tento význam je v citovaném hesle ve skutečnosti uveden až jako třetí. Druhý uvedený význam — definován jako „jakékoliv slovo, které je v textu považováno za ‚ústřední', ale ne na základě statistického měření" („[a]ny word that is considered 'focal' in a text, but not through statistical measures") — však odpovídá Stubbsem zmiňovanému kulturnímu konceptu klíčových slov, a není tedy pro korpusovou lingvistiku relevantní.

postup vychází převážně z praxe užívané v ÚČNK a pracuje se softwarovými nástroji a daty přístupnými právě na webových stránkách ČNK. Tyto nástroje jsou popsány dále, ještě předtím jsou v následujících odstavcích představeny některé vybrané práce související s analýzou klíčových slov — odborný článek, v němž byla metoda poprvé exaktně popsána, několik teoreticky zaměřených prací a dva příklady výzkumů užívajících analýzu klíčových slov v souvislosti s popisem individuálního jazyka, tedy idiolektu či jeho výseku.

První náznaky statistického pojetí klíčových slov (viz kap. 2.2) se objevovaly od 50. let 20. století v rámci statistické stylistiky, více se však do povědomí odborné veřejnosti toto pojetí dostalo až s rozvojem modernějších korpusových programů od druhé poloviny 90. let 20. století (Culpeper & Demmen 2015, s. 91–94). Poprvé byla kvantitativní metoda analýzy klíčových slov v podobě, kterou známe dnes, exaktně použita ve výzkumu britského lingvisty Mikea Scotta, autora korpusového softwaru *WordSmith Tools*, který ji v roce 1997 popsal v článku *PC analysis of key words — And key key words*.[23] Scott zde nepřímo navázal na Stubbse, jenž ve svém korpusovém výzkumu vyšel z Firthova a Williamsova kulturního pojetí klíčových slov (viz kap. 2.2), kterýžto koncept empiricky podložil kvantitativní analýzou kolokací (Stubbs 1996, s. 165–181; Scott 1997, s. 235). Oproti Stubbsovi však Scott ve svých analýzách zvolil opačný směr — nevycházel od předem vybraných konkrétních slov, nýbrž od celých textů, v nichž zjišťoval, která slova jsou klíčová. Takovýto postup není závislý na tom, o jakých jednotkách se badatel předem domnívá, že jsou významné, ale naopak umožní odhalit takové, které jsou opravdu významné (statisticky) nezávisle na badatelových předpokladech (Culpeper & Demmen 2015, s. 90). Tento postup také vyžaduje rozšířené pojetí kontextu — Scott na rozdíl od Stubbse neomezoval kontext na bezprostřední okolí slova (ve smyslu KWIC), ale za kontext každého slova považoval celý text (ve smyslu dokumentu), v němž se dané slovo vyskytuje. Termín *klíčové slovo* zde definoval jako „slovo, které se v daném textu vyskytuje s neobvyklou frekvencí,“[24] což neznamená jednoduše frekvenci vysokou, ale neobvyklou ve srovnání s jiným, tedy referenčním korpusem (Scott 1997., s. 234–236).[25] Vlastní průběh analýzy pak rozdělil do čtyř kroků (které však provádí software automaticky). Za prvé, vytvoří se frekvenční seznam

[23] Jeden ze tří hlavních nástrojů softwaru *WordSmith Tools*, nazvaný *KeyWords*, slouží právě k automatizované analýze klíčových slov a vytváření jejich seznamů. V uvedeném výzkumu byla použita první verze tohoto nástroje.

[24] „A key word may be defined as a *word which occurs with unusual frequency in a given text*“ (zvýraznění v originále).

[25] Slova, jejichž frekvence je ve zkoumaném textu při srovnání s referenčním korpusem neobvykle nízká, označuje Scott jako *negativní klíčová slova* (Scott 1997; s. 243).

všech slov (v tomto případě wordů) v referenčním korpusu. Za druhé, vytvoří se frekvenční seznam všech slov v textu, jehož klíčová slova chce badatel identifikovat. Za třetí, frekvence každého slova v seznamu slov zkoumaného textu se porovná s frekvencí téhož slova v seznamu slov referenčního korpusu. „Pokud jsou procenta [relativní frekvence][26] podobná, může být položka ignorována. V případě velkého rozdílu ve frekvenci je však možné položku označit za klíčovou. Vlastní výpočet ‚klíčovosti' se provádí pomocí statistiky chí-kvadrát, ale je důležité pochopit, že základem je pojem výjimečnosti. Jinými slovy, pokud se slovo v našem textu vyskytuje mimořádně často, bude klíčové."[27] Konečně za čtvrté, když jsou identifikovány všechny potenciálně klíčové položky, jsou seřazeny podle hodnot své relativní klíčovosti (Scott 1997, s. 236). Poté Scott metodu ilustroval na příkladu vybraného článku z *Guardianu*, který srovnal s referenčním korpusem složeným z textů *Guardianu* z let 1992–1994. Výsledky prezentoval v tabulce s klíčovými slovy článku seřazenými podle jejich relativní klíčovosti a s uvedenými hodnotami jejich absolutních i relativních frekvencí ve zkoumaném textu i referenčním korpusu. Na závěr popisu metody zdůraznil potřebu statistických prahů pro určení signifikance, a to jednak pro minimální hodnotu výsledků chí-kvadrát testu, jednak pro minimální absolutní frekvenci výskytů (ibid., s. 237). Ve druhé části článku představil další dva koncepty související s klíčovými slovy — tzv. *klíčová klíčová slov* (*key key words*), tedy „slova, která jsou klíčová ve velkém množství textů určitého druhu,"[28] a tzv. *společníky klíčových slov* (*associates of key words*), tedy „slova, která jsou klíčová ve stejných textech jako dané klíčové klíčové slovo"[29] (ibid., s. 237n.) — i metody jejich identifikace. Tyto koncepty mají potenciál zejména v korpusově založené kritické analýze diskurzu, např. při odhalování předsudečnosti v žurnalistických textech (ibid., s. 243), avšak pro analýzu lexikální roviny idiolektu v předkládaném výzkumu nejsou stěžejní, a tudíž není potřeba jim zde věnovat větší pozornost (podobné koncepty a metody jsou však od 10. let 21. století užívány a rozvíjeny ve výzkumech v rámci projektu NHM, viz kap 2.2.2).

[26] Ve Scottově článku jsou hodnoty relativní frekvence uváděny v procentech, nikoliv v počtu výskytů na milion (či jiné dané číslo).

[27] „If the percentage is similar, the item may be ignored. Where there is a great disparity in frequency, however, it is possible to identify an item as key. The actual calculation of 'keyness' is done using the chi-square statistic, but the important point to grasp here is that the notion underlying it is one of outstandingness. In other words, if a word occurs outstandingly frequently in our text, it will be key."

[28] „[W]ords which are key in a large number of texts of a given type."

[29] „[W]ords found to be key in the same texts as a given key key word."

Od prvního Scottova uvedení a popisu metody analýzy klíčových slov se s rozšířením jeho softwaru *WordSmith Tools* postupně objevovalo čím dál více výzkumů užívajících tuto metodu při zkoumání různých typů textů (Culpeper & Demmen 2015, s. 90n., 94n.). Různorodé cíle výzkumů užívajících analýzu klíčových slov vyžadovaly větší či menší úpravy metodologické procedury. Rozšíření metody dále vedlo ke vzniku rozsáhlejších prací pojednávajících koncepty klíčových slov a klíčovosti z teoretičtějších perspektiv. Z významných teoreticky zaměřených prací lze uvést monografii *Textual Patterns: Key words and corpus analysis in language education* (Scott & Tribble 2006), zabývající se kromě obecné teorie též možností využití klíčových slov ve vzdělávání, a sborník *Keyness in Text* (Bondi & Scott 2010), v němž jednotlivé studie rozebírají koncept klíčovosti jak na obecné rovině, tak v konkrétních diskurzech a žánrech. Analýze klíčových slov je věnována pozornost také v příručkách dalších lingvistických disciplín — využití této metody v sociolingvistice je diskutováno např. v monografii *Sociolinguistics and Corpus Linguistics* (Baker 2010, s. 26n., 133–141); důležitost metody pro analýzu diskurzu je zdůrazněna např. v úvodu sborníku *Corpora and Discourse Studies: Integrating Discourse and Corpora*, kde je také hojně užívána v jednotlivých studiích (Baker & McEnery 2015a), dále je užívána též ve studiích v knize *Patterns and Meanings in Discourse: Theory and practice in corpus-assisted discourse studies (CADS)* (Partington & Duguid & Taylor 2013).

Užívání analýzy klíčových slov se tedy ustálilo zejména při zkoumání politického diskurzu (viz kap. 2.2.2), mediálních, odborných a literárních textů (Culpeper & Demmen 2015, s. 90n., 94n.). Při těchto analýzách se zkoumaný korpus skládá obvykle z textů řazených k jednomu žánru, vydávaných v jednom periodiku nebo týkajících se jednoho tématu. Metoda však má značný potenciál také pro zkoumání individuálního jazyka jedince, kdy se zkoumaný korpus skládá z textů jediného původce.

Jednou z prvních prací využívajících analýzu klíčových slov pro výzkum „individuálního textového profilu“ (pojem *idiolekt* se v práci neužívá) je článek lingvisty Davida Coniama z roku 2004 nazvaný *Concordancing Oneself: Constructing Individual Textual Profiles*. Přestože v článku není analýza dat popsána příliš detailně a konkrétně metoda analýzy klíčových slov je komentována naprosto minimálně, jedná se o práci přínosnou především z hlediska metodologie přípravy a výběru korpusů pro zkoumání jazyka jedince. Coniam pro svůj výzkum sestavil dva žánrově i oborově blízké autorské korpusy: první se skládal z jeho vlastních akademických textů (tematicky zaměřených na komputační lingvistiku a jazykové testování) a druhý z akademických textů lingvistky Amy

Bik May Tsui (tematicky zaměřených na vzdělávání učitelů a analýzu diskurzu). Oba zkoumané korpusy srovnával s referenčním korpusem, konkrétně s psanou částí *BNC Sampler Corpus*, což je reprezentativní subkorpus o rozsahu cca jedné padesátiny *British National Corpus* zachovávající proporce textových typů celého korpusu (Coniam 2004, s. 274n.). Coniam zkoumal tři aspekty textových profilů obou autorů: slovní zásobu, osobnost autora v textu (tedy do jaké míry užívá autorský singulár, autorský plurál a pasivní konstrukce) a tzv. *hedging* (tedy slova či fráze vyjadřující jistou míru pochybnosti či naopak jistoty ve vztahu k autorovým tvrzením). Pro srovnání slovní zásoby textů obou autorů využil nástroje *WordList* a *KeyWords* z již zmiňovaného softwaru *WordSmith Tools* a také program *kfNgram,* ve kterém provedl analýzu nejfrekventovanějších tetragramů (ibid., s. 277–281). Na provedených analýzách slovní zásoby textů je poměrně diskutabilní (přinejmenším z dnešního pohledu) autorovo rozhodnutí odstranit z frekvenčních seznamů slov (vytvořených pomocí *WordList*) všechna gramatická slova, která však autor explicitně nevymezil ani nevyjmenoval. Výsledné frekvenční seznamy slov jsou tudíž velmi podobné výsledným seznamům klíčových slov identifikovaných pomocí log-likelihood testu. Nabízí se tedy otázka, proč autor vůbec generoval frekvenční seznamy slov, když analýza klíčových slov dokáže gramatická slova „eliminovat“ sama o sobě; pokud by se přesto ve výsledném seznamu klíčových slov nějaká gramatická slova objevila, jednalo by se taktéž o hodnotnou informaci stojící za případnou kvalitativní interpretaci.

Právě metodologické nedostatky vyskytující se ve výzkumech užívajících analýzu klíčových slov diskutuje Jonathan Culpeper ve svém článku *Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet* z roku 2009 a sám nabízí metodologii preciznější. Ani v Culpeperově článku se sice neužívá pojem *idiolekt*, ale předmětem jeho výzkumu byl individuální jazyk literárních postav — konkrétně analyzoval promluvy šesti postav Shakespearova dramatu *Romeo a Julie*. Tyto analýzy autorovi posloužily především pro ilustraci teoretických aspektů klíčových slov a metodologie jejich analýzy (Culpeper 2009, s. 30–32). Culpeper vytyčil pět otázek, před kterými stojí badatel provádějící analýzu klíčových slov, a s odkazy na dřívější výzkumy jiných autorů nabídl možnosti, jaká rozhodnutí v těchto otázkách může badatel učinit. První otázka se týká toho, jak velký by měl být referenční korpus. Jednoduchá odpověď by zněla, že čím větší, tím lepší, ale Culpeper s odkazy na výzkumy, v nichž byl výsledný seznam klíčových slov téměř identický bez ohledu na to, zda referenční korpus obsahoval 100 milionů slov, nebo 1 milion slov, poukázal, že od jistého počtu slov již

velikost referenčního korpusu nehraje příliš roli. Naproti tomu badatelovo rozhodnutí hraje významnou úlohu ve druhé otázce, tedy z jakých textů (textových typů) by se referenční korpus měl skládat. Culpeper na příkladech ukázal, že „volba referenčního korpusu ovlivní, zda jsou všechna výsledná klíčová slova relevantní pro konkrétní aspekt zkoumaného textu (či textů). Čím jsou si zkoumaný a referenční korpus bližší [z hlediska diskurzu, žánru, tématu apod.], tím větší je pravděpodobnost, že výsledná klíčová slova budou odrážet něco specifického pro zkoumaný korpus.“[30] Třetí otázkou je, jaká by měla být stanovena hodnota minimální absolutní frekvence výskytu slov zahrnutých do seznamu klíčových slov. Tato hodnota závisí především na velikosti zkoumaného korpusu — dle Culpepera bývá zvykem ji stanovit na 10 výskytů pro „velké“ korpusy, pro „menší“ korpusy se užívá méně, typicky 5 výskytů. Čtvrtá otázka spočívá v tom, jaký test statistické signifikance by měl být použit pro výpočet klíčovosti. Jak již bylo citováno v kap. 2.2, obvykle se užívají testy log-likelihood a chí-kvadrát, jejichž výsledky bývají podobné, avšak log-likelihood je spolehlivější i s menšími korpusy. Konečně pátá otázka také souvisí se statistickou signifikancí a týká se toho, jaká by měla být stanovena úroveň pravděpodobnosti náhody, tzv. *p-value*. Culpeper ve svém výzkumu stanovil p-value 0,01, což znamená, že ve výsledném seznamu klíčových slov byla pouze slova, „u nichž se předpokládalo, že jejich rozdíly [mezi zkoumaným a referenčním korpusem] mají 1% nebo menší pravděpodobnost, že se jedná o náhodu.“[31] Tato hodnota také v praxi ovlivní, kolik slov bude ve výsledku identifikováno jako klíčových, tedy čím nižší je p-value, tím méně slov bude výsledný seznam obsahovat (Culpeper 2009, s. 34–36).

Analýza klíčových slov má svou tradici také v českém lingvistickém prostředí, kde se jí z teoretického i praktického hlediska věnuje především Václav Cvrček (viz kap. 2.2.2). *Příručka ČNK* termín *klíčové slovo* definuje jako jednotku, „jejíž frekvence v rámci textu je výrazně vyšší, než bychom mohli očekávat na základě frekvence této jednotky v referenčním korpusu,“ a uvádí, že „[k]líčovými slovy jsou nejčastěji vlastní jména (často typická pro konkrétní text), žánrově specifické tvary (např. verba 1. os. pl. prez. pro kuchařky a návody), lexikální signály tématu textu, příp. doby jeho vzniku“ (heslo *Klíčové slovo (keyword)*, in:

[30] „[T]he choice of the reference corpus will affect whether you acquire keyword results that are all relevant to the particular aspect of the text(s) you are researching. The closer the relationship between the target corpus and the reference corpus, the more likely the resultant keywords will reflect something specific to the target corpus.“

[31] „Thus, words whose differences were considered to have a 1% chance or less of being a fluke would be included as keywords; words with more of a chance of being a fluke would be excluded.“

Příručka ČNK). O statistické signifikanci již byla řeč výše, v *Příručce ČNK* je k tomu ještě zdůrazněna potřeba „dbát na zjištění míry relevance (tzv. effect-size), která vyjadřuje, do jaké míry jsou frekvence ve zkoumaném textu a v referenčním korpusu odlišné, a tedy relevantní pro následnou analýzu“ (ibid.). V rámci ČNK se jako effect-size metrika užívá DIN, který může dosahovat hodnot od -100 („daný jev se ve zkoumaném textu nevyskytuje, je pouze v referenčním korpusu“) do 100 (daný jev „se vyskytuje pouze ve zkoumaném textu“), v praxi lze za prominentní (klíčové) jednotky považovat ty s hodnotou vyšší než 75 (heslo *DIN*, in: *Příručka ČNK*). ČNK poskytuje svým uživatelům dva nástroje umožňující provádět analýzu klíčových slov: samostatnou aplikaci *KWords* a funkci *Analýza klíčových slov*, což je jeden z modulů v rozhraní *KonText*.

Starším z obou nástrojů je aplikace *KWords*. Její první verzi (v současnosti již neaktualizovanou) vytvořili Václav Cvrček a Pavel Vondřička, zveřejněna byla v roce 2013 a dodnes je přístupná na adrese https://kwords.korpus.cz/legacy. Autory aktuální, druhé verze jsou Václav Horký, Pavel Vondřička a Václav Cvrček, zveřejněna byla v listopadu 2023 a je přístupná na adrese https://kwords.korpus.cz. Aplikace je přístupná všem uživatelům, tedy i těm bez registrace v ČNK. *KWords* uživateli umožňuje analyzovat jakýkoliv jím nahraný text, přičemž jako referenční korpus může být použit také jakýkoliv uživatelem nahraný text, nebo jeden z vybraných korpusů ČNK — pro češtinu to jsou *SYN2020*, *Synecek2020* a *Totalita*, pro jiné jazyky (aktuálně jich je celkem 37) to jsou jazykové větve paralelního korpusu *InterCorp*. Vedle klíčových slov identifikuje *KWords* také slova nesoucí tematickou koncentraci textu. Mimoto umožňuje aplikace provádět též tzv. *keymorph analýzu* (tato metoda byla poprvé uvedena ve Fidler & Cvrček 2019, viz kap. 2.2.2), při níž se zkoumá podíl jednotlivých tvarů na celku nadřazené jednotky, typicky se tedy užívá ke zjištění, zda není v rámci lemmatu nápadně nadreprezentován jeden slovní tvar. Před provedením každé jednotlivé analýzy lze nastavit řadu parametrů, např. poziční atribut, statistický test, sílu efektu atd. Výsledky poskytují kromě samotného seznamu klíčových slov také další údaje, konkrétně o disperzi klíčových slov ve zkoumaném textu, o vztazích mezi klíčovými slovy textu (tzv. *keyword links*) a konkordance jednotlivých klíčových slov.

Novějším nástrojem je funkce *Analýza klíčových slov*, která byla zveřejněna v únoru 2024. Tato funkce je součástí rozhraní *KonText* jakožto jeden ze čtyř modulů pro zadávání dotazů a je přístupná pouze registrovaným uživatelům ČNK. Na rozdíl od aplikace *KWords* neumožňuje tento modul analyzovat klíčová slova ve vlastních nahraných textech,

nýbrž vždy v konkrétním korpusu, který uživatel zvolí z nabídky všech (sub)korpusů ČNK (k nimž má přístup), ze stejné nabídky (sub)korpusů uživatel volí též referenční korpus. Před odesláním dotazu je potřeba nastavit několik parametrů, konkrétně poziční atribut, metriku pro řazení výsledků (log-likelihood, chí-kvadrát, nebo DIN) a minimální (eventuálně též maximální) absolutní frekvenci hledaného výrazu ve zkoumaném korpusu. Jelikož je *Analýza klíčových slov* modulem pro zadávání dotazů (podobně jako např. základní modul *Konkordance*), neprobíhá analýza v obou korpusech automaticky zcela plošně, ale pouze podle zadaného hledaného výrazu. V něm však lze používat regulární výrazy, a tak vyhledáním defaultního výrazu . * uživatel získá všechny výsledky (resp. dle zvolené metriky řazených prvních 1000 výskytů). Výsledný seznam má podobu tabulky zobrazující klíčová slova (případně jiný zadaný poziční atribut), hodnoty všech tří používaných metrik a hodnoty absolutní i relativní frekvence ve zkoumaném i referenčním korpusu, tuto tabulku lze uložit ve formátech CSV, XLSX, XML a JSONL. Pro každý vyhledaný výraz lze pomocí pozitivního filtru zobrazit jeho konkordanci jak ve zkoumaném, tak v referenčním korpusu, s takto zobrazenými výsledky lze dále pracovat stejně jako v modulu *Konkordance* (tedy např. provést analýzu kolokací apod.).

2.2.2 Klíčová slova a politický diskurz

Václav Klaus byl od roku 1989 do roku 2013 aktivním politikem a k politickým tématům se veřejně vyjadřuje i po ukončení své politické kariéry dodnes. Z tohoto důvodu lze texty shromážděné v korpusu *VK* považovat za součást politického diskurzu, respektive za texty podílející se na jeho konstituování. Předkládaný výzkum je primárně korpusově lingvistický, avšak kvůli politicky diskurzní povaze zkoumaných dat lze výsledky jejich korpusových analýz potenciálně využít v rámci analýzy diskurzu. V kap. 4.3.3 se na případové studii Klausových vystoupení na setkáních *Euro Business Breakfast* pokusím ukázat jednu z možností, jak by diskurzní analýza dat v korpusu *VK* mohla vypadat, a to konkrétně s využitím klíčových slov. V následujících odstavcích je shrnuto několik výzkumů užívajících metodu analýzy klíčových slov při analýze politického diskurzu. Tyto výzkumy mi byly přínosné pro koncepci vlastní metodologie a inspirativní při interpretaci výsledků.

Studie *Who Benefits When Discourse Gets Democratised? Analysing a Twitter Corpus around the British Benefits Street Debate* z roku 2015 představuje výzkum Paula Bakera a Tonyho McEneryho, v němž analyzovali korpus sestavený z tweetů reagujících na

televizní debatu o dokumentární sérii *Benefits Street*, která se věnovala občanům Velké Británie pobírajícím dávky státní sociální podpory (anglicky běžně označované jako *benefits*). Výchozím bodem této diskurzní analýzy byla identifikace 100 nejsignifikantnějších klíčových slov zkoumaného korpusu vzešlých ze srovnání s referenčním korpusem sestaveným z odpovídajícího počtu náhodných tweetů z období krátce před odvysíláním zmiňované debaty. U všech klíčových slov byly následně provedeny kolokační a konkordanční analýzy, na jejichž základě byly vymezeny tři hlavní diskurzy projevující se v korpusu: „líní chudí", „chudí jako oběti" a „bohatí bohatnou". Každý z těchto diskurzů byl podroben kvalitativně orientované kritické analýze poukazující na souvislosti, jak jsou konkrétní klíčová slova a jejich kolokace užívány k utváření obrazu občanů pobírajících dávky (Baker & McEnery 2015b, s. 245–250, 261–263).

Českému prostředí jsou blízké výzkumy Masako Fidler a Václava Cvrčka, kteří od roku 2011 spolupracují v rámci projektu *The Needle-in-a-Haystack Method* (NHM), jehož cílem je propojovat kvantitativní a kvalitativní metody analýzy textu. Na počátku projektu stál zejména zájem o analýzu klíčových slov a snaha tuto metodu zefektivnit pro využití v diskurzní analýze. Autoři tak postupně rozšířili koncept klíčovosti a v řadě článků představili několik vylepšení původní metody.[32] Analyzovaným materiálem v těchto výzkumech byly různé texty českého politického a mediálního diskurzu.[33]

V článku *A Data-Driven Analysis of Reader Viewpoints: Reconstructing the Historical Reader Using Keyword Analysis* z roku 2015 představili Fidler a Cvrček analýzu klíčových slov novoročních projevů československého prezidenta Gustáva Husáka z let 1975 až 1989, přičemž zkoumaný korpus srovnali se dvěma různými referenčními korpusy z ČNK: s korpusem *Totalita* (diachronní korpus žurnalistických a propagandistických textů socialistického Československa z 50. až 70. let 20. století) a korpusem *SYN2010* (vyvážený synchronní reprezentativní korpus psané češtiny let 2005–2009). Rozdíly mezi výslednými seznamy klíčových slov odhalily, co v Husákových projevech jako klíčové vnímali doboví recipienti (*Totalita* jako referenční korpus), a co je v nich spíše typické pro jazyk socialismu obecně (*SYN2010* jako referenční korpus). Tento postup ukázal, jak volba referenčního

[32] Přehled všech výstupů projektu je k dispozici na adrese https://sites.brown.edu/needle-in-haystack/about/research/.

[33] Vztah mezi politickým a mediálním diskurzem je obousměrný — na jedné straně texty mediálního diskurzu mohou politiku ovlivňovat či spoluutvářet, na druhé straně texty politického diskurzu bývají pro média zásadním zdrojem, na němž jsou některé žánry přímo závislé. Oba řády diskurzu se tedy často prolínají, a nelze mezi nimi vždy jednoznačně rozlišovat (Merta 2007, s. 18n.).

korpusu ovlivňuje výsledky analýz a jak se toho dá využít pro různé výzkumné cíle, pokud je badatel schopen výsledky adekvátně interpretovat (Fidler & Cvrček 2015, s. 201–205, 219n.). Význam tohoto výzkumu spočívá také v uvedení effect-size metriky klíčovosti DIN (viz kap. 2.2.1); v příloze článku autoři podrobně popsali způsob jejího výpočtu a její výhody oproti jiným metrikám (ibid., s. 226–231).

Také následné výzkumy v rámci projektu NHM přinesly další metodologická vylepšení a přidaly k analýze klíčových slov „nadstavbové" úrovně analýzy. Tyto pokročilejší metody mají své uplatnění především ve výzkumech orientovaných primárně na analýzu diskurzu, tudíž jsou zde uváděné články popsány jen stručně se zaměřením na rozvíjení metodologie související s konceptem klíčovosti.

Již zmiňovaná metoda keymorph analýzy (viz kap. 2.2.1) byla představena v článku *Keymorph analysis, or how morphosyntax informs discourse* publikovaném roku 2019, kde byla užita na materiálu korpusu československých a českých prezidentských projevů *Speeches* z ČNK. Keymorph analýza zaměřující se na morfosyntaktické rysy s minimem lexikálních informací tak v jazycích s bohatou flexí poskytuje klíč k obecnějším vlastnostem diskurzu a informuje o způsobu, jakým jsou situace a jejich účastníci důsledně reprezentováni (Fidler & Cvrček 2019).

Syntéza dílčích metod zkoumajících klíčovost (prominenci) různých jednotek, tzv. *multi-level discourse prominence analysis* (MLDPA), byla představena ve studii *Going Beyond "Aboutness": A Quantitative Analysis of Sputnik Czech Republic* z roku 2018,[34] kde byly analyzovány texty z české mutace publicistického webu ruské zpravodajské agentury Sputnik. Tato víceúrovňová analýza zkoumá klíčové slovní tvary, klíčová lemmata, klíčové morfémy a věty s vysokou hustotou klíčových lemmat. Kombinace poznatků dílčích analýz tak umožňuje v textech odhalit nejen dominantní témata, ale též způsoby reprezentace situací a účastníků diskurzu a společné vlastnosti vět, v nichž se klíčová lemmata shlukují. MLDPA tak poskytuje celkový obraz toho, co vše je v textech zkoumaného korpusu obzvlášť nápadné v kontrastu s jazykovými jednotkami a vzorci běžnými v textech referenčního korpusu (Fidler & Cvrček 2018b).

Zatím posledním metodologickým rozšířením analýzy klíčových slov vzešlým z projektu NHM je tzv. *market basket analysis* (MBA), která byla představena v článku *No*

[34] Zhuštěnější podoba výzkumu byla prezentována v souvisejícím článku *More than keywords: Discourse prominence analysis of the Russian Web Portal Sputnik Czech Republic* z roku 2019 (Cvrček & Fidler 2019).

keyword is an island: In search of covert associations publikovaném v roce 2022, kde byla aplikována na téma migrace ve vybraných českých online médiích. MBA je technika vytěžování dat pocházející z marketingu, kde se užívá k procházení mnoha transakcí, aby bylo možné identifikovat zboží, jež je často kupováno společně; pokud MBA odhalí daný minimální počet transakcí, v nichž se určité položky vyskytují společně, pak mezi těmito položkami vytvoří asociativní vazbu. Podobně dokáže MBA vytvářet asociativní vazby také mezi klíčovými slovy vyskytujícími se společně v jednotlivých textech ve zkoumaném korpusu. Při analýze diskurzu tedy MBA eliminuje nedostatky postupů tradičně užívaných návazně na analýzu klíčových slov — konkordanční analýzy (je při práci s rozsáhlým množstvím dat časově i interpretačně náročná), kolokační analýzy (bere v potaz pouze jednotky vyskytující se v těsné blízkosti, a vypovídá tedy spíše o charakteristikách lexikonu než diskurzu) a keyword links (nevyhodnocují významnost či sílu vztahů mezi danými klíčovými slovy) — a dokáže je alespoň do určité míry suplovat. Kvalitativní interpretace získaných výsledků tak lze díky této metodě provádět při velkém rozsahu dat stejně efektivně a spolehlivě jako při jejich menším rozsahu (Cvrček & Fidler 2022).

3 Metodologie výzkumu

V předkládaném výzkumu převažuje kvantitativní přístup, který je podle potřeby doplněn přístupem kvalitativním. Následující dvě podkapitoly detailně popisují metodologii celého výzkumu, aby bylo možné provedené postupy replikovat, případně odhalit jejich slabá místa a potenciálně je vylepšit. Krok za krokem je tedy pozornost věnována nejprve sběru dat, poté analýze dat. V duchu hnutí otevřené vědy jsou soubory s výsledky analýz veřejně přístupné na adrese https://osf.io/7btnc/?view_only=95c8e2f540764097b22ef8effa0de8dd. Zdrojová data nemohou být z důvodu ochrany autorských práv publikována, všechna jsou však snadno dohledatelná a volně přístupná na internetu.

3.1 Sběr dat

Základním předpokladem úspěšné analýzy idiolektu jedince je především vhodný vzorek jeho jazykové produkce, který umožní vytvoření autorského korpusu. Takový korpus analýzu idiolektu technicky usnadní a empiricky zpřesní její výsledky, musí však splňovat několik zásadních podmínek platných pro korpusy obecně — musí být dostatečně velký a reprezentativní a musí obsahovat pouze autentické texty. V případě autorského korpusu, tedy korpusu jediného člověka, je situace poněkud specifická, protože rozsah dostupných textů je ze své podstaty relativně omezený a jejich reprezentativnost je ovlivněna tím, jaké typy textů daný jedinec produkuje (respektive jaké jsou k dispozici). I přesto je možné uvedené podmínky splnit, je ale důležité před vytvořením korpusu jasně vymezit vzorek pro sběr dat (Čermák 2017, s. 23).

Jak bylo zmíněno v úvodu, jedním z důvodů, proč byl pro analýzu idiolektu v tomto výzkumu vybrán právě Václav Klaus, je skutečnost, že již několik desetiletí konzistentně zveřejňuje velké množství vlastních rozličných textů, a patří tak v tomto ohledu k nejproduktivnějším osobnostem české politiky. Díky počtu, různorodosti a relativní dostupnosti těchto textů je Klaus vhodným adeptem na vytvoření autorského korpusu. Tyto aspekty však s sebou kromě pozitiv nesou také určité metodologické výzvy a technické obtíže.

Již během prvních rešerší k výzkumu začalo být zřejmé, že vzhledem k velkému množství Klausových textů a jejich rozptýlení v online prostoru i tištěných zdrojích bude potřeba analyzovaný vzorek omezit jen na určitý výběr z nich. Ideálním stavem by

samozřejmě bylo, kdyby Klausův autorský korpus ve své konečné podobě obsahoval všechny jeho veřejně dostupné texty, ale sestavení takového korpusu by přesahovalo ambice předkládané diplomové práce.

Obzvláště užitečnou platformou shromažďující Klausovy texty se ukázaly být jeho oficiální webové stránky na adrese www.klaus.cz spuštěné na podzim roku 2000,[35] ale obsahující též texty starší (viz níže). Tyto stránky už od svého vzniku slouží nejen jako komunikační kanál, kterým Klaus veřejnosti zprostředkovává informace o sobě a svých aktivitách, ale také jako jedinečný prostor jeho sebeprezentace, neboť kromě textů určených primárně pro tyto stránky obsahují i texty původně publikované jinde (např. články v novinách či rozhovory v časopisech), respektive vytvořené k některým konkrétním událostem (např. prezidentské projevy či přednášky na odborných konferencích). Pro jejich široký záběr a reprezentativní povahu byly tedy Klausovy webové stránky zvoleny jako zdroj textů k vytvoření autorského korpusu.

Kromě vlastních Klausových textů stránky obsahují také texty, u kterých je Klaus pouze spoluautorem (např. společná prohlášení) či jejichž autorem není vůbec (např. sdělení tiskového odboru prezidentské kanceláře). Aby byla splněna zmiňovaná podmínka autenticity zkoumaného jazykového materiálu jedince, nebyly tyto texty do sběru dat zahrnuty.

Původním záměrem výzkumu bylo neomezovat se při analýze idiolektu jen na některý modus jazyka, ale zahrnout projevy psané i mluvené, a to včetně přechodových typů psané mluvenosti a mluvené psanosti (jejíž podstata prakticky vylučuje výskyt takovýchto textů na oficiálních webových stránkách, což se při sběru dat potvrdilo; Hoffmannová 2017). Díky tomu by bylo možné sledovat potenciální rozdíly v idiolektu mezi jednotlivými mody. U nezanedbatelného množství textů (především u velké části rozhovorů) však nebylo možné modus spolehlivě určit; u prokazatelně mluvených textů (diskuse a některé rozhovory) zase situaci komplikovaly evidentní redakční úpravy Klausových promluv, přičemž míra i povaha těchto zásahů se mezi jednotlivými texty značně lišila. Pro zachování autenticity jazykového materiálu a pro konzistentnost prováděných analýz tudíž nebyly do sběru dat zahrnuty texty, jejichž modus nebylo možné jednoznačně určit, a nakonec ani texty „čistě“ mluvené.

[35] KLAUS, Václav. Aktuální reakce Václava Klause věnovaná prvním čtenářům této stránky. *Klaus.cz* [online]. 11. 10. 2000. URL: https://www.klaus.cz/clanky/1833. Cit. 29. 11. 2023.

Konečně bylo potřeba vzorek textů vymezit časově. Aby byl autorský korpus v rámci výše uvedených podmínek co největší, byl počáteční časový bod určen nejstarším textem dostupným na Klausových webových stránkách, který je datován dne 2. března 1995.[36] Konečný časový bod byl stanoven na 31. října 2023, neboť počátkem listopadu 2023 již probíhal sběr dat. Texty tudíž pokrývají časové rozpětí cca 28,5 let.

Texty použité za účelem sestavení autorského korpusu Václava Klause lze tedy vymezit následujícími čtyřmi podmínkami. Byly shromážděny

1) pouze texty zveřejněné na webových stránkách www.klaus.cz;
2) pouze texty, jejichž jediným (uvedeným) autorem je Václav Klaus;
3) pouze texty v modu psaném a v modu psané mluvenosti (tj. texty původně psané, ale určené k mluvenému projevu);
4) pouze texty zveřejněné do 31. října 2023 včetně.

Pro vlastní realizaci sběru dat podle uvedených vymezujících podmínek se nabízely dvě možnosti — automatizovaný strojový web scraping, nebo manuální kopírování a ukládání textů. Výhodou první možnosti by sice byla časová úspora při samotném ukládání dat, ta by však dále vyžadovala důkladnou kontrolu a velká část z nich následně také manuální úpravy, a to především ze dvou důvodů. Za prvé, mnoho jednotlivých stránek na Klausově webu, které by měly teoreticky odpovídat právě jednomu textu, obsahuje kromě Klausova vlastního textu také neoddělený text průvodní, jehož autorem Klaus není. Typicky se jedná třeba o informaci o zaslání dopisu, který je následně citován v plném znění.[37] Přítomnost takovýchto průvodních textů v korpusu by porušila výše uvedenou druhou podmínku. Za druhé, třídění textů do rubrik na webových stránkách není vždy zcela konzistentní, respektive ne vždy podává dostatečnou a/nebo přesnou informaci potřebnou ke kategorizaci textů (a tedy jejich opatření patřičnými metadaty) pro korpus. Typickým případem jsou např. texty v rubrice nazvané *Dokumenty*, která kromě dokumentů v užším smyslu (převážně dopisy) obsahuje také předmluvy ke sborníkům (v korpusu řazeny k registru *oborová literatury*), odpovědi na anketní otázky (v korpusu řazeny k registru *publicistika*) apod. Z těchto důvodů jsem se rozhodl pro manuální kopírování a ukládání textů. Tato možnost je sice časově náročnější, ale lidské vyhodnocení textů by mělo při jejich

[36] KLAUS, Václav. Poznámky k dnešní fázi společenské přeměny. *Klaus.cz* [online]. 2. 3. 1995. URL: https://www.klaus.cz/clanky/1628. Cit. 8. 5. 2024.

[37] Např. Prezident republiky Václav Klaus kondoloval rodině generála Fajtla. *Klaus.cz* [online]. 4. 10. 2006. URL: https://www.klaus.cz/clanky/1632. Cit. 29. 11. 2023.

selekci a ukládání zcela minimalizovat pravděpodobnost, že se do korpusu dostanou takové, které nesplňují výše stanovené podmínky. Tento manuální postup také umožňuje provádět patřičnou kategorizaci textů a zaznamenání všech údajů (metadat) relevantních pro vytvářený korpus rovnou při jejich ukládání.

Pracovní postup sběru dat probíhal následovně. Na Klausových webových stránkách jsem v rubrikách *Články a eseje*, *Ekonomické texty*, *Projevy a vystoupení*, *Dokumenty*, *Co Klaus neřekl*, *Excerpta z četby* a *Tisková sdělení* (texty v ostatních rubrikách nesplňují vždy minimálně jednu z výše uvedených podmínek) postupně otvíral všechny jednotlivé stránky. Na každé stránce jsem vyhodnotil, zda obsahuje text (případně alespoň část textu nebo naopak více různých textů) splňující podmínky pro zařazení do Klausova autorského korpusu. V záporném případě jsem stránku zavřel a pokračoval na další; v kladném případě jsem daný text zkopíroval a uložil jako *plain text* do TXT souboru s kódováním Unicode (UTF-8). Každý uložený soubor jsem pojmenoval v podobě „XY-yyyy/mm/dd“, kde „XY“ označuje registr či žánr, k němuž text náleží (viz níže), a „yyyy/mm/dd“ označuje datum ve formátu ‚rok/měsíc/den‘, kdy byl text publikován. Pokud bylo ve stejný den publikováno více textů stejného registru či žánru, bylo na konec názvu souboru přidáno ještě písmeno abecedy (např. *TS-2006/09/01/b* označuje druhé tiskové sdělení publikované dne 1. září 2006). Následně jsem do tabulky ve formátu XLSX zaznamenal každý uložený soubor a uvedl k němu všechny údaje sloužící jako korpusová metadata. Každý řádek tabulky odpovídá právě jednomu uloženému souboru, každý sloupec tabulky odpovídá jednomu strukturnímu atributu. Zaznamenávané strukturní atributy a jejich hodnoty jsou následující:

1) název souboru / doc.id – jedinečné hodnoty v rámci osmi podtypů:
    i. CKN – Co Klaus neřekl
    ii. ČE – Články a eseje
    iii. D – Dokumenty
    iv. ET – Ekonomické texty
    v. PV – Projevy a vystoupení
    vi. SK – Statě a kapitoly
    vii. TS – Tisková sdělení
    viii. VČ – Výpisky z četby
2) název textu / doc.title – jedinečné hodnoty
3) datum publikování / doc.date – jedinečné hodnoty

4) období / doc.period – pět hodnot:
    i. premiér
    ii. poslanec
    iii. předseda PSP
    iv. prezident
    v. exprezident
5) modus / doc.modus – dvě hodnoty:
    i. psaný
    ii. psaný (psaná mluvenost)
6) registr / doc.registr – čtyři hodnoty:
    i. dokumenty
    ii. oborová literatura
    iii. publicistika
    iv. veřejná vystoupení
7) typ textu / doc.txtype – jedenáct hodnot:
    i. dementi
    ii. dopis
    iii. posudek
    iv. projev
    v. přednáška
    vi. recenze
    vii. stať / kapitola
    viii. tiskové sdělení
    ix. výpisky z četby
    x. zápisky z cest
    xi. článek / esej
8) typ zdroje / doc.medium – sedm hodnot:
    i. kniha / sborník
    ii. newsletter
    iii. noviny
    iv. rozhlas
    v. televize
    vi. webové stránky
    vii. časopis

9) zdroj / doc.source – jedinečné hodnoty
10) poznámka / doc.comment – jedinečné hodnoty (jedná se o doplňující informace, které by mohly být potenciálně užitečné při práci s konkordancí, nemusí být u každého textu).

Uvedeným postupem jsem shromáždil celkem 2 313 různých textů, které jsem jednotlivě uložil jako textové soubory a které jsem zaznamenal do tabulky i s příslušnými metadaty. Všechny soubory i tabulku jsem předal zaměstnancům ÚČNK, kteří z tohoto datasetu vytvořili *Korpus textů Václava Klause* (v této práci dále označován jako *VK*), k němuž mi udělili přístup v rozhraní *KonText* na webu ČNK na adrese www.korpus.cz/kontext. Samotný korpus zatím není uživatelům ČNK veřejně přístupný, ale v budoucnu pravděpodobně zpřístupněn bude.

V rámci technického zpracování shromážděných textů v ÚČNK byly všechny texty automaticky anotovány a lemmatizovány. Při anotaci byly k textovým datům přiřazeny jak strukturní, tak poziční atributy. Strukturní atributy se vždy váží k jednomu dokumentu (textovému souboru), a kromě výše vypsaných deseti, zaznamenávaných ve zmiňované tabulce, jsou ve výsledném korpusu *VK* uváděny ještě další čtyři:

11) jazyk / doc.lang – jedna hodnota:
    i. cs
12) zdrojový jazyk / doc.src_lang – jedna hodnota:
    i. cs
13) autor / doc.author – jedna hodnota:
    i. Václav Klaus
14) rok publikování / doc.pubDateYear – 29 hodnot:
    i. {1995–2023}.

Výsledný korpus *VK* obsahuje celkem 1 750 891 pozic (tokenů), z toho 1 475 640 pozic bez interpunkce (počet zjištěn pomocí dotazu *[!tag="Z.*"]*). Pozicím byly při anotaci přiřazeny poziční atributy, kterých je celkem jedenáct. Tyto poziční atributy a počty jejich typů jsou uvedeny v následující tabulce:[38]

[38] Vysvětlivky k pozičním atributům jsou dostupné na adrese http://wiki.korpus.cz/doku.php?id=pojmy:atributy_pozicni&rev=1641224995 (heslo *Poziční atributy*, in: *Příručka ČNK*). Cit. 1. 5. 2024.

	počet typů s interpunkcí	počet typů bez interpunkce
word	98 892	98 830
lc	88 814	88 752
sforma	98 896	98 831
lemma	36 604	36 555
lemma_lc	35 893	35 844
sublemma	37 942	37 892
sublemma_lc	37 225	37 175
tag[39]	2 013	2 007
pos	15	14
case	8	8
verbtag[40]	77	77

Pro úplnost lze ještě dodat, že korpus *VK*, byť zatím není veřejně přístupný, je označen následujícími štítky sloužícími k procházení korpusů v ČNK: *čeština*, *specializovaný*, *synchronní*, *psaný*.

Kromě samotného korpusu *VK* analyzuji Klausův idiolekt také na materiálu dvou korpusů veřejně přístupných v ČNK obsahujících mimo jiných také jeho texty; konkrétně se jedná o specializované korpusy *ParlCorp* a *Speeches*.

Korpus *ParlCorp* byl zveřejněn v roce 2021 a obsahuje monologické projevy pronesené na půdě Poslanecké sněmovny Parlamentu České republiky (PSP) v období let 1993–2021. Základem korpusu jsou stenoprotokoly sněmovních jednání veřejně přístupné na adrese www.psp.cz. Jedná se o korpus lemmatizovaný a anotovaný. Strukturní atributy se zde váží k jednotlivým textům a mluvčím, díky čemuž je možné vytvořit dva subkorpusy pro vzájemné srovnávání (viz kap. 3.2). Celý korpus obsahuje téměř 32,7 milionu pozic bez

[39] V korpusu *VK* byl použit stejný tagset jako v korpusu *SYN2020* (tagset *cs_cnc2020*). Kompletní přehled struktury a hodnot tohoto aktuálního tagu je dostupný na adrese http://wiki.korpus.cz/doku.php?id=seznamy:tagy&rev=1650299545 (heslo *Morfologické značky (tagy) a jejich hodnoty*, in: *Příručka ČNK*). Cit. 1. 5. 2024.

[40] Verbtag obsahuje podrobnější informace (nad rámec tagu) o gramatických kategoriích slovesa, a to jak u sloves jednoduchých, tak složených. Kompletní přehled struktury a hodnot aktuálního verbtagu je dostupný na adrese http://wiki.korpus.cz/doku.php?id=seznamy:verbtagy_detail&rev=1668772417 (heslo *Morfologické kategorie a hodnoty v atributu verbtag a jejich značkování*, in: *Příručka ČNK*). Cit. 1. 5. 2024.

interpunkce, kterým byly při anotaci přiřazeny poziční atributy *word*, *lemma* a *tag* (heslo *ParlCorp: Korpus českých parlamentních projevů*, in: *Příručka ČNK*).

Korpus *Speeches* byl zveřejněn v roce 2015 a obsahuje oficiální prezidentské projevy pronesené u příležitosti periodicky se opakujících výročí a svátků (novoroční projev, projev ke dni vzniku samostatného československého státu, vánoční poselství a výročí protektorátu) v období let 1918–2015. Mluvčími jsou v korpusu československí a čeští prezidenti, případně jiní politici mluvící v zastoupení prezidenta. Také se jedná o korpus lemmatizovaný a anotovaný. Díky strukturním atributům popisujícím jednotlivé projevy je opět možné jednoduše vytvořit dva subkorpusy pro vzájemné srovnávání (viz kap. 3.2). Celkově korpus obsahuje (kvůli své úzké specializaci) pouze 217 314 pozic bez interpunkce, kterým byly při anotaci přiřazeny poziční atributy *word*, *lemma*, *tag*, *lc* a *lemma_lc* (heslo *Korpus prezidentských projevů Speeches*, in: *Příručka ČNK*).

3.2 Analýza dat

Hlavní metodou analýzy dat v předkládaném výzkumu je korpusově lingvistická kvantitativní metoda zvaná *analýza klíčových slov*.[41] Z teoretické perspektivy byl pojem *klíčové slovo* rozebírán v kap. 2.2, metoda jejich analýzy pak v kap. 2.2.1. Na tomto místě tedy jen připomenu, že v této práci označuje pojem *klíčové slovo* takové slovo, které se ve zkoumaném korpusu (či zkoumaném textu obecně) vyskytuje statisticky signifikantně častěji, než by se očekávalo na základě frekvence téhož slova v referenčním korpusu. Podstata metody analýzy klíčových slov tedy spočívá ve srovnání relativní frekvence každého slova v seznamu slov zkoumaného korpusu s relativní frekvencí téhož slova v seznamu slov referenčního korpusu. Klíčovost každého slova ve zkoumaném korpusu je vypočítána na základě testů statistické signifikance (obvykle log-likelihood či chí-kvadrát) v kombinaci s informací o míře relevance rozdílu (v praxi ČNK se užívá *difference index* neboli *DIN*).

V následujících odstavcích popisuji procedury jednotlivých analýz, jejichž výsledky jsou prezentovány v praktické části práce. Všechny kroky vlastních analýz byly prováděny v rozhraní *KonText*. Získané výsledky mající podobu seznamu slov (o rozsahu maximálně 1000 položek) jsem následně ukládal jako tabulky ve formátu XLSX, aby bylo možné s nimi

[41] Anglicky *keyword analysis*.

snadněji pracovat při interpretacích výsledků a aby mohly být souhrnně zveřejněny. V rozhraní *KonText* byly analýzy prováděny s využitím modulu pro zadávání dotazů nazvaného *Analýza klíčových slov* (viz kap. 2.2.1). V rámci ČNK se jedná o zcela novou funkci, která byla zveřejněna v únoru 2024, zaměstnanci ÚČNK mi však umožnili přístup již k její beta verzi koncem roku 2023. Od té doby jsem modul testoval, prováděl v něm zkušební analýzy a podával vývojářům zpětnou vazbu. Na základě mých připomínek byly ještě před zveřejněním modulu opraveny některé chyby a implementována drobná vylepšení. Všechny výsledky prezentované v praktické části byly získány prací s již veřejně dostupnou verzí modulu. Několik dalších technických problémů však bylo odhaleno až v průběhu finálních analýz předkládaného výzkumu. Některé z nich se vývojářům podařilo rychle odstranit (za což jim velice děkuji), jiné bohužel nikoliv. Tyto přetrvávající problémy se nejvíce dotýkají práce se subkorpusy, především když jsou užívány jakožto referenční korpusy, kvůli čemuž jsem musel částečně upravit některé původně plánované metodologické procedury popsané níže. Pro lepší přehlednost zde uvádím původní podobu procedur, provedené změny jsou popsány na patřičných místech praktické části (konkrétně v kap. 4.1), vliv těchto změn na získané výsledky a důsledky pro jejich interpretaci jsou diskutovány v závěru práce.

Analýza klíčových slov v *Korpusu textů Václava Klause*

Jádro předkládaného výzkumu tvoří analýza klíčových slov v korpusu *VK* (kap. 4.1). Ten jsem srovnával s referenčním korpusem *SYN v12*, což je aktuálně nejnovější verze (a tedy i nejrozsáhlejší, celkem obsahuje cca 5,175 miliard pozic bez interpunkce) korpusu *SYN*, která byla zveřejněna na konci roku 2023. Korpus *SYN v12* obsahuje všechny synchronní psané korpusy řady *SYN* (tedy pět korpusů reprezentativních a tři korpusy publicistické) a navíc publicistiku z let 2010–2022 (heslo *Korpus SYN verze 12*, in: *Příručka ČNK*). Nebylo by však vhodné srovnávat mezi sebou korpusy *VK* a *SYN v12* celé, neboť klíčovost není inherentní vlastností slov, nýbrž je závislá na vztahu zkoumaného a referenčního materiálu. Aby tak měly výsledky analýz patřičnou vypovídací hodnotu o Klausově idiolektu a aby se daly vhodně interpretovat, je zapotřebí srovnávat texty opravdu srovnatelné, tedy takové, jež jsou si v určitých ohledech blízké. Pokud by se srovnávaly např. Klausovy ekonomické přednášky s poezií, identifikovaná klíčová slova by byla charakteristická pro daný obor a žánr spíše než pro Klausův idiolekt. Podobně pokud by se srovnávaly např. Klausovy publicistické texty z let 2020–2023 s publicistikou z let 2003–2013,

identifikovaná klíčová slova by byla příliš podmíněna aktuálními, dříve neexistujícími jevy či událostmi (pandemie covidu-19, ruská invaze na Ukrajinu), a jejich klíčovost by tedy byla zkresleně vysoká, ba přímo absolutní (hodnota DIN 100). Ani v jednom případě by taková klíčová slova nebyla pro zkoumání Klausova idiolektu příliš relevantní.

Aby tedy byly prováděné analýzy a jejich výsledky smysluplné, vytvořil jsem s využitím strukturních atributů v korpusu *VK* řadu subkorpusů, podle nichž jsem následně vytvořil odpovídající subkorpusy také v korpusu *SYN v12*. Metadata v korpusu *VK* umožňují vytvářet subkorpusy na základě mnoha kombinací hodnot strukturních atributů. Po testování různých možností jsem se rozhodl se zřetelem k hlavnímu cíli výzkumu vytvořit subkorpusy podle atributů *doc.period* a *doc.registr*. Subkorpusy vytvořené pouze podle hodnot jednoho z těchto atributů a jejich velikosti jsou uvedeny v následujících dvou tabulkách:

doc.period	**počet pozic včetně interpunkce**
premiér	42 393
poslanec	69 945
předseda PSP	333 669
prezident	774 866
exprezident	530 018

doc.registr	**počet pozic včetně interpunkce**
dokumenty	142 405
oborová literatura	314 230
publicistika	670 610
veřejná vystoupení	623 646

Ani tyto subkorpusy však nebyly z hlediska hlavního cíle výzkumu zcela adekvátní pro srovnávání, a proto jsem vytvořil subkorpusy kombinací hodnot obou atributů. Výsledné subkorpusy a jejich velikosti (v počtu pozic včetně interpunkce) jsou uvedeny v následující tabulce:

	premiér	*poslanec*	*předseda PSP*	*prezident*	*exprezident*
dokumenty	N/A[42]	159	11 740	116 972	13 534
oborová literatura	13 049	20 519	60 900	129 627	90 135
publicistika	N/A	28 506	151 394	210 895	279 815
veřejná vystoupení	29 344	20 761	109 635	317 372	146 534

[42] „N/A“ označuje prázdnou množinu, subkorpusy kombinující takové hodnoty atributů tedy nebyly vytvořeny.

Kvůli velikosti a složení korpusu *SYN v12* nebylo možné vytvořit odpovídající referenční subkorpusy takto jednoduše z hodnot dvou atributů, ale bylo nutné dle potřeby kýžených registrů a časových období kombinovat různé hodnoty strukturních atributů *doc.syn* (původní korpus řady *SYN*), *doc.txtype_group* (skupina textových typů), *doc.txtype* (textový typ) a *doc.pubyear* (rok vydání). Výsledné subkorpusy a jejich velikosti (v počtu pozic včetně interpunkce) jsou uvedeny v následující tabulce:

	1995–1997	*1998–2003*	*2003–2013*	*2013–2022*
administrativa	164 917	416 188	571 453	526 983
oborová literatura	28 637 346	53 817 146	207 063 259	274 458 957
publicistika[43]	27 452 658	57 545 355	94 154 864	533 546 332

Celkem tedy bylo v korpusu *VK* analyzováno 18 subkorpusů, které byly srovnávány s 12 subkorpusy v korpusu *SYN v12*, a to podle následujícího klíče:

zkoumané období	**referenční období**
premiér	*1995–1997*
poslanec	*1998–2003*
předseda PSP	*1998–2003*
prezident	*2003–2013*
exprezident	*2013–2022*

zkoumaný registr	**referenční registr**
dokumenty	*administrativa*
oborová literatura	*oborová literatura*
publicistika	*publicistika*
veřejná vystoupení	*publicistika*

Každá z 18 uskutečněných analýz byla provedena na základě stejného zadaného vyhledávacího dotazu, jenž měl následující parametry:

parametr	**hodnota**
Atribut	lemma_lc [lowercase lemma]
Hledaný výraz	.*
Třídit podle	DIN (Difference index)
Min. frekvence	5
Max. frekvence	—

[43] Při volbě subkorpusů v rámci registru *publicistika* jako referenčních korpusů docházelo během analýz k technické chybě při výpočtu pomocných dat, kvůli níž analýzy neproběhly. Řešení tohoto problému je popsáno v kap. 4.1.

Poziční atribut *lemma_lc* byl zvolen především z toho důvodu, že hlavní pozornost při analýze Klausova idiolektu je v předkládaném výzkumu upřena na jeho lexikální rovinu, a tedy na lexémy jakožto jednotky primárně nesoucí význam (viz kap. 2.2). Souhrnnou informaci o sdělovaných významech poskytnou lemmata lépe než jednotlivé slovní tvary, tedy atribut *word* nebo *lc* (heslo *Lemma*, in: *Příručka ČNK*). Atribut *lemma_lc*, který oproti atributu *lemma* nerozlišuje velká a malá písmena, byl zvolen proto, aby se neutralizovaly všechny potenciální odchylky způsobené automatickou lemmatizací, které by mohly vést ke zkreslení frekvencí u slov lemmatizovaných s odlišnými počátečními písmeny. Nevýhodou sice je nerozlišování některých proprií a apelativ, avšak ani toto rozlišování není v důsledku automatické lemmatizace vždy spolehlivé.[44] Jako hledaný výraz byla ponechána defaultní kombinace .* umožňující analýzu klíčovosti všech jednotek (lemmat) ve zkoumaném korpusu. Výsledky byly tříděny podle effect-size metriky *DIN*, která vyjadřuje, nakolik jsou rozdíly relativních frekvencí ve zkoumaném a referenčním korpusu relevantní pro následnou analýzu. Jako minimální frekvence ve zkoumaném korpusu byla ponechána defaultní hodnota pěti výskytů, neboť se jedná o běžnou praxi pro korpusy, které nejsou ani příliš velké, ani příliš malé.[45] Na základě zkoumání výsledků zkušebních analýz se tato hodnota ukázala při srovnání s jinými jako uspokojivá. Maximální frekvence ve zkoumaném korpusu byla ponechána defaultně bez omezení, aby nebyly zbytečně vyřazeny jakékoliv potenciálně klíčové jednotky.

Výsledků získaných analýzami jednotlivých subkorpusů v korpusu *VK* je vzhledem k velikosti zkoumaného korpusu a rozmanitosti dat v něm obsažených značné množství. Tato skutečnost na jednu stranu poskytuje empiricky robustní základ pro případné navazující analýzy, na druhou stranu komplikuje hlubší kvalitativní interpretace výsledků. Kromě toho proměnlivé složení referenčních subkorpusů (i přes výše popsanou snahu o jejich adekvátnost vůči subkorpusům zkoumaným) nezaručuje vždy ve všech ohledech konzistentní výsledky. Z těchto důvodů jsem se za účelem adekvátnějšího dosažení hlavního

[44] Problematiku velkých a malých písmen při automatické lemmatizaci lze v korpusu *VK* ilustrovat na příkladu slova *černý*, které je v případě příjmení Černý lemmatizováno jako *Černý*, avšak v zeměpisných názvech Černá Hora a Černé moře je lemmatizováno jako *černý*, aniž by se jednalo o atribut apelativní. Dalším příkladem může být zkratka CEP (Centrum pro ekonomiku a politiku), která je na rozdíl od uzualizovaných zkratek typu EU (lemmatizována jako *EU*) lemmatizována jako *cep*, a v rámci tohoto lemmatu tedy splývá s cepem ve významu nářadí či zbraně.

[45] Jsem si vědom vágnosti tohoto vymezení, avšak např. ani Culpeper (2009, s. 35n.) není při diskutování této problematiky konkrétnější.

cíle výzkumu rozhodl využít skutečnosti, že dva konkrétní typy Klausových textů jsou již v rámci ČNK obsaženy a zpracovány ve dvou specializovaných korpusech.

Analýza klíčových slov v korpusech *ParlCorp* a *Speeches*

Pro zkoumání Klausova idiolektu pomocí analýzy klíčových slov mají korpusy *ParlCorp* a *Speeches* (viz kap. 3.1) dvě hlavní přednosti. Za prvé, každý z korpusů obsahuje vždy texty pouze jednoho žánru (projev) vzniklé vždy ve víceméně stejné komunikační situaci (sněmovní jednání; vystoupení k významnému výročí či svátku). A za druhé, v každém korpusu byly všechny texty zpracovány (anotovány) stejným způsobem. Výsledky analýzy klíčových slov „uvnitř" těchto korpusů by tak měly podat spolehlivější informace přímo o Klausově idiolektu bez potenciálního zkreslení „vnějšími" vlivy. To by spolu s menším množstvím výsledků (v korpusu *ParlCorp* byly provedeny tři analýzy, v korpusu *Speeches* dvě) mělo umožnit vyvození adekvátních kvalitativních závěrů, byť za cenu pokrytí jen relativně úzkého výseku Klausova idiolektu.

V korpusu *ParlCorp* jsem podle strukturních atributů *sp.speaker* (mluvčí), *doc.period* (období) a *sp.role* (funkce/role) vytvořil celkem čtyři subkorpusy: *VKst–premiér*, *VKst–předseda PSP*, *VKst–prezident* a *!VKst*. Václav Klaus starší a jeho syn Václav Klaus mladší, který byl poslancem v letech 2017–2021, sdílejí u atributu *sp.speaker* stejnou hodnotu *Václav Klaus*. Subkorpusy projevů Václava Klause staršího jsem tedy kromě hodnot atributů *sp.speaker* a *sp.role* vymezil též ještě hodnotami *1993–1996*, *1996–1998*, *1998–2002* a *2002–2006* u atributu *doc.period*, aby bylo zajištěno, že jeho subkorpusy nebudou zahrnovat také texty Václava Klause mladšího. Subkorpusy Václava Klause staršího tak sestávají z jeho projevů na půdě PSP ve funkcích předsedy vlády (111 527 pozic včetně interpunkce), předsedy PSP (48 842 pozic včetně interpunkce) a prezidenta (4 009 pozic včetně interpunkce), a to konkrétně od 1. ledna 1993 do 16. října 2003. Subkorpus *!VKst* byl vytvořen pomocí negativní podmínky *!within sp.speaker="Václav Klaus"*, a skládá se tedy z textů všech ostatních mluvčích kromě Václava Klause staršího a Václava Klause mladšího, jehož texty nemohly být při splnění této negativní podmínky zahrnuty. Celkem tento subkorpus obsahuje 38 367 780 pozic včetně interpunkce.

Obdobně jsem v korpusu *Speeches* podle strukturního atributu *doc.speaker* (mluvčí) s hodnotou *Klaus, Václav* a podle atributu *doc.occasion* (příležitost projevu) vytvořil subkorpus *VK–NYA* (12 613 pozic včetně interpunkce) zahrnující konkrétně jedenáct

novoročních projevů (nejstarší z nich, z roku 1993, Klaus přednesl jakožto předseda vlády v zastoupení prezidenta) a subkorpus *VK–RD* (14 037 pozic včetně interpunkce) zahrnující deset projevů ke dni vzniku samostatného československého státu 28. října. Dále jsem vytvořil subkorpus *!VK–NYA/CM* (147 550 pozic včetně interpunkce) označením všech ostatních hodnot atributu *doc.speaker* a hodnot *Christmas mesaage* a *New Year's Address* u atributu *doc.occasion* (vánoční poselství a novoroční projev lze považovat za varianty téhož) a se stejnými mluvčími ještě též subkorpus *!VK–RD* (72 246 pozic včetně interpunkce), aby byly v subkorpusech zachovány jen stejné typy textů jako v Klausových subkorpusech.

V obou korpusech jsem analýzy klíčových slov provedl na základě zadání vyhledávacího dotazu s následujícími parametry:

parametr	**hodnota**
Atribut	lemma [*ParlCorp*] / lemma_lc [*Speeches*]
Hledaný výraz	.*
Třídit podle	DIN (Difference index)
Min. frekvence	5
Max. frekvence	—

Analýza slovesných tvarů

Pro analýzu slovesných tvarů jsem se rozhodl pro srovnávání korpusu *VK* s korpusem *SYN2020*, což je aktuálně poslední vydaný korpus řady *SYN*, a to především z důvodu jeho reprezentativnosti ve smyslu pokrytí všech variet psané češtiny[46] (také spolu s důvodem, že se jedná o první a zatím jeden z mála korpusů ČNK obsahujících poziční atribut *verbtag*, který umožňuje provést analýzu slovesných tvarů výrazně snadněji a více do hloubky). Korpus *SYN2020* obsahuje cca 100 milionů pozic bez interpunkce a je rovnoměrně rozdělen na třetiny mezi tři skupiny textových typů (strukturní atribut *txtype_group*): beletrii, oborovou literaturu a publicistiku (heslo *Korpus SYN2020*, in: *Příručka ČNK*). Takováto reprezentativnost referenčního korpusu je při zkoumání klíčovosti konkrétního gramatického rysu důležitější než jeho celková velikost, neboť v něm lze předpokládat

[46] Různá pojetí reprezentativnosti jsou pojednána v *Příručce ČNK* (heslo *Reprezentativnost korpusu*, in: *Příručka ČNK*).

zastoupení úplnějšího počtu gramatických struktur vyskytujících se ve všech typech psaných komunikátů.

Celkem jsem provedl deset analýz klíčových slovesných tvarů — analýzu celého korpusu *VK*, analýzy pěti subkorpusů vytvořených podle hodnot strukturního atributu *doc.period* a analýzy čtyř subkorpusů vytvořených podle hodnot strukturního atributu *doc.registr* (viz výše). Jak celý korpus, tak všech devět subkorpusů jsem srovnával vždy s celým korpusem *SYN2020*, neboť na základě zkušebních analýz se ukázalo, že srovnávání vždy se stejným referenčním korpusem poskytuje hodnotnější informaci o proměnách Klausova užívání konkrétního gramatického rysu. Hodnoty klíčovosti slovesných tvarů se totiž při analýzách ukázaly být volbou referenčního korpusu podmíněny mnohem méně, než tomu bylo u hodnot klíčovosti lexikálních jednotek.

Parametry dotazu na analýzu slovesných tvarů byly pro všech deset provedených analýz následující:

parametr	**hodnota**
Atribut	verbtag
Hledaný výraz	.*
Třídit podle	DIN (Difference index)
Min. frekvence	5
Max. frekvence	—

Během zkušebních analýz byla odhalena technická chyba, kdy se s nastaveným atributem *verbtag* při analýzách subkorpusů (nikoliv však celých korpusů) počítala dvojnásobná absolutní frekvence, což ovlivňovalo také výsledné hodnoty DIN. Na odstranění chyby se pracuje, avšak díky vývojářům ÚČNK, kteří mi dotčené subkorpusy vytvořili jako samostatné „korpusy“, jsem mohl analýzy provést bez rizika zkreslení výsledků.

Analýza afixace

Zkoumání Klausovy afixace jsem založil na srovnání celého korpusu *VK* s celým korpusem *SYN v12*, protože pro dosažení cíle této analýzy je vhodné, aby byly zkoumaný i referenční korpus co největší. Zkoumaný korpus by měl být co největší z toho důvodu, že mi při

analýzách klíčovosti vybraných slovotvorných prostředků (prefixu *ne-* a sufixu *-ismus*) nejde o časovou ani žánrovou specifičnost. Lze totiž předpokládat, že i přes vysoké souhrnné frekvence každého afixu mohou být absolutní frekvence jejich konkrétních realizací (slovních tvarů či lemmat) velmi nízké. S tím souvisí hypotetická možnost, že by se např. jeden konkrétní klausismus vyskytoval v každém z pěti subkorpusů odpovídajících jednotlivým obdobím právě dvakrát — při nastavení minimální frekvence na hodnotu 5 by se tak ve výsledcích objevil pouze při analýze celého korpusu (a dokonce by mohl by mít vysokou hodnotu DIN), avšak neobjevil by se ve výsledcích analýzy žádného ze subkorpusů. Referenční korpus by měl být co největší především proto, aby pokryl stejné (nebo i větší) časové období jako zkoumaný korpusu, čímž se zvýší šance, že se v něm potenciální klausismy vyskytnou také, a nebudou tak ve výsledcích přeceněny na hodnotu DIN 100. Tento trend se potvrdil porovnáním výsledků zkušebních analýz, kdy byly (při zachování stejných parametrů dotazu) jako referenční korpusy použity korpusy *SYN v12* a *SYN2020*. Zatímco při srovnání s korpusem *SYN v12* obsahovaly výsledky nula (prefix *ne-*) a dvě (sufix *-ismus*) položky s hodnotou DIN 100, při srovnání s korpusem *SYN2020* výsledky obsahovaly šestnáct (prefix *ne-*) a deset (sufix *-ismus*) položek s hodnotou DIN 100; kromě těchto položek s absolutní klíčovostí byl zbytek výsledků srovnatelný. Tyto zkušební analýzy tak ukázaly, že hodnota DIN 100 je situací „zvláštní, která je vždy hodna speciálního pozoru“ (heslo *DIN*, in: *Příručka ČNK*).[47]

Kvůli velké produktivnosti prefixu *ne-* (heslo *ne-*, in: Šimandl 2016, s. 385–387) se pochopitelně v Klausových textech vyskytuje také mnoho běžných negovaných slov, které nelze označit za klausismy, což může potenciálně komplikovat interpretaci výsledků analýz. Pokud však klausismy chápeme jako slova pro Klause charakteristická, nikoliv nutně typická ve smyslu frekvence (z lexikologického hlediska bychom je mohli označit za okazionalismy či neologismy vymezené osobou autora; Martincová 2017), pak lze předpokládat, že je automatická lemmatizace při budování korpusu *VK* označila za samostatná lemmata začínající sekvencí písmen „ne“, nikoliv za prefigované tvary jiných lemmat. Tento předpoklad jsem ověřil testováním pomocí pokročilých dotazů v modulu *Konkordance*. První dotaz byl na všechny slovní tvary začínající sekvencí „ne“, jejichž lemmata také začínají sekvencí „ne“ a které zároveň nejsou v superlativu (kvůli předponě *nej-*), tento dotaz měl podobu *[lc="ne.*" & lemma_lc="ne.*" & tag!=".....3.*"]*. Druhý dotaz byl na

[47] Hodnota DIN 100 navíc neodráží, jak velké mohou být u dotčených položek rozdíly absolutních frekvencí ve zkoumaném korpusu, což může být poměrně zásadní informace.

všechny slovní tvary začínající sekvencí „ne“, jejichž lemmata však nezačínají sekvencí „ne“ a které zároveň nejsou v superlativu, tento dotaz měl podobu *[lc="ne.*" & lemma_lc!="ne.*" & tag!=".....3.*"]*. Srovnání seznamů lemmat získaných těmito dotazy (byť ze své podstaty subjektivní) uvedený předpoklad potvrdilo — prakticky žádné tvary intuitivně pociťované jako klausismy (s absolutní frekvencí alespoň 3) nebyly přiřazeny k lemmatům nezačínajícím sekvencí „ne“.

Analýzu klíčových slov s prefixem *ne-* jsem tedy provedl pomocí dotazu s následujícími parametry:

parametr	**hodnota**
Atribut	lemma_lc
Hledaný výraz	ne.*
Třídit podle	DIN (Difference index)
Min. frekvence	3
Max. frekvence	—

Minimální frekvenci jsem stanovil na hodnotu 3, která je nižší než 5 (z již popsaných důvodů), avšak vyšší než 2, aby byly eliminovány případné hapaxy, které by se mohly objevit kvůli některým textům (nebo pasážím textů), jež se na Klausově webu (a tím pádem i v korpusu *VK*) vyskytují dvakrát.

Sufix *-ismus*[48] funguje v systému češtiny specifičtěji než prefix *ne-*, a tudíž analýza klíčových slov, jež ho obsahují, i interpretace jejích výsledků s sebou nenese tolik komplikací. Analýza byla provedena na základě dotazu s následujícími parametry:

parametr	**hodnota**
Atribut	lemma_lc
Hledaný výraz	.*i[sz]m.*
Třídit podle	DIN (Difference index)
Min. frekvence	3
Max. frekvence	—

[48] Analýza počítá též s pravopisnou variantou *-izmus* (heslo *-ismus/-izmus*, in: Šimandl 2016, s. 278–280), ta se však v korpusu *VK* vyskytuje jen celkem devětkrát v osmi různých tvarech, z čehož dva tvary jsou polská slova. Zbylých šest tvarů odpovídá třem lemmatům, která se píší s písmenem „s“.

Hledaný výraz *.*i[sz]m.** (a nikoliv *.*i[sz]mus*) byl zvolen opět s ohledem na nedokonalou lemmatizaci klausismů vyskytujících se v korpusu *VK* v jiných tvarech než v nominativu singuláru, kdy např. každému z tvarů *rightismus*, *rightismu* a *rightismem* bylo přiřazeno samostatné lemma.

Diskurzní analýza

Jak již bylo zmíněno v kap. 1.2, pro diskurzní analýzu jsem zvolil texty vytvořené pro jednu pravidelně se opakující událost, díky čemuž tyto texty sdílejí stejné „vnější" charakteristiky (s výjimkou data vzniku). Takto úzce vymezený soubor textů umožňuje sledovat, jak se v průběhu let proměňovala Klausova slovní zásoba, respektive klíčová slova, v jedné konkrétní komunikační situaci. Jako případovou studii jsem vybral projevy na setkáních *Euro Business Breakfast* (EBB) pořádaných pravidelně (aktuálně jednou měsíčně) ekonomickým časopisem *Euro*, kde Klaus vystupuje již po mnoho let téměř bez výjimek jednou ročně. Na Klausově webu je publikováno (a tedy obsaženo v korpusu *VK*) celkem šestnáct projevů z EBB, a to z období let 2005–2023.

Pro účely diskurzní analýzy byla klíčová slova analyzována vzájemným srovnáváním jednotlivých subkorpusů v korpusu *VK*. Celkem jsem vytvořil 32 subkorpusů, které mezi sebou fungují jako 16 dvojic. Subkorpusy byly vytvořeny pomocí podmínky *within doc.id* a názvů konkrétních dokumentů (souborů) obsahujících dané projevy, které jsem dohledal pomocí vlastní tabulky obsahující metadata korpusu (viz kap. 3.1). Každý jeden ze šestnácti projevů na EBB tvoří právě jeden subkorpus nazvaný v podobě „EBB-yyyy:mm:dd", kde „yyyy:mm:dd" označuje datum, kdy byl projev přednesen, ve formátu ‚rok:měsíc:den'. Dalších šestnáct subkorpusů pojmenovaných v podobě „EBB-yyyy:mm:dd!" se skládá vždy z ostatních patnácti projevů, než je projev odkazovaný v názvu daného subkorpusu. Subkorpusy mezi sebou byly srovnávány podle následujícího klíče:

zkoumaný subkorpus	**referenční subkorpus**
EBB-2005:12:05	*EBB-2005:12:05!*
...	...
EBB-2023:04:04	*EBB-2023:04:04!*

Postupně každý ze šestnácti zkoumaných subkorpusů byl analyzován na základě dotazu s následujícími parametry:

parametr	hodnota
Atribut	lemma_lc
Hledaný výraz	.*
Třídit podle	DIN (Difference index)
Min. frekvence	2
Max. frekvence	—

Při volbě pozičního atributu jsem nejprve provedl analýzy všech subkorpusů s nastaveným atributem *lc*, protože konkrétní slovní tvary mohou hrát při diskurzní analýze významnou roli. Jelikož však v předkládaném výzkumu není cílem na získané výsledky navázat komplexní kvalitativní analýzou diskurzu, při níž bych do hloubky zkoumal např. konkordanční řádky či kolokace všech identifikovaných klíčových slov, nebyly výsledky pro související vedlejší cíl (zda jsou identifikovaná klíčová slova podmíněna pouze tématy projevů, anebo zda vypovídají také o Klausově diskurzní praxi) příliš uspokojivé. Proto jsem provedl analýzy všech subkorpusů znovu s nastaveným atributem *lemma_lc*, přičemž výsledky těchto analýz se ukázaly být pro dosažení vytyčeného cíle vhodnější. Minimální frekvenci jsem nastavil na hodnotu 2, tedy nejnižší možnou bez započtení hapaxů, neboť zkoumané subkorpusy jsou velmi malé (od 929 do 2 458 pozic včetně interpunkce).

4 Praktická část

V této kapitole prezentuji kvantitativní výsledky analýz, provedených během výzkumu podle metodologických procedur popsaných v kap. 3.2. Praktická část je rozdělena na tři podkapitoly, které jsou dále děleny na oddíly. První dvě podkapitoly souvisejí s hlavním cílem výzkumu. Nejprve jsou představeny výsledky analýz klíčových slov v korpusu *VK*, kde je každý z pěti oddílů věnován jednomu subkorpusu podle období Klausova veřejného působení. Následuje podkapitola s výsledky analýz klíčových slov ve dvou specializovaných korpusech dostupných v rámci ČNK, kde je první oddíl věnován korpusu sněmovních projevů *ParlCorp* a druhý oddíl korpusu prezidentských projevů *Speeches*. Třetí podkapitola souvisí s vedlejšími cíli výzkumu a postupně v ní rozebírám výsledky dalších dílčích analýz Klausova idiolektu, konkrétně slovesných tvarů, afixace a diskurzní analýzy jeho projevů na setkáních *Euro Business Breakfast*.

Aby se se získanými výsledky dalo lépe nakládat a jejich konečná podoba byla přehlednější, byly výsledky všech 48 analýz klíčových slov provedených v rozhraní *KonText* exportovány a uloženy jako jednotlivé XLSX soubory. S těmito soubory jsem dále pracoval v programu *Microsoft Excel*, kde jsem listy s výsledky jednotlivých analýz seskupoval do souborů (sešitů) odpovídajících souvisejícím podkapitolám a oddílům praktické části. Pro účely adekvátních interpretací výsledků byly v některých sešitech přidány další listy. Všechny provedené úkony jsou popsány a vysvětleny na patřičných místech následujících podkapitol, respektive oddílů. Výsledné soubory jsou přístupné k prohlížení a stažení na adrese https://osf.io/7btnc/?view_only=95c8e2f540764097b22ef8effa0de8dd.

4.1 Analýza klíčových slov v *Korpusu textů Václava Klause*

Před prezentací vlastních výsledků analýz je nutné zmínit jednu technickou komplikaci, která ovlivnila část výsledků prezentovaných v této podkapitole, a popsat změny v metodologii, kterými jsem se snažil dopady této komplikace minimalizovat.

Veškeré analýzy pro předkládaný výzkum byly prováděny v rozhraní *KonText* s využitím modulu *Analýza klíčových slov*. Tento modul je zveřejněn teprve krátkou dobu a průběžně prochází dalším vývojem. Přestože je tedy již převážně funkční a registrovaní uživatelé ČNK jej mohou využívat k různým účelům (jak se ostatně snaží ukázat tato diplomová práce), nadále v něm může docházet k jistým technickým chybám, které se zatím

nepodařilo odhalit, případně odstranit, jako je např. již zmiňovaný problém s dvojnásobnými absolutními frekvencemi v subkorpusech při nastavení atributu *verbtag*.

Předkládaný výzkum byl nejvíce ovlivněn jiným technickým nedostatkem týkajícím se také subkorpusů. V teoretické i metodologické části této práce bylo ukázáno, jak důležitá je při analýze klíčových slov volba vhodného referenčního korpusu. Nabídka korpusů zpřístupněných v rámci ČNK je pro účely předkládaného výzkumu dostatečně široká a výběr referenčního korpusu užitečného pro dosažení hlavního cíle výzkumu nepředstavoval problém. Aby však byly získané výsledky z hlediska tohoto cíle opravdu relevantní, bylo nutné ve zvoleném korpusu *SYN v12* vytvářet subkorpusy, které budou v určitých ohledech (období a registr) blízké konkrétním zkoumaným subkorpusům vytvořeným v korpusu *VK* (viz kap. 3.2).

Zatímco při srovnávání s referenčními subkorpusy v rámci registrů *administrativa* a *oborová literatura* proběhly všechny analýzy bez problémů, při srovnávání s referenčními subkorpusy v rámci registru *publicistika* vždy došlo k chybě při výpočtu pomocných dat, kvůli které tyto analýzy vůbec neproběhly. Zdá se, že příčinou není velikost referenčního subkorpusu jako taková (některé subkorpusy v rámci registru *oborová literatura* jsou větší než některé subkorpusy v rámci registru *publicistika*), avšak celkový počet všech tokenů hodnoty strukturního atributu použité pro vymezení subkorpusu (hodnota *NMG: publicistika* strukturního atributu *doc.txtype_group* zahrnuje v korpusu *SYN v12* celkem přes 5,5 miliardy tokenů).

Po testování alternativních řešení s různými jinými referenčními subkorpusy (především z jednotlivých korpusů řady *SYN*) jsem se nakonec rozhodl subkorpusy v rámci registrů *publicistika* a *veřejná vystoupení* v korpusu *VK* srovnávat s celým korpusem *SYN v12* jakožto referenčním korpusem. Toto řešení samozřejmě není dokonalé, především kvůli příliš širokému časovému rozsahu (nejstarší texty v korpusu pocházejí z konce 50. let 20. století, avšak celkově je zde textů z doby před 90. lety jen zlomek) a kvůli zahrnutí registrů *beletrie* a *oborová literatura* (oba registry dohromady však svým rozsahem odpovídají cca jen jedné osmině rozsahu publicistiky, což je v tomto případě příhodné). Volba celého korpusu *SYN v12* jako referenčního korpusu však zachovává dva aspekty důležité pro srovnávání. Za prvé, zkoumaný i referenční korpus jsou anotovány a lemmatizovány podle stejných pravidel, díky čemuž je analýza klíčových slov podle pozičního atributu *lemma_lc* přesnější, než kdyby byly jako referenční korpusy použity starší korpusy řady *SYN*. Za druhé, referenční korpus díky tomu obsahuje také publicistiku

z let 2020–2023, a tak se v něm také vyskytují slovní tvary a lemmata týkající se pandemie covidu-19, které by se při srovnávání s korpusy řady SYN vyskytovaly pouze ve zkoumaných subkorpusech, což by vedlo k absolutnímu zkreslení hodnot klíčovosti těchto jednotek.

V souborech s výsledky pro oddíly v této podkapitole jsem za účelem jejich zpřehlednění a snazší vyvození adekvátních závěrů provedl následující úkony:

1) Listy s výsledky analýz jednotlivých subkorpusů ‚období–registr' jsem sloučil do jednoho sešitu pro celý odpovídající nadřazený subkorpus ‚období'.
2) V každém listu jsem o jeden řádek dolů odsadil všechny položky s hodnotou DIN nižší než 75. Všechny položky s hodnotou DIN 75 a vyšší jsem ohraničil a nadpisy všech sloupců jsem nastavil jako filtry pro řazení ohraničených položek podle parametrů uvedených v nadpisech (defaultně DIN).
3) V každém listu jsem ohraničené položky barevně rozdělil podle jejich hodnot DIN do třech pásem klíčovosti: DIN 100 (první pásmo), DIN 99,99–90,00 (druhé pásmo) a 89,99–75,00 (třetí pásmo). Položky druhého a třetího pásma jsem zkopíroval do stejného listu vedle původní ohraničené tabulky, kde jsem je seřadil sestupně podle hodnot absolutní frekvence.[49]
4) V každém sešitu jsem vytvořil nový list pro celý odpovídající subkorpus ‚období', do něhož jsem vložil všechny položky, které mají hodnotu DIN 75 a vyšší ve více než jednom z odpovídajících subkorpusů ‚období–registr'. Každý z těchto přidaných listů obsahuje tři sloupce: takto vybrané položky, součet jejich absolutních frekvencí z daných subkorpusů a určení, ve kterých subkorpusech má daná položka hodnotu DIN 75 a vyšší (pokud ve všech, pak má tento parametr hodnotu *all*). Nadpisy sloupců jsem nastavil jako filtry pro řazení a položky jsem seřadil sestupně podle hodnot absolutní frekvence.[50]

[49] Tento krok jsem učinil, protože ve značné části subkorpusů bylo identifikováno příliš mnoho lemmat s hodnotou DIN 75 a vyšší (nejvíce v subkorpusu *exprezident – publicistika*, konkrétně 895 položek), což výrazně stěžovalo kvalitativní vyhodnocení výsledků. Díky tomuto kroku jsou však v každém listu viditelné frekvenční špičky všech pásem klíčovosti najednou. To poskytuje informace jak o klíčovosti, tak o absolutních frekvencích, přičemž kombinace obou informací je pro adekvátní vyhodnocení výsledků stěžejní.

[50] K tomuto kroku jsem přistoupil, aby byla dostupná informace o disperzi klíčovosti daných lemmat a aby bylo možné výsledky dále třídit, respektive filtrovat podle jejich rovnoměrného rozložení, zajištěného alespoň tímto základním způsobem (heslo *Frekvence*, in: *Příručka ČNK*).

Pro účely celé této podkapitoly (a tedy i pro zodpovězení hlavní výzkumné otázky) jsem vytvořil nový XLSX soubor (sešit), nazvaný *Kap. 4.1*, do něhož jsem z jednotlivých sešitů pro subkorpusy ‚období' zkopíroval listy zachycující disperzi klíčových lemmat, vytvořené podle výše popsaného čtvrtého kroku. Poté jsem na základě podobného postupu, avšak aplikovaného již na listy zachycující disperzi, vytvořil v tomto sešitu čtyři další listy: *2+_key_periods* obsahuje lemmata, jež se vyskytují ve dvou a více listech ‚období' zachycujících disperzi (299 lemmat), *3+_key_periods* ve třech a více (115 lemmat), *4+_key_periods* ve čtyřech a více (37 lemmat), a konečně *all_key_periods* obsahuje ta lemmata, která jsou klíčová vždy alespoň ve dvou subkorpusech ‚období–registr' v každém z pěti subkorpusů ‚období'. Těchto lemmat je celkem devět, a jedná se o následující (v závorkách jsou uvedeny součty jejich hodnot absolutní frekvence ze všech subkorpusů ‚období–registr', v nichž mají hodnotu DIN 75 a vyšší):

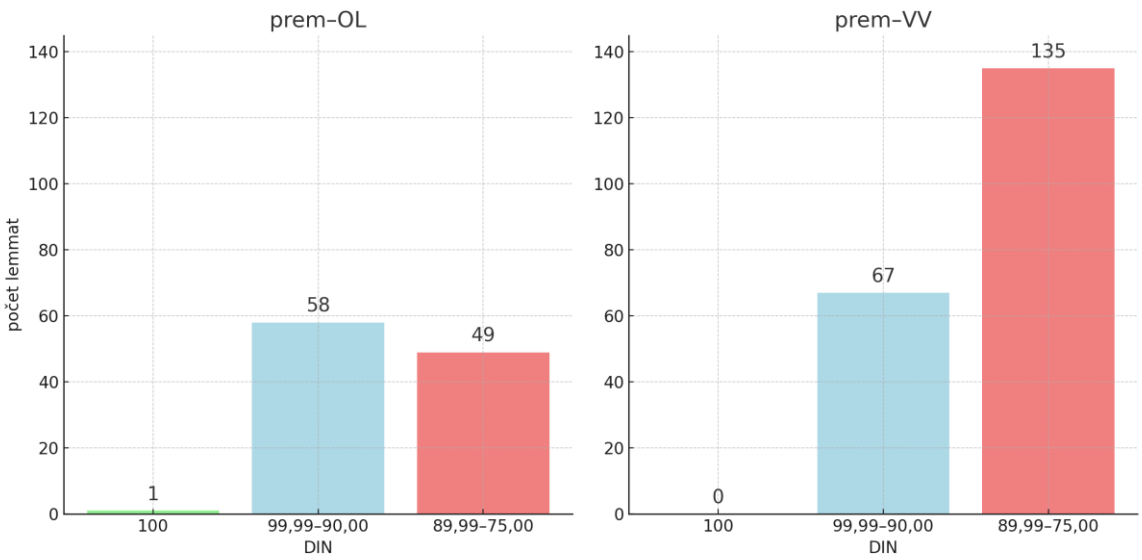
> *politický* (4 066), *ekonomický* (3 325), *ekonomika* (2 444), *měnový* (1 157), *ekonom* (880), *ekonomie* (672), *transformace* (621), *deficit* (473), *HDP*[51] (290).

Z kvantitativního hlediska zaměřeného na klíčovost – nikoliv na méně vypovídající absolutní a relativní frekvence – lze tyto lexikální jednotky označit za charakteristické pro idiolekt Václava Klause napříč časem, veřejnými funkcemi i registry.

4.1.1 Subkorpus *premiér*

Subkorpus *premiér* se jako jediný ze subkorpusů odpovídajících jednotlivým obdobím neskládá z textů všech čtyř registrů, avšak pouze z textů registrů *oborová literatura* a *veřejná vystoupení*. Počty lemmat v jednotlivých pásmech klíčovosti obou subkorpusů zobrazuje *Graf 1*.

[51] Vzhledem k použití pozičního atributu *lemma_lc* pro analýzy se ve výsledcích objevuje v podobě *hdp*, avšak všechny výskyty této zkratky v celém korpusu *VK* mají pochopitelně podobu *HDP*. Toto platí také pro dále uváděné zkratky typu *ODS* apod.



Graf 1 – Subkorpus *premiér*

Lemmat, která jsou klíčová v obou těchto subkorpusech, je celkem 40. Již letmý pohled na 10 nejfrekventovanějších z nich dává představu, jaká slovní zásoba je charakteristická pro texty Klausova premiérského období (přesněji v letech 1995–1997), byť jsou tyto výsledky ovlivněny tím, že analyzované texty pocházejí jen ze dvou registrů (v závorkách za lemmaty jsou zde i v následujících oddílech uváděny součty hodnot absolutní frekvence ze všech odpovídajících subkorpusů ‚období–registr'):

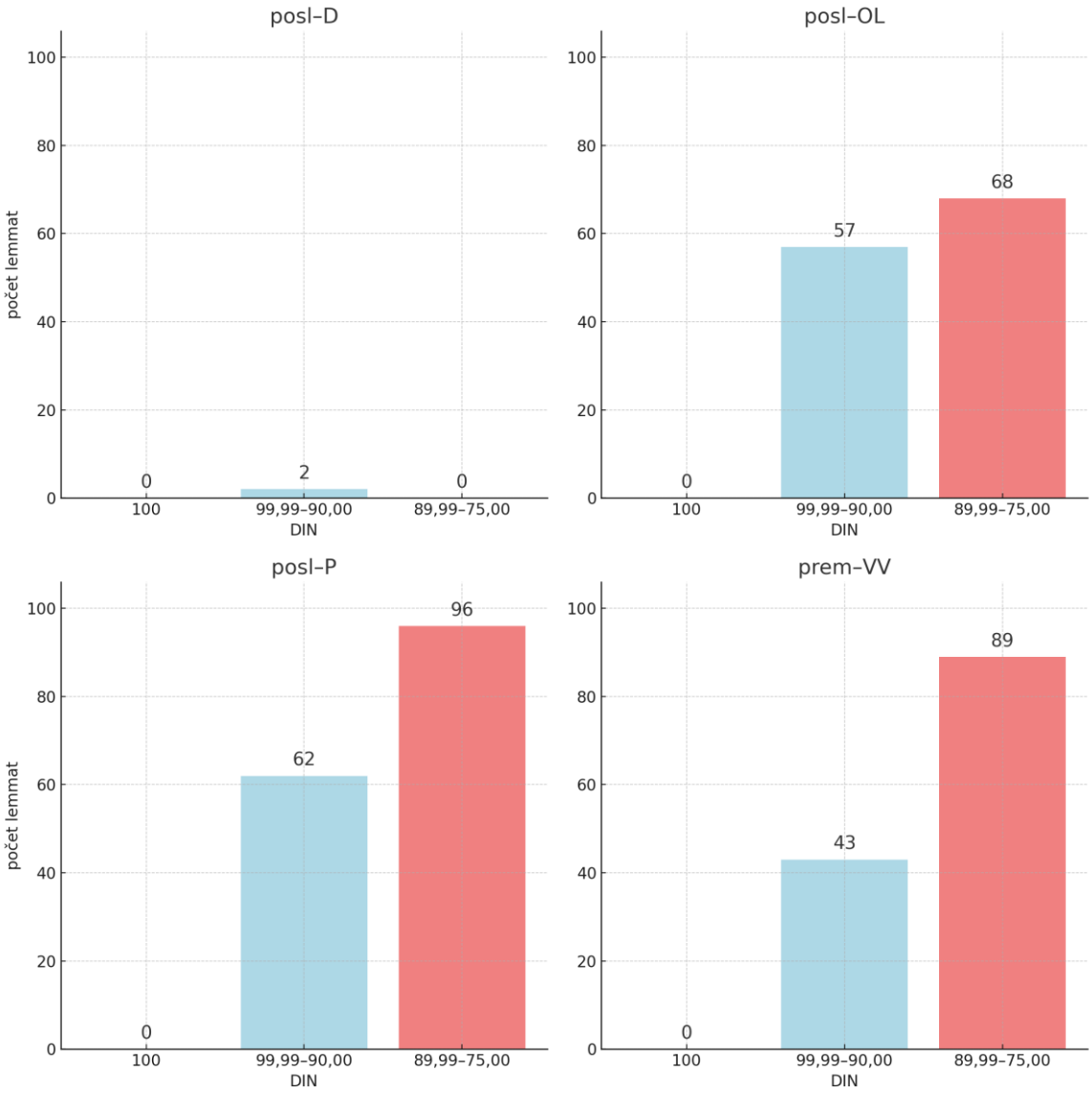
> *ekonomický* (148), *ekonomika* (138), *politický* (86), *politika* (67), *transformační* (47), *měnový* (42), *transformace* (41), *důsledek* (39), *bilance* (37), *privatizace* (37).

V podobném duchu se nese také zbylých 30 lemmat. Kromě jednotek odvozených od lexémů *politika* a *ekonomika* / *ekonomie*, které, jak již bylo ukázáno, patří k nejklíčovějším ve všech subkorpusech ‚období', jsou pro subkorpus *premiér* charakteristická lemmata týkající se polistopadové transformace republiky a privatizace státního majetku.

4.1.2 Subkorpus *poslanec*

Specifičností subkorpusu *poslanec* je fakt, že se jedná o jediný subkorpus ‚období', který není časově souvislý, neboť obsahuje texty z let 1998, 2002 a 2003, a prakticky tedy „obklopuje" subkorpus *předseda PSP*. Přestože subkorpus *poslanec* obsahuje texty všech registrů, jeden z nich – *dokumenty* – je zastoupen jen dvěma texty o celkovém počtu pouhých

159 pozic včetně interpunkce. V takto malém subkorpusu byla identifikována jen dvě klíčová lemmata s hodnotou DIN vyšší než 75, poměrně nepřekvapivě se jedná o lemmata *poslanecký* (DIN 99,87; absolutní frekvence 6) a *sněmovna* (DIN 99,83; absolutní frekvence 5). Počty lemmat v jednotlivých pásmech klíčovosti všech čtyř subkorpusů zobrazuje *Graf 2*.



Graf 2 – Subkorpus *poslanec*

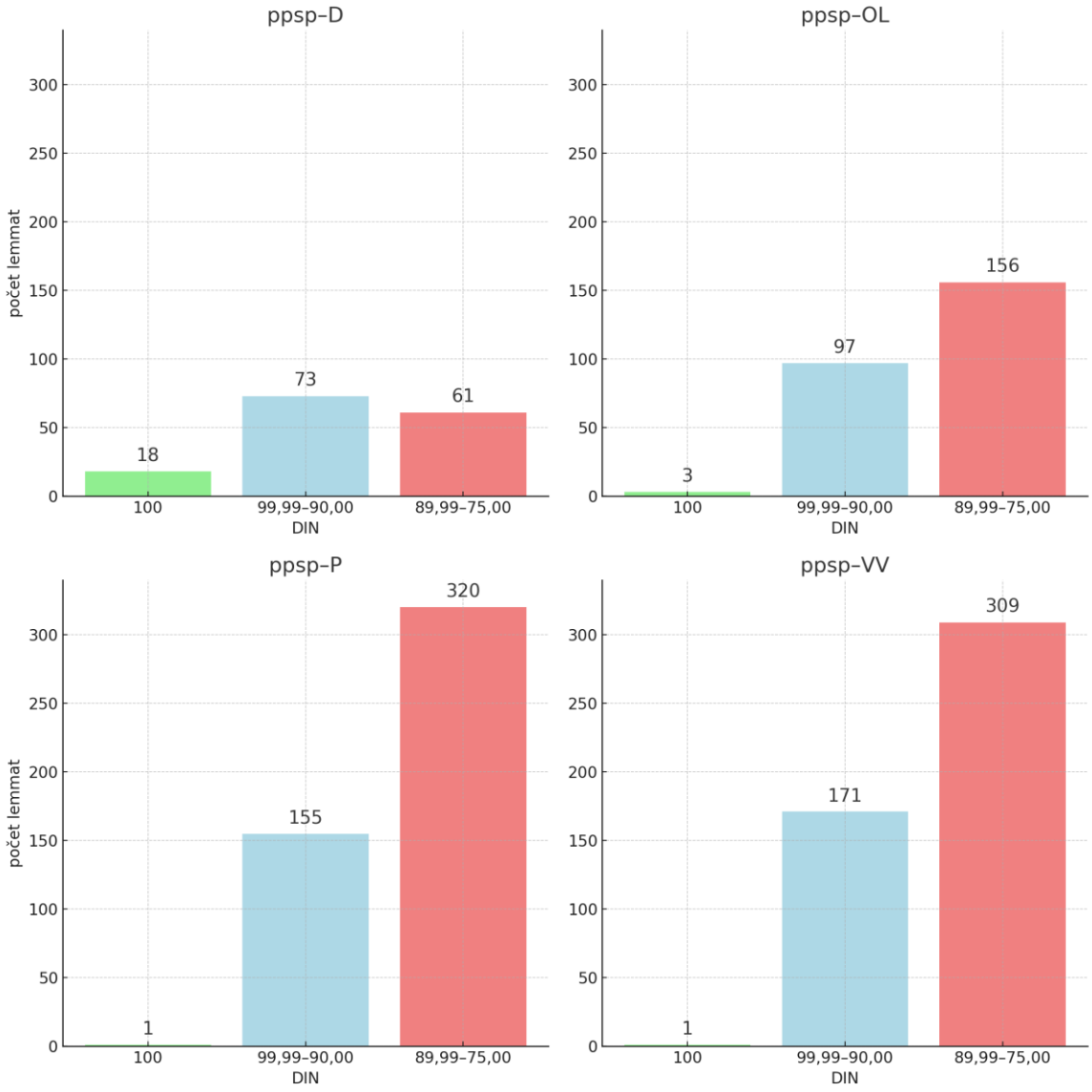
Vzhledem k uvedené zanedbatelné velikosti subkorpusu *poslanec – dokumenty* nejsou žádná lemmata klíčová ve všech čtyřech subkorpusech. Celkem 17 lemmat je však klíčových ve zbylých třech subkorpusech. Následujících 10 je z nich nejfrekventovanějších:

politika (140), *nikoli* (81), *rozpočtový* (74), *deficit* (64), *daňový* (47), *diskuse* (47), *měnový* (45), *komunismus* (41), *realita* (40), *transformace* (36).

Přestože hlavní tematické rámce lemmat zůstávají víceméně zachovány, začíná se projevovat posun zaměření od transformace k otázkám rozpočtu a daní. Za pozornost stojí výskyt synsémantika *nikoli*, jehož poměrně vysoká frekvence může naznačovat, že Klausova tendence k jazykové negaci se neprojevuje pouze slovotvornými prostředky (viz kap. 4.3.2), ale též gramaticko-lexikálními.

4.1.3 Subkorpus *předseda PSP*

Subkorpus *předseda PSP* je výrazně větším subkorpusem ‚období' než předchozí dva, což se odráží také na tom, že je zde poprvé výrazněji zastoupen registr *dokumenty*. Počty lemmat v jednotlivých pásmech klíčovosti všech čtyř subkorpusů ‚období–registr' zobrazuje *Graf 3*.



Graf 3 – Subkorpus *předseda PSP*

Pouze tři lemmata jsou zastoupena ve všech čtyřech uvedených subkorpusech jako klíčová:

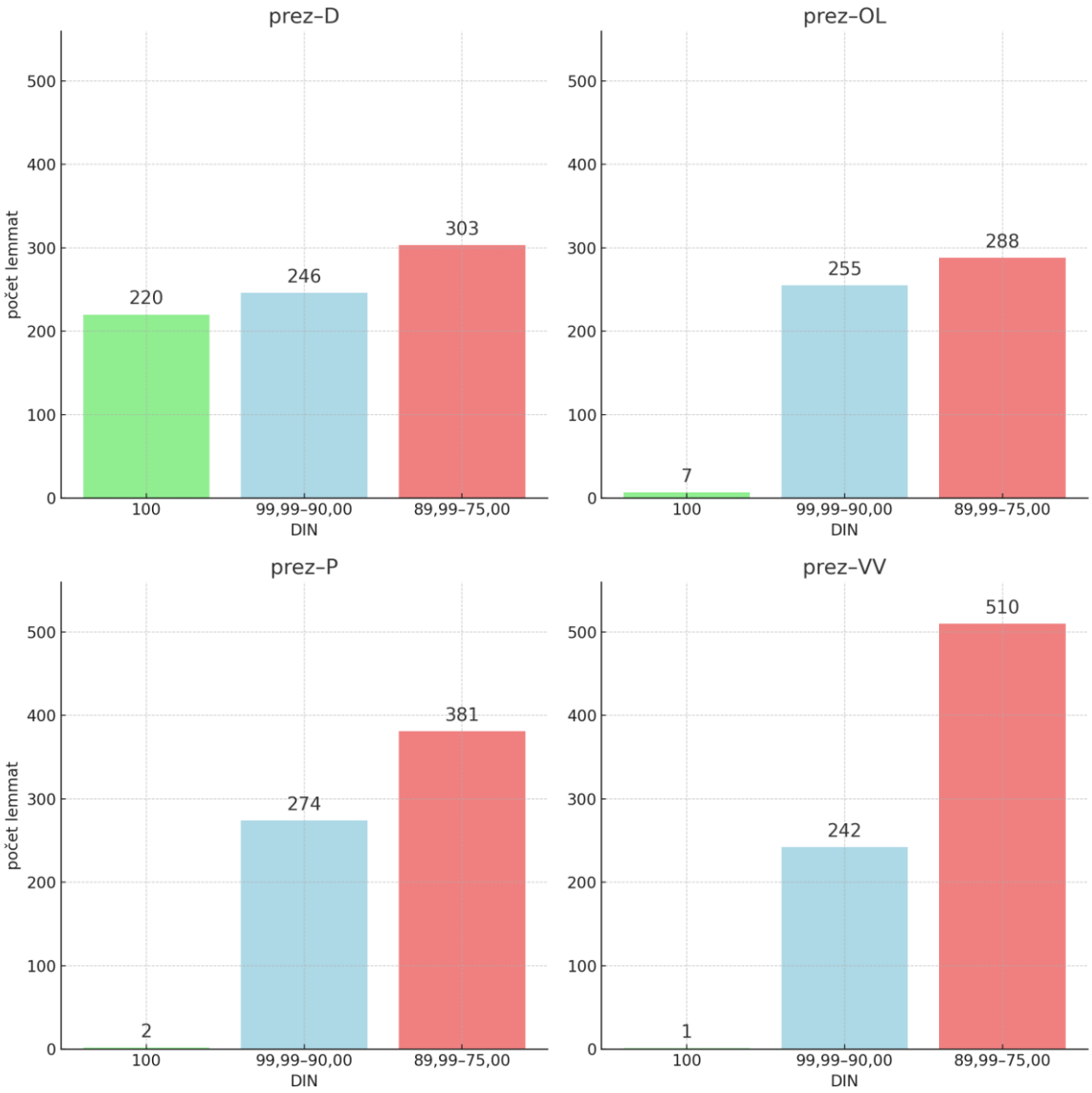
politický (979), *dnešní* (483), *demokratický* (197).

Společným klíčovým lemmatem v registrech *dokumenty*, *oborová literatura* a *publicistika* je *politik* (234), v registrech *dokumenty*, *oborová literatura* a *veřejná vystoupení* je to *ODS* (277) a v registrech *dokumenty*, *publicistika* a *veřejná vystoupení* se jedná o *nikoli* (243), *postoj* (202), *onen* (180), *svobodný* (129), *nepochybně* (59), *pozorně* (32), *odlišnost* (24). Nejvíce společných klíčových lemmat, konkrétně 63, mají registry *oborová literatura*, *publicistika* a *veřejná vystoupení*, přičemž zdaleka nejfrekventovanějšími jsou v nich opět lemmata týkající se politiky, ekonomiky i ekonomie a transformace.

4.1.4 Subkorpus *prezident*

Jak bylo ukázáno v kap. 3.2, subkorpus *prezident* je co do počtu pozic jednoznačně největším ze všech pěti subkorpusů ‚období'. Není tedy divu, že v jeho jednotlivých registrech je identifikováno mnoho klíčových lemmat, což je dáno nejen samotným rozsahem subkorpusu, ale také specifičností některých prezidentských žánrů (textových typů).

Tento fakt se odráží např. v nevídaně vysokém počtu (220) lemmat s absolutní hodnotou klíčovosti DIN 100 v registru *dokumenty*, jak je na první pohled patrné z *Grafu 4*. Na tomto příkladu lze ilustrovat úskalí při výběru vhodného referenčního korpusu pro srovnávání s různými typy zkoumaných korpusů. Všechny zkoumané subkorpusy v rámci registru *dokumenty* byly srovnávány se subkorpusy v rámci registru *administrativa*. Zatímco v ostatních subkorpusech ‚období' fungovalo toto párování relativně přiměřeně, zde došlo k jednoznačnému zkreslení. Příčinou je, že z celkového počtu 392 textů v subkorpusu *prezident – dokumenty* představuje 378 z nich dopisy (zbylých 14 jsou tisková sdělení), přičemž se většinou jedná o kondolence, blahopřání apod. Naproti tomu většinu referenčního subkorpusu *2003–2013 – administrativa* (v korpusu *SYN v12*) tvoří texty legislativní povahy, výroční zprávy apod. Na základě tohoto nesouladu je mimo jiné vidět, že ke zkoumaným (sub)korpusům obsahujícím některé velmi specifické žánry (např. otevřený dopis), se adekvátní referenční korpusy vybírají hůře než k jiným, obecnějším.



Graf 4 – Subkorpus *prezident*

Celkem 45 lemmat je klíčových ve všech prezidentských subkorpusech zároveň. Už při uvedení 10 nejfrekventovanějších z nich je vidět částečný posun směrem od ekonomických témat:

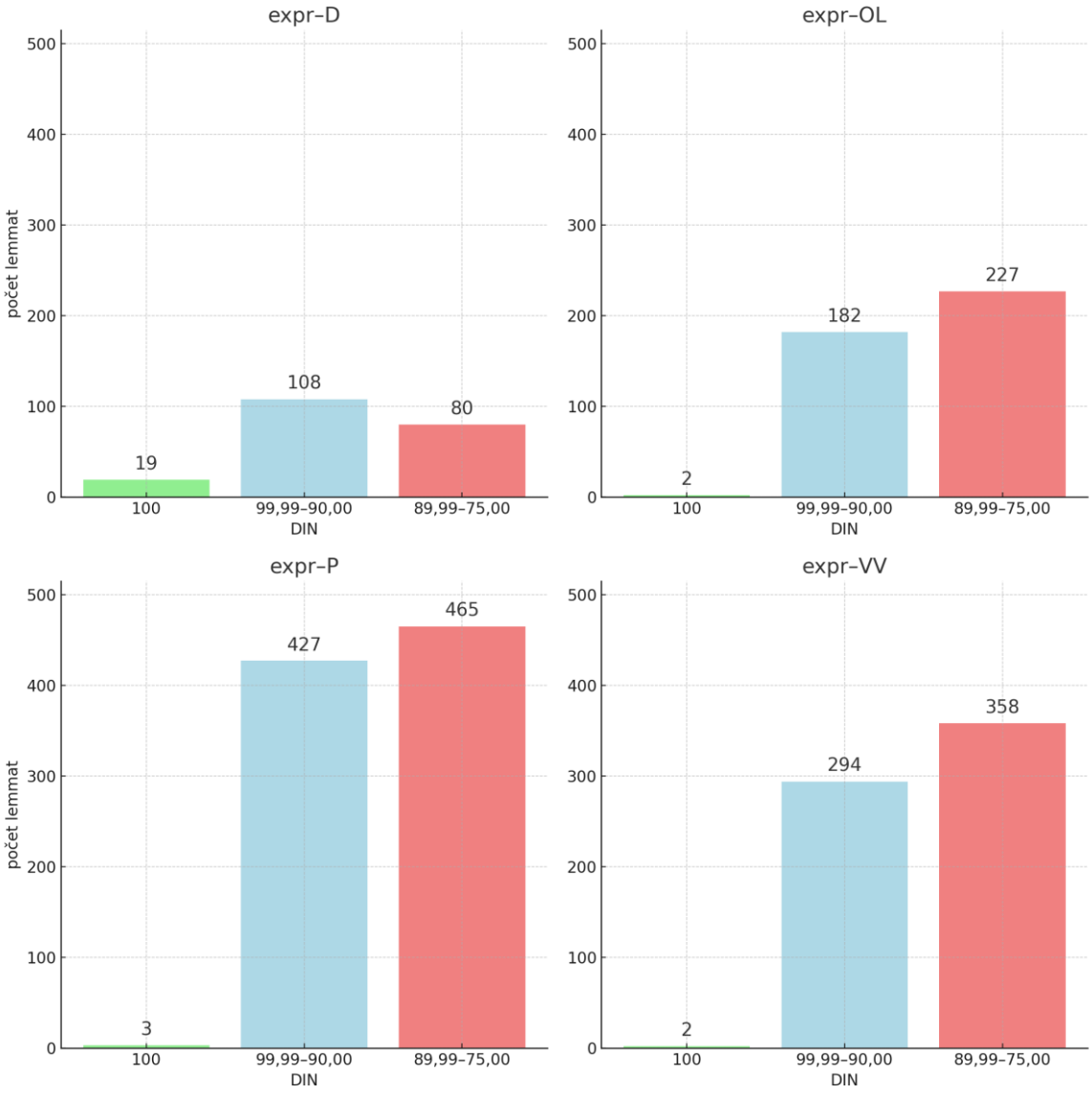
> *politický* (1689), *dnešní* (1231), *nikoli* (748), *demokracie* (627), *demokratický* (526), *ústava* (515), *komunismus* (476), *svobodný* (352), *desetiletí* (327), *komunistický* (321).

Vybraná další z těchto 45 klíčových lemmat se týkají, velmi obecně řečeno, vyjadřování různých názorů (*ideologie*, *racionální*, *argumentace*) či prosazování se a podvolování se (*suverenita*, *totalitní*, *prosazení*). Dvě lemmata odkazují na globální oteplování, respektive změnu klimatu napřímo (*oteplování*, *klimatický*), při pohledu na

kolokace lemmatu *doktrína* jej však lze k tomuto tématu přiřadit také. Ke globálnímu oteplování se sice Klaus vyjadřoval již dříve, avšak až v prezidentském období lze pozorovat, že se pro něj stalo tématem zásadním napříč všemi registry.

4.1.5 Subkorpus *exprezident*

Hlavní specifičností subkorpusu *exprezident* je skutečnost, že se jedná o jediný subkorpus ‚období' s texty, které Klaus již nepublikoval jakožto politik. Tato odlišnost se může potenciálně významně odrazit v charakteristické slovní zásobě daných textů. *Graf 5* poskytuje přehled o rozložení lemmat v jednotlivých pásmech klíčovosti všech čtyř odpovídajících subkorpusů ‚období–registr'.



Graf 5 – Subkorpus *exprezident*

Disperze klíčových lemmat se ukázala být v rámci subkorpusu *exprezident* značně nerovnoměrná, neboť jen 13 jich je klíčových ve všech čtyřech subkorpusech, z nichž položka s nejnižší hodnotou absolutní frekvence je součástí odkazů na webové stránky (*www*). Zbylých 12 je následujících:

> *politický* (1149), *dnešní* (1138), *ekonomický* (1072), *demokracie* (463), *ekonom* (365), *éra* (327), *desetiletí* (302), *výrok* (298), *komunistický* (247), *postoj* (243), *revoluce* (186), *IVK* (126).

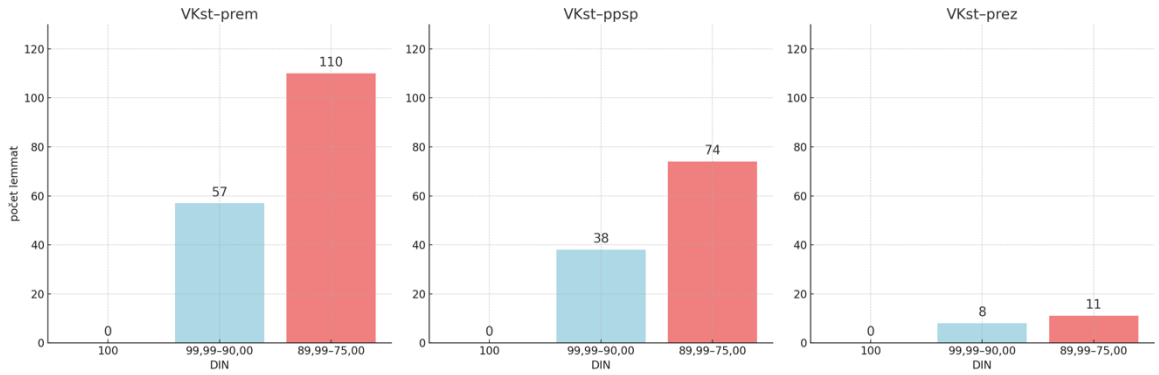
Kromě prakticky všudypřítomných lemmat *politický*, *ekonomický* a *ekonom* zde působí poměrně dominantně lemmata týkající se času, ať už přímo (*dnešní*, *éra*, *desetiletí*), či nepřímo jako odkazy na dobu minulou (*komunistický*, *revoluce*). Předpoklad o významné proměně idiolektu v důsledku ukončení aktivní politické činnosti se pohledem na lemmata s rovnoměrnou disperzí mezi registry nepotvrdil. Lemmat, která jsou klíčová ve více než jednom registru subkorpusu *exprezident*, je však celkem 437, tudíž v některých registrech, především v publicistice, k posunu došlo.

4.2 Analýza klíčových slov v dalších korpusech

4.2.1 Korpus *ParlCorp*

Tři listy s výsledky pro tento oddíl jsem sloučil do jednoho sešitu v souboru *Kap. 4.2.1* a provedl jsem s nimi stejné úkony, jaké byly popsány v kap. 4.1.

Analýza klíčových slov (lemmat) vytvořených subkorpusů Václava Klause staršího v korpusu *ParlCorp* by měla poskytnout především informaci o tom, jaká slovní zásoba charakterizuje Klausovy projevy v PSP, ať už hovořil jako poslanec (a zároveň i předseda vlády či předseda PSP), jako předseda vlády nebo jako prezident. *Graf 6* vzhledem k relativně malému množství identifikovaných klíčových lemmat ve třech vytvořených subkorpusech naznačuje, že by se mohla potvrdit domněnka o relevantních výsledných hodnotách klíčovosti při srovnávání „uvnitř“ jednoho korpusu, kdy by nemělo docházet ke zkreslení vlivem vnějších charakteristik textů.



Graf 6 – Korpus *ParlCorp*

Pohled na konkrétní výsledky oproti předpokladům ale neposkytuje příliš jasnou představu o charakteristikách Klausových sněmovních projevů. Žádné lemma nebylo identifikováno jako klíčové ve všech třech subkorpusech, což sice není vzhledem k malému rozsahu subkorpusu *VKst–prezident* překvapivé, ale o to překvapivější naopak je, že 7 z 8 klíčových lemmat vyskytujících se ve dvou subkorpusech, se nachází právě v prezidentském subkorpusu, a to vždy spolu se subkorpusem *VKst–předseda PSP*. Navíc 6 z těchto lemmat je technické povahy a týká se sněmovních procesů (*prosit*, *zahájit*, *ruka*, *hlásit*, *končit*, *přerušovat*), jedno je příjmení poslankyně (*Dundáčková*). Podobné povahy je také značná část lemmat klíčových vždy jen v jednom ze subkorpusů. Lemmata s vysokou hodnotou klíčovosti navíc často nejsou rovnoměrně rozložena ani v rámci jednoho subkorpusu. Např. 7 z 10 lemmat s nejvyšší hodnotou klíčovosti v subkorpusu *VKst–premiér* (*okurka*, *zintenzívnění*, *faleš*, *antibalíček*, *odsíření*, *pánevní* a *Amsterodam*) se vyskytuje vždy jen v jednom textu.

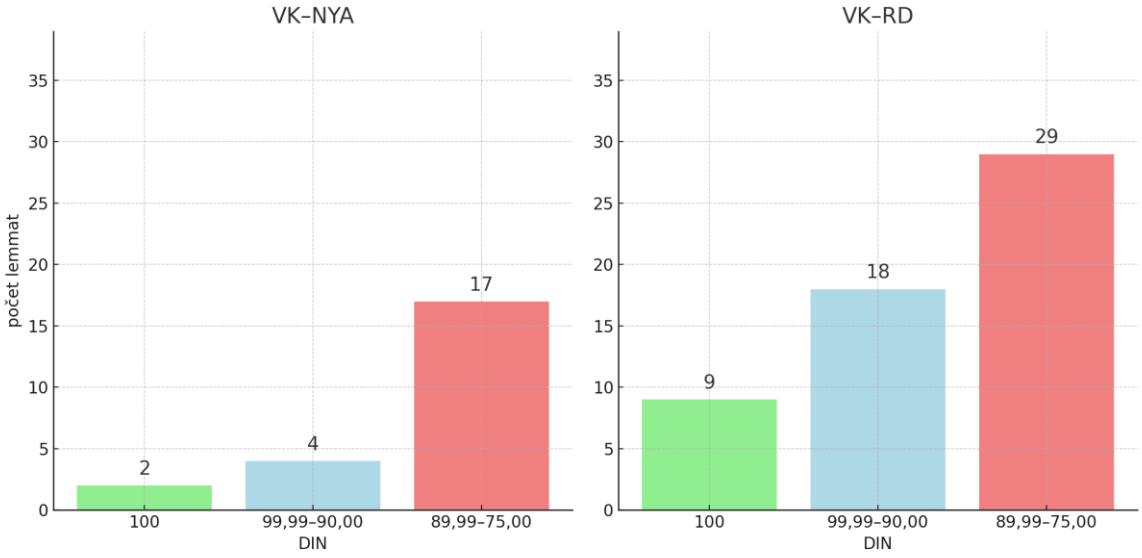
Jediným lemmatem, které se jako klíčové vyskytuje v subkorpusech *VKst–premiér* a *VK–předseda PSP*, je adjektivum *evidentní* se souhrnnou hodnotou absolutní frekvence 36. Při prozkoumání těchto výskytů se ukazuje, že jeho zdaleka nejčastějším kolokátem je lemma *naprosto*, jež mu bezprostředně předchází v 17 z těchto 36 výskytů, což dělá z kolokace *naprosto evidentní* dobrého kandidáta na frázi charakteristickou pro Klausovy sněmovní projevy. Modul *Analýza klíčových slov* sice neumí analyzovat klíčovost víceslovných jednotek, avšak při vyhledání tohoto spojení v modulu *Konkordance* s vymezením všech Klausových textů v korpusu *ParlCorp* je jeho relativní frekvence 103,42 i.p.m., zatímco relativní frekvence téhož spojení v subkorpusu *!VKst* je pouhých 1,3 i.p.m.

I když tedy výsledky analýz klíčových slov v korpusu *ParlCorp* neposkytují o zkoumaném idiolektu tolik informací, jak bylo očekáváno, s využitím doplňujících dotazů se podařilo identifikovat alespoň jednu kolokaci, která je v kontextu sněmovních projevů jednoznačně charakteristická pro Václava Klause.

4.2.2 Korpus *Speeches*

Také pro tento oddíl jsem sloučil dva listy s výsledky do jednoho sešitu (soubor *Kap. 4.2.2*) a provedl jsem s nimi stejné úkony, jaké byly popsány v kap. 4.1.

Podobně jako u analýz klíčových slov v korpusu *ParlCorp*, také o analýze v korpusu *Speeches* se lze domnívat, že by měla poskytnout „čistější“ informaci o charakteristikách Klausovy slovní zásoby v rámci konkrétního typu textu. Rozložení lemmat v pásmech klíčovosti pro subkorpusy Klausových novoročních projevů a projevů 28. října je znázorněno v *Grafu 7*.



Graf 7 – Korpus *Speeches*

Vzhledem k malému rozsahu obou subkorpusů nebylo identifikováno příliš mnoho klíčových lemmat, přičemž pouze jediné lemma bylo identifikováno jako klíčové v obou subkorpusech, a to *unie* (28). Při bližším pohledu na výskyty tohoto lemmatu je vidět, že o Evropské unii se Klaus zmiňoval průběžně v obou typech projevů pravidelně po celou dobu svého prezidentského mandátu. Z žánrového hlediska je poměrně překvapivý výskyt klíčových lemmat s negativními konotacemi v subkorpusu novoročních projevů, konkrétně *bohužel* (9), *dluh* (8), *pokles* (6), *líbivý* (5) a *neúspěch* (5). V subkorpusu projevů 28. října

byla s vysokou hodnotou klíčovosti identifikována lemmata, kterými Klaus popisuje okolnosti a prostředí slavnostního setkání – *sál* (9) *večer* (9), *televizní* (9), *rozhlasový* (8), *posluchač* (5), *Vladislavský* (5) – což značí, že jiní prezidenti při této příležitosti toto nečinili.

4.3 Další dílčí analýzy

4.3.1 Slovesné tvary

Analýza klíčových slovesných tvarů je v celém předkládaném výzkumu metodou, která se od ostatních odlišuje patrně nejvíce. Důvodem je její zaměření na čistě gramatický jev, což má vliv také na to, že zobrazovanými výsledky v tabulce nejsou slovní tvary nebo lemmata, ale hodnoty tzv. *verbtagu*, tedy značky gramatických kategorií slovesa. Tato značka má v rámci ČNK šest pozic: typ slovesa, typ slovesného tvaru, slovesný rod, osobu, číslo a čas.

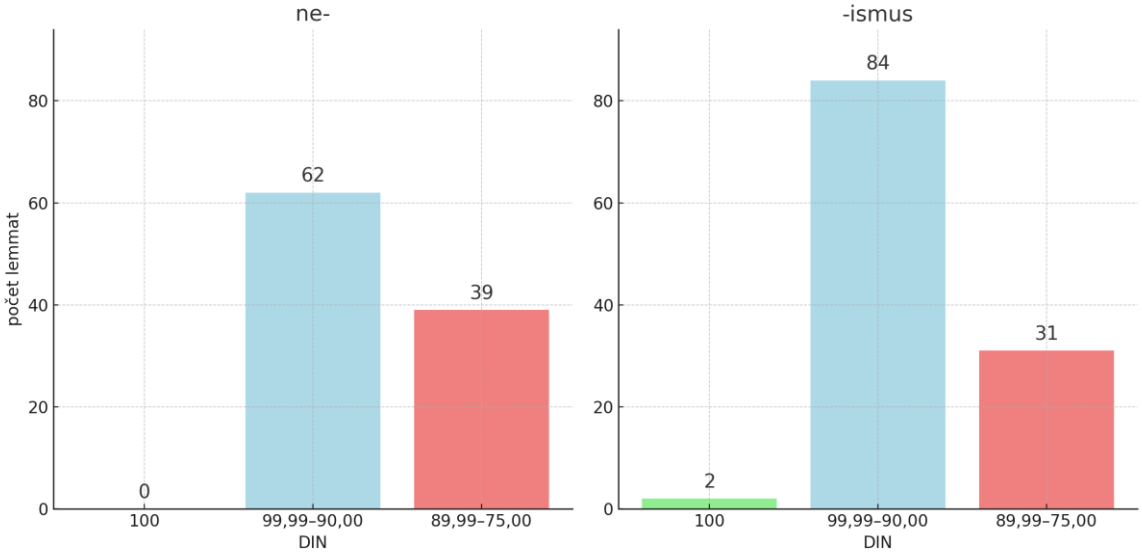
Jak bylo popsáno v kap. 3.2, provedl jsem celkem 10 analýz slovesných tvarů, přičemž jako první jsem analyzoval celý korpus *VK*. Výsledky této analýzy identifikovaly jedinou klíčovou položku (tedy s hodnotou DIN vyšší než 75), a to *VDP1SP*, což je značka pro sloveso plnovýznamové v indikativu pasiva první osoby singuláru v prézentu. Tato značka měla hodnotu DIN 90,58, zatímco druhá nejklíčovější značka *VDP1SR* (sloveso plnovýznamové v indikativu pasiva první osoby singuláru v minulém čase) měla hodnotu DIN 72,35. Výsledky analýz všech devíti jednotlivých subkorpusů prokázali, že verbtag *VDP1SP* byl identifikován jako klíčový v každém z nich, jedná se tedy o slovesný tvar silně charakteristický pro Václava Klause.

Při bližším pohledu na konkrétní výskyty tohoto verbtagu v celém korpusu *VK* (v němž má verbtag *VDP1SP* absolutní frekvenci 765) je zřejmé, že se na jeho vysoké klíčovosti podílí jedno konkrétní sloveso, a to *přesvědčit*, konkrétně tedy tvar *jsem přesvědčen*, který má absolutní frekvenci 590 výskytů. Na tomto výsledku je pozoruhodné, že při srovnávání stejných korpusů podle atributu *lemma_lc* má lemma *přesvědčit* hodnotu klíčovosti DIN pouhých 66,88, a nebylo by tak identifikováno jako klíčové, ale díky atributu verbtag

Tento výsledek tak z poněkud jiné perspektivy potvrdil předchozí zjištění Josefa Šlerky, který *jsem přesvědčen, že* identifikoval jako nejčastější kolokaci v Klausových politických projevech (viz pozn. 3).

4.3.2 Afixace

Dle očekávání identifikovaly analýzy prefixu *ne-* a sufixu *-ismus* vyšší desítky klíčových lemmat s těmito afixy, jak je vidět z *Grafu 8*.



Graf 8 – Korpus *VK*

Jelikož všechna identifikovaná lemmata mají v rámci druhého pásma klíčovosti hodnoty DIN poměrně vysoké, je užitečné kromě samotné klíčovosti sledovat také absolutní frekvence všech lemmat.

V rámci druhého pásma klíčovosti má následujících 14 lemmat s prefixem *ne-* frekvenci vyšší než 10: *nerovnováha* (184), *nevyhnutelně* (74), *nesvoboda* (71), *neekonom* (39), *nevyhnutelnost* (17), neudržitelnost (16), *neefektivnost* (13), *nepřijatelnost* (12), *nechápání* (11), *nepřejícnost* (11) *neřešení* (11), *nedemokracie* (10), *nepokora* (10) *nepružnost* (10)

Následujících 11 (respektive 10, kvůli nedokonalé lemmatizaci tvaru *rightismu*) lemmat se sufixem *-ismus* bylo identifikováno s nejvyššími hodnotami DIN: rightismu (11), rightismus (4), intergovernmentalismus (14), supranacionalismus (30), transnacionalismus (17), covidismus (21), evropeismus (110), gradualismus (27), evropanismus (3), genderismus (31), euronaivismus (5).

4.3.3 Diskurzní analýza — případová studie *Euro Business Breakfast*

Pro účely diskurzní analýzy se bohužel oproti očekáváním navrhnutá metodologie neprokázala být zcela nejvhodnější. Provedených 16 analýz klíčových slov sice dalo některé zajímavé výsledky, ale zdá se, že i v takto malých subkorpusech je hodnota minimální frekvence 2 příliš nízká. Získaných výsledků tak bylo příliš mnoho s příliš malými rozdíly hodnot DIN i absolutních frekvencí, aby mohly být samy o sobě relevantní. Výsledky takto provedených analýz sice mohou poskytnout alespoň nějaký výchozí bod pro případné návazné konkordanční a kolokační analýzy, ale pro takto malé subkorpusy je lze provést i bez analýzy klíčových slov, která tak nenabízí příliš přidané hodnoty.

Závěr

Závěry výzkumu

Hlavním cílem předkládaného výzkumu byla komplexní analýza idiolektu Václava Klause a sledování jeho potenciálních proměn v čase a v závislosti na veřejných funkcích, které Klaus vykonával. Při analýzách jednotlivých zkoumaných aspektů Klausova idiolektu jsem využíval především kvantitativní přístup, založený na metodách korpusové lingvistiky, který jsem podle potřeby doplňoval kvalitativním přístupem s využitím poznatků zejména lexikologie a analýzy diskurzu.

Nyní se pokusím zformulovat odpověď na hlavní výzkumnou otázku:

Jaké lexikální prostředky jsou charakteristické pro idiolekt Václava Klause?

Pro idiolekt Václava Klause, respektive pro jeho veřejnou podobu, jsou charakteristické především lexikální prostředky související s jeho profesí ekonoma a s jeho dlouholetou politickou kariérou. Lexikální rovinu Klausova idiolektu jsem zkoumal pomocí korpusové metody analýzy klíčových slov, díky níž bylo možné určit, které lexikální prostředky jsou v Klausových textech statisticky signifikantní ve srovnání s referenčním jazykovým materiálem, a tím pádem jsou charakteristické právě pro Václava Klause. Pro účely tohoto výzkumu jsem z mnoha materiálů z Klausových oficiálních webových stránek sestavil *Korpus textů Václava Klause*.

Díky víceúrovňové aplikaci metody analýzy klíčových slov se mi podařilo identifikovat devět lexikálních jednotek, které jsou v Klausových textech klíčové již dlouhodobě a napříč různými registry, a dá se tedy říci, že představují základ veřejné podoby jeho idiolektu. Jedná se o následující lexikální jednotky: *politický*, *ekonomický*, *ekonomika*, *ekonom*, *ekonomie*, *měnový, transformace*, *deficit* a *HDP*.

Kromě těchto dlouhodobě stálých lexikálních jednotek se mi s využitím metody analýzy klíčových slov podařilo zmapovat také částečné proměny lexikální roviny Klausova idiolektu v čase, respektive v závislosti na veřejných funkcích, které vykonával. V době, kdy byl Klaus předsedou vlády, byly v jeho textech klíčové lexikální jednotky týkající se především ekonomické transformace republiky a privatizace státního majetku. V období, kdy Klaus působil ve funkci poslance a předsedy Poslanecké sněmovny, začaly být v jeho textech klíčovější lexikální jednotky související s rozpočtovou a daňovou politikou, ale také s politickými směry a ideologiemi. Během vykonávání prezidentského úřadu již nebyly

v Klausových textech tolik klíčové ekonomické pojmy, ale více se v jeho idiolektu projevovaly lexikální prostředky související s vyjadřováním názorů a ideologií, s prosazováním moci a poprvé systematičtěji též s globálním oteplováním, respektive klimatickou změnou. Od doby, co Klaus ukončil svou aktivní politickou kariéru, se mezi nejklíčovější lexikální jednotky v jeho textech opět zařadily ekonomické lexémy, celkově je však jeho lexikální zásoba pestřejší a odlišnější mezi jednotlivými registry.

Kromě zmiňovaných lexikálních prostředků, které jsou sémanticky dosti výrazné, se v Klausově idiolektu opakovaně jako statisticky klíčové objevují také sémanticky méně nápadné lexikální jednotky, jako např. *dnešní*, *postoj* nebo *nikoli*. Díky analýze Klausových sněmovních projevů v korpusu *ParlCorp* se podařilo zjistit, že v kontextu těchto projevů je pro Klausův idiolekt charakteristická fráze *naprosto evidentní*. Naproti tomu analýza Klausových prezidentských projevů v korpusu *Speeches* poukázala na skutečnost, že Evropská unie patří k nejklíčovějším lexikálním jednotkám právě v textech z prezidentského období.

Dále jsem pro svůj výzkum vytyčil čtyři vedlejší cíle. Prvním z nich bylo samotné vytvoření *Korpusu textů Václava Klause*. Tento cíl se podařilo splnit, byť oproti původnímu záměru nakonec výsledný korpus obsahuje texty shromážděné pouze z jednoho zdroje, kterým jsou Klausovy oficiální webové stránky. Přesto lze říci, že s velikostí cca 1,5 milionu pozic (bez interpunkce) se jedná o dostatečně velký autorský korpus, v rámci ČNK první specializovaný na texty jednoho politika.

Druhým vedlejším cílem výzkumu byla analýza vybraného gramatického rysu Klausova idiolektu. V souvislosti s tímto cílem jsem položil vedlejší výzkumnou otázku:

Jaké slovesné tvary jsou charakteristické pro idiolekt Václava Klause?

Analýzou klíčových slovesných tvarů bylo zjištěno, že pro idiolekt Václava Klause je silně charakteristické užívání sloves ve tvaru indikativu pasiva první osoby singuláru v prézentu. Další zkoumání ukázalo, že na takto významné statistické signifikanci tohoto slovesného tvaru se podílí konkrétně tvar *jsem přesvědčen*. Díky tomu, že tento slovesný tvar je silně klíčový napříč všemi obdobími i registry, lze jednoznačně říci, že je pro Klausův idiolekt nejcharakterističtější.

Třetím vedlejším cílem výzkumu byla analýza vybraného slovotvorného rysu Klausova idiolektu. V souvislosti s tímto cílem jsem zformuloval vedlejší výzkumnou otázku:

> **V jaké míře jsou pro idiolekt Václava Klause charakteristické lexémy s prefixem *ne-* a lexémy se sufixem *-ismus*?**

Analýzou lemmat obsahujících prefix *ne-* a lemmat obsahujících sufix *-ismus* se podařilo empiricky ověřit pozorování publicistů, kteří si dlouhodobě všímají Klausovy specifické slovotvorby s užíváním těchto afixů. Jednak je počet různých lexémů s těmito afixy v Klausově idiolektu značný (vyšší desítky), jednak jejich klíčovost dosahuje vysokých hodnot. Byť se v případě některých lexémů jedná spíše o okazionalismy, jejichž absolutní frekvence jsou velmi nízké, a nelze tedy říci, že by byly typické ve smyslu frekvence, souhrnně je možné konstatovat, že oba afixy představují v Klausově idiolektu velmi charakteristický slovotvorný rys.

Čtvrtým vedlejším cílem byla korpusově založená diskurzní analýza vybrané případové studie, konkrétně Klausových projevů na pravidelných setkáních *Euro Business Breakfast*. Výzkumná otázka k tomuto cíli zněla:

> **Souvisejí klíčová slova v projevech Václava Klause na setkáních *Euro Business Breakfast* pouze s jejich tématy, anebo vypovídají také o Klausově diskurzní praxi?**

Tento vedlejší cíl se bohužel nepodařilo splnit. Navrhnutá metodologie by sice mohla mít potenciál, avšak bylo by pravděpodobně potřeba změnit jisté parametry provedených analýz, aby bylo získaných výsledků méně, zato byly relevantnější.

Limity výzkumu

Za nejvýraznější limit předkládaného výzkumu považuji skutečnost, že jsem si vytyčil cíle až příliš komplexní a široce rozkročené, kvůli čemuž nakonec nebylo možné dostatečně se věnovat každému z vybraných aspektů Klausova idiolektu a provést důkladnější interpretace získaných výsledků. Jedná se především o případovou studii pro diskurzní analýzu, ale také

třeba o analýzy klíčových slov v korpusech *ParlCorp* a *Speeches*, kde by např. podrobnější návazné konkordanční a kolokační analýzy mohly poskytnout více relevantních výsledků. Domnívám se, že by vzhledem k časové náročnosti samotného sestavení korpusu a k potřebě navrhnutí důmyslné metodologie pro analýzu idiolektu s využitím nového softwaru pro analýzu klíčových slov stačilo zaměřit se pouze na velmi úzce vymezenou část Klausova idiolektu.

Dalším, avšak těžko ovlivnitelným limitem výzkumu se ukázaly být určité technické nedostatky nového nástroje v rozhraní *KonText*, modulu *Analýza klíčových slov*. Některé z nich se podařilo odstranit ještě během výzkumu, za což ještě jednou velmi děkuji vývojářům z ÚČNK. Jiné se bohužel zatím odstranit nepodařilo, a tak jim musela být uzpůsobena metodologie, jak bylo popsáno v praktické části. Co se však týče hlavního zmiňovaného problému, tedy nefunkčnosti publicistických subkorpusů korpusu *SYN v12* jako referenčních korpusů, výsledky se nakonec nezdály být nijak významně negativně ovlivněny. Během testovacích analýz se výsledky při srovnávání s celým korpusem *SYN v12* prokázaly být přinejmenším stejně uspokojivé jako výsledky srovnávání s publicistickými subkorpusy jednotlivých korpusů řady *SYN*, většinou dokonce působily výrazně lépe. Přestože pravděpodobně nelze dělat přílišné generalizace, v tomto případě se ukázalo, že stejná pravidla lemmatizace zkoumaného a referenčního korpusu ovlivní analýzu klíčových slov pozitivním směrem více než jejich časová a žánrová blízkost.

Diskuse

I přes výše zmíněné nedostatky se domnívám, že předkládaný výzkum je dostatečně kvalitní a jsem přesvědčen, že má tři hlavní přínosy.

Prvním přínosem je vybudování *Korpusu textů Václava Klause*, který představuje dobrou materiálovou základnu jak pro další potenciální výzkumy, jež by se zabývaly jazykem Václava Klause, tak pro jeho případné rozšíření o další zdroje než jen webové stránky.

Druhým přínosem je testování nového nástroje *Analýza klíčových slov* v rozhraní *KonText* na webu ČNK. Pokud je mi známo, tento výzkum je prvním, v němž byl tento nástroj plně využíván, a mohl tak přispět k dílčím vylepšením a opravám.

Konečně třetí přínos je metodologický. Přestože korpusový výzkum idiolektu s využitím metody analýzy klíčových slov není ničím novým, jak ukázal přehled vybrané literatury v teoretické části, snažil jsem se nabyté teoretické poznatky zužitkovat a navrhnout některé nové drobné metodologické postupy, jak by taková analýza mohla vypadat, a na co by měla dbát. Byť jsem příliš široký záběr výzkumu zmiňoval jako jeho limit, na druhou stranu jsem díky těmto „odbočkám" mohl alespoň částečně ukázat různé možnosti, k čemu všemu se dá využít nový modul *Analýza klíčových slov*.

Z hlediska zkoumaného subjektu doufám, že tento výzkum pomohl alespoň částečně nahlédnout charakteristiky idiolektu Václava Klause a že případně někdy poslouží jako výchozí bod pro další bádání, neboť se jedná o téma, v němž lze nalézt ohromné množství zajímavého jazykového materiálu.

Seznam použité literatury

Monografie

BAKER, Paul. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

BAKER, Paul & HARDIE, Andrew & MCENERY, Tony. (2006). *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

BAKER, Paul & MCENERY, Tony (eds.). (2015a). *Corpora and Discourse: Integrating Discourse and Corpora*. Basingstoke: Palgrave.

COLERIDGE, Samuel Taylor. (1884). *Biographia Literaria*. 2nd ed. London: George Bell and Sons.

COULTHARD, Malcolm & JOHNSON, Alison & WRIGHT, David. (2017). *An Introduction to Forensic Linguistics: Language in Evidence*, 2nd ed. London / New York: Routledge.

CVRČEK, Václav & LAUBEOVÁ, Zuzana & LUKEŠ, David & POUKAROVÁ, Petra & ŘEHOŘKOVÁ, Anna & ZASINA, Adrian Jan. (2020). *Registry v češtině*. Praha: Lidové noviny.

ČERMÁK, František. *Jazyk a jazykověda*. (2011). Přehled a slovníky. 4. vyd. Praha: Karolinum.

ČERMÁK, František. (2017). *Korpus a korpusová lingvistika*. Praha: Karolinum.

ČERMÁK, František. (2010). *Lexikon a sémantika*. Praha: Lidové noviny.

ČERMÁK, František (ed.). (2007). *Slovník Karla Čapka*. Praha: Lidové noviny.

ČERMÁK, František & CVRČEK, Václav (eds.). (2009). *Slovník Bohumila Hrabala*. Praha: Lidové noviny.

DAVID, Jaroslav & ČECH, Radek & DAVIDOVÁ GLOGAROVÁ, Jana & RADKOVÁ, Lucie & ŠÚSTKOVÁ, Hana. (2013). *Slovo a text v historickém kontextu: perspektivy historickosémantické analýzy jazyka*. Brno: Host.

DERRIDA, Jacques. (1995). *Points…: Interviews, 1974–1994*. Stanford: Stanford University Press.

FIDLER, Masako & CVRČEK, Václav (eds.). (2018a). *Taming the Corpus: From Inflection and Lexis to Interpretation*. Cham: Springer.

PARTINGTON, Alan & DUGUID, Alison & TAYLOR, Charlotte. (2013). *Patterns and Meanings in Discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam / Philadelphia: John Benjamins.

PAUL, Hermann. (1891). *Principles of the History of Language*. London: Longmans, Green, and Co.

SCOTT, Mike & TRIBBLE, Christopher. (2006). *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam / Philadelphia: John Benjamins.

STUBBS, Michael. (1996). *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford: Blackwell Publishers.

TOGNINI-BONELLI, Elena. (2001). *Corpus Linguistics at Work*. Amsterdam / Philadelphia: John Benjamins.

WIERZBICKA, Anna. (1997). *Understanding Cultures through Their Key Words: English, Russian, Polish, German, and Japanese*. New York: Oxford University Press.

ZIPF, George Kingsley. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, Massachusetts: Addison-Wesley.

Články, studie, kapitoly

BAKER, Paul. (2004). Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis. In: *Journal of English Linguistics*, 32 (4), pp. 346–359. Thousand Oaks: Sage Publications.

BAKER, Paul & MCENERY, Tony. (2015b). Who Benefits When Discourse Gets Democratised? Analysing a Twitter Corpus around the British Benefits Street Debate. In: BAKER, Paul & MCENERY, Tony (eds.). *Corpora and Discourse: Integrating Discourse and Corpora*, pp. 244–265. Basingstoke: Palgrave.

BARLOW, Michael. (2013). Individual differences and usage-based grammar. In: *International Journal of Corpus Linguistics*, 18 (4), pp. 443–478. Amsterdam / Philadelphia: John Benjamins.

BLOCH, Bernard. (1948). A Set of Postulates for Phonemic Analysis. In: *Language*, 24 (1), pp. 3–46. Linguistic Society of America.

CONIAM, David. (2004). Concordancing Oneself: Constructing Individual Textual Profiles. In: *International Journal of Corpus Linguistics*, 9 (2), pp. 271–298. Amsterdam / Philadelphia: John Benjamins.

CULPEPER, Jonathan. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. In: *International Journal of Corpus Linguistics*, 14 (1), pp. 29–59. Amsterdam / Philadelphia: John Benjamins.

CULPEPER, Jonathan & DEMMEN, Jane. (2015). Keywords. In: BIBER, Douglas & REPPEN, Randi (eds.). *The Cambridge Handbook of English Corpus Linguistics*, pp. 90–105. Cambridge: Cambridge University Press.

CVRČEK, Václav & FIDLER, Masako. (2019). More than keywords: Discourse prominence analysis of the Russian Web Portal *Sputnik Czech Republic*. In: SALAMUROVIČ, Aleksandra & BERROCAL Martina (eds.). *Political discourse in Central, Eastern and Balkan Europe*, pp. 93–117. Amsterdam / Philadelphia: John Benjamins.

CVRČEK, Václav & FIDLER, Masako. (2022). No keyword is an island: In search of covert associations. In: *Corpora*, 17 (2), pp. 259–290. Edinburgh: Edinburgh University Press.

EIBL, Otto & GREGOR, Miloš & MACKOVÁ, Alena. (2013). O čem a jak hovořil Václav Klaus ve svých veřejných projevech. In: *Politologický časopis*, 20 (4), s. 392–418. Brno: FSS MUNI.

FIDLER, Masako & CVRČEK, Václav. (2015). A Data-Driven Analysis of Reader Viewpoints: Reconstructing the Historical Reader Using Keyword Analysis. In: *Journal of Slavic Linguistics*, 23 (2), pp. 197–239. Nova Gorica: Slavic Linguistics Society.

FIDLER, Masako & CVRČEK, Václav. (2018b). Going Beyond "Aboutness": A Quantitative Analysis of *Sputnik Czech Republic*. In: FIDLER, Masako & CVRČEK, Václav (eds.). *Taming the Corpus: From Inflection and Lexis to Interpretation*, pp. 195–225. Cham: Springer.

FIDLER, Masako & CVRČEK, Václav. (2019). Keymorph analysis, or how morphosyntax informs discourse. In: *Corpus Linguistics and Linguistic Theory*, 15 (1), pp. 39–70. Berlin: De Gruyter Mouton.

HAUGEN, Einar. (1972). From idiolect to language. In: SCHERABON FIRCHOW, Evelyn & GRIMSTAD, Kaaren & HASSELMO, Nils & O'NEIL, Wayne A. (eds.). *Studies by Einar Haugen. Presented on the Occasion of his 65th Birthday*, pp. 415–421. The Hague / Paris: Mouton.

LEHEČKOVÁ, Eva. (2016) Kritická analýza diskurzu v kontextu korpusové a kognitivní lingvistiky. In: HASIL, Jiří (ed.). *Přednášky z 59. běhu Letní školy slovanských studií*, s. 65–76. Praha: FF UK.

MILOM, Vannessa L. (2022). Corpus Linguistic Analysis of the Idiolects of Gollum and Sméagol. In: *Journal of Linguistics and Literature*, 5 (1), pp. 1–5. Newark: Science and Education Publishing.

MOLLIN, Sandra. (2009). "I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations. In: *International Journal of Corpus Linguistics*, 14 (3), pp. 367–392. Amsterdam / Philadelphia: John Benjamins.

SCOTT, Mike. (1997). PC analysis of key words — And key key words. In: *System*, 25 (2), pp. 233–245. Amsterdam: Elsevier.

STUBBS, Michael. (2010). Three concepts of keywords. In: BONDI, Marina & SCOTT, Mike (eds.). *Keyness in Texts*, pp. 21–42. Amsterdam / Philadelphia: John Benjamins.

Kvalifikační práce

LEŠKO, Marek. (2012). *The individual textual profile: a corpus-based study of idiolect*. Praha: Filozofická fakulta Univerzity Karlovy.

MERTA, Viktor. (2007). *Analýza promluv aktérů řádu mediovaného politického diskurzu*. Olomouc: Filozofická fakulta Univerzity Palackého v Olomouci.

Slovníky

REJZEK, Jiří. (2015). *Český etymologický slovník*. 3. vyd. Praha: Leda.

ŠIMANDL, Josef (ed.). (2016). *Slovník afixů užívaných v češtině*. Praha: Karolinum.

Encyklopedická hesla

HAZEN, Kirk. Idiolect. (2006). In: BROWN, Keith (ed.). *Encyclopedia of Language & Linguistics*. 2nd ed., vol. 5, pp. 512–513. Amsterdam: Elsevier.

WRIGHT, David. Idiolect. (2018). In: ARONOFF, Mark (ed.). *Oxford Bibliographies in Linguistics*. New York: Oxford University Press.

Online zdroje

BARBER, Alex & GARCIA RAMIREZ, Eduardo (2021). Idiolects. In: ZALTA, Edward N. (ed.). *The Stanford Encyclopedia of Philosophy*. URL: https://plato.stanford.edu/archives/sum2021/entries/idiolects/. Cit. 5. 4. 2024.

HOFFMANNOVÁ, Jana (2017). Projevy mluvené a psané. In: KARLÍK, Petr & NEKULA, Marek & PLESKALOVÁ, Jana (eds.). *CzechEncy – Nový encyklopedický slovník češtiny*. URL: https://www.czechency.org/slovnik/PROJEVY%20MLUVENÉ%20A%20PSANÉ. Cit. 5. 4. 2024.

KRČMOVÁ, Marie (2017). Idiolekt. In: KARLÍK, Petr & NEKULA, Marek & PLESKALOVÁ, Jana (eds.). *CzechEncy – Nový encyklopedický slovník češtiny*. URL: https://www.czechency.org/slovnik/IDIOLEKT. Cit. 5. 4. 2024.

MARTINCOVÁ, Olga (2017). Okazionalismus . In: KARLÍK, Petr & NEKULA, Marek & PLESKALOVÁ, Jana (eds.). *CzechEncy – Nový encyklopedický slovník češtiny*. URL: https://www.czechency.org/slovnik/OKAZIONALISMUS. Cit. 6. 5. 2024.

NEKVAPIL, Jiří (2017). Sociolingvistika. In: KARLÍK, Petr & NEKULA, Marek & PLESKALOVÁ, Jana (eds.). *CzechEncy – Nový encyklopedický slovník češtiny*. URL: https://www.czechency.org/slovnik/SOCIOLINGVISTIKA. Cit. 5. 4. 2024.

DIN. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=pojmy:din&rev=1569573129. Cit. 17. 4. 2024.

Frekvence. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=pojmy:frekvence&rev=1614015385. Cit. 18. 5. 2024.

Klíčové slovo (keyword). In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=pojmy:keyword&rev=1421871839. Cit. 10. 4. 2024.

Korpus prezidentských projevů Speeches. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=cnk:speeches&rev=1443694178. Cit. 4. 5. 2024.

Korpus SYN2020. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=cnk:syn2020&rev=1697209338. Cit. 6 5. 2024.

Korpus SYN verze 12. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=cnk:syn:verze12&rev=1703837693. Cit. 4. 5. 2024.

KWIC. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=pojmy:kwic&rev=1707383726. Cit. 10. 4. 2024.

Lemma. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=pojmy:lemma&rev=1641228720. Cit. 6. 5. 2024.

Morfologické kategorie a hodnoty v atributu verbtag a jejich značkování. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=seznamy:verbtagy_detail&rev=1668772417. Cit. 1. 5. 2024.

Morfologické značky (tagy) a jejich hodnoty. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=seznamy:tagy&rev=1650299545. Cit. 1. 5. 2024.

ParlCorp: Korpus českých parlamentních projevů. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=cnk:parlcorp&rev=1622898246. Cit. 4. 5. 2024.

Poziční atributy. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=pojmy:atributy_pozicni&rev=1641224995. Cit. 1. 5. 2024.

Referenční korpus. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=pojmy:referencni&rev=1610833298. Cit. 12. 4. 2024.

Reprezentativnost korpusu. In: *Příručka ČNK*. URL: http://wiki.korpus.cz/doku.php?id=pojmy:reprezentativnost&rev=1669722582. Cit. 13. 5. 2024.

Seznam grafů