**Univerzita Karlova**

**Filozofická fakulta**

Ústav českého jazyka a teorie komunikace

# Diplomová práce

Bc. Leona Špetová

## Srovnání nejfrekventovanějších lemmat v psané a mluvené češtině

Comparison of the most frequent lemmas in written and spoken Czech

2022

Vedoucí práce: Mgr. Dominika Kováříková, M.A., Ph.D.

Ráda bych poděkovala Mgr. Dominice Kováříkové, M.A., Ph.D., za ochotu, cenné připomínky a především za řadu velice přínosných konzultací.

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

V Praze 22. 12. 2022

Leona Špetová

## Abstrakt

Diplomová práce se zabývá 20 nejfrekventovanějšími lemmaty v psaném korpusu SYN2020 a v mluveném korpusu ORAL. Cílem práce je zjistit, jaké funkce a významy mají jednotlivá lemmata, a především prozkoumat, zda se tyto významy a funkce liší v mluvené a psané komunikaci. Potřebný materiál je získán z korpusu psané češtiny SYN2020 a z korpusu mluvené češtiny ORAL. Na náhodném vzorku z obou korpusů je provedena sémantická analýza zkoumaných lemmat. Výsledky porovnání 20 nejfrekventovanějších lemmat v psaném a mluveném korpusu ukazují, že nejfrekventovanějšími lemmaty v obou korpusech jsou výrazy s gramatickou funkcí, které se podílejí na výstavbě a organizaci textu. Tato lemmata mají často více významů. Z analýzy vyplývá, že nejfrekventovanějšími výrazy v korpusu SYN2020 jsou předložky a spojky, v korpusu ORAL zájmena a deiktické výrazy. Nejčastější výrazy v korpusu ORAL mají často také pragmatickou funkci, tyto výrazy je možné označit za diskurzní markery. Porovnáním nejfrekventovanějších lemmat lze sledovat odlišné charakteristiky psané a mluvené komunikace.

## Klíčová slova

mluvený jazyk, psaný jazyk, korpusový výzkum, frekvence, lexikální analýza

## Abstract

The diploma thesis deals with the 20 most frequent lemmas in the SYN2020 corpus of written Czech and the ORAL corpus of informal spoken Czech. The aim of this thesis is to find out what functions and meanings studied lemmas have and, above all, to investigate whether these meanings and functions differ in spoken and written communication. The data for the research were extracted from the SYN2020 corpus of written Czech and the ORAL corpus of informal spoken Czech. A semantic analysis of the lemmas is based on a random sample from both corpora. The results of the comparison of the 20 most frequent lemmas in the written and spoken corpora show that the most frequent lemmas in both corpora are expressions with a grammatical function, which are involved in the construction and organization of the text. These lemmas often have multiple meanings. The analysis shows that the most frequent expressions in the SYN2020 corpus are prepositions and conjunctions and in the ORAL corpus pronouns and deictic expressions. The most frequent expressions in the ORAL corpus often also have a pragmatic function. These expressions can be described as discourse markers. By comparing the most frequent lemmas we can observe different characteristics of written and spoken communication.

## Keywords

spoken language, written language, corpus-based research, frequency, lexical analysis

## Seznam použitých zkratek

AF: absolutní frekvence

ARF: average reduced frequency

ASSČ: Akademický slovník současné češtiny

ČNK: Český národní korpus

GSL: General Service List

NGSL: New General Service List

new-GSL: New General Service List

SSČ: Slovník spisovné češtiny

SSJČ: Slovník spisovného jazyka českého

RF: relativní frekvence

# Obsah

Úvod ..... 9

Teoretická část ..... 12

1 Frekvence v lingvistice ..... 12

1.1 Význam a využití frekvence ..... 12

1.2 Způsoby zachycení frekvence ..... 14

1.3 Frekvenční slovníky češtiny ..... 15

1.4 Frekvenční seznamy slov ..... 17

2 Vztah mluvenosti a psanosti ..... 20

2.1 Hranice mezi mluveností a psaností ..... 20

2.2 Jazykový paradox a problém neexistence metajazyka ..... 23

3 Korpusový výzkum mluveného a psaného jazyka ..... 25

4 Slovní zásoba psané a mluvené komunikace ..... 28

4.1 Distribuce slovních druhů ..... 28

4.2 Výskyt osobních a (pseudo)ukazovacích zájmen ..... 30

4.3 Interaktivní prostředky ..... 31

4.4 Repetitivnost a omezenost prostředků ..... 31

4.5 Jevy související s promluvou v reálném čase a sdíleným kontextem ..... 32

5 Podobné práce a výzkumy ..... 33

Praktická část ..... 37

6 Vlastní výzkum ..... 37

6.1 Popis výzkumu ..... 37

6.1.1 Zkoumaná lemmata ..... 37

6.1.2 Získání dat ..... 39

6.2 Analýza materiálu ..... 40

6.2.1 Frekvenční špička korpusu SYN2020 ..... 41

6.2.2 Frekvenční špička korpusu ORAL ..... 43

6.2.3 Srovnání frekvenčních špiček korpusu SYN2020 a korpusu ORAL .....45
6.2.3.1 Lemmata nacházející se ve frekvenční špičce obou korpusů.....48
6.2.3.2 Lemmata nacházející se pouze ve frekvenční špičce korpusu SYN2020 ..63
6.2.3.3 Lemmata nacházející se pouze ve frekvenční špičce korpusu ORAL .....68
6.3 Shrnutí výsledků výzkumu .....79
Závěr .....83
Seznam použité literatury .....86

# Úvod

Přibližně v sedmdesátých letech minulého století začaly vznikat první korpusy mluveného jazyka, které značně ovlivnily výzkumy týkající se mluvené komunikace. Snadná dostupnost velkého množství dat umožnila jak kvalitativní, tak kvantitativní výzkumy. Vznik mluvených korpusů podnítil badatele k tomu, aby se soustředili na mluvený jazyk, kterému do té doby nebyla věnována taková pozornost jako jazyku psanému.

I přes velké množství prací, které od té doby prohloubily poznatky o mluvené komunikaci, se mnoho lingvistů shodne, že toho o tomto typu komunikace víme poměrně málo a je potřeba tyto poznatky stále rozšiřovat (o nedostatečném prozkoumání mluvené komunikace píše např. Carter – McCarthy, 2017; Hoffmannová – Zeman, 2017).

V této diplomové práci zkoumám frekvenční špičku[1] (tedy 20 nejfrekventovanějších lemmat) psaného korpusu SYN2020 a mluveného korpusu ORAL. Cílem této práce je odpovědět na následující otázky, které je možné shrnout do tří bodů:

1) Která lemmata se vyskytují ve frekvenční špičce korpusu SYN2020 a ORAL? Proč jsou tato lemmata tak frekventovaná, proč jsou frekventovanější než jiná lemmata?
2) Jaké jsou významy a funkce nejfrekventovanějších lemmat v korpusu SYN2020 a ORAL? V čem se tato lemmata shodují a v čem se liší? Užívají se v korpusu SYN2020 a ORAL shodně, nebo odlišně?
3) Co lze na základě nejfrekventovanějších lemmat v korpusu SYN2020 a ORAL vyvodit ohledně psané a mluvené komunikace?

Práce by měla přispět k prohloubení poznatků o psané a mluvené komunikaci. Její výsledky mohou být využity např. jako materiály pro lexikografické zpracování zkoumaných lemmat či jako materiály pro mezijazykové srovnání. Práce může být také využita jako podklady pro výuku češtiny jako cizího jazyka.

Tato práce je rozdělena do dvou hlavních částí – teoretické a praktické. Teoretickou část tvoří pět kapitol. V první kapitole se věnuji frekvenci v lingvistice. Postupně představím, co je to frekvence, a uvedu její využití. Následuje stručný popis významných českých frekvenčních slovníků. V závěru první kapitoly se věnuji frekvenčním seznamům slov. Uvedu, co jsou seznamy slov, k čemu se používají a které faktory by měly být zváženy při jejich tvorbě.

[1] Podrobněji o pojmu *frekvenční špička* viz kapitola 6.

Ve druhé kapitole se věnuji vztahu mezi mluveností a psaností. V první části této kapitoly se zabývám hranicí mezi mluvenou a psanou komunikací. V souvislosti s tím se věnuji prototypům psanosti a mluvenosti. Druhá část této kapitoly se zabývá paradoxem reprezentace mluveného jazyka prostřednictvím grafického systému. V rámci toho zmiňuji problematiku adekvátního metajazyka pro popis mluvené a psané komunikace.

Třetí kapitola se věnuje dvěma přístupům, které se zabývají otázkou (korpusového) výzkumu mluveného a psaného jazyka. Hlavním rozdílem obou přístupů je názor na to, jak moc se mluvená komunikace liší od psané a do jaké míry má být mluvená komunikace popisována a zkoumána ve vztahu k psané komunikaci.

Ve čtvrté kapitole se zabývám slovní zásobou psané a mluvené komunikace. Na základě odborné literatury jsem vymezila pět jazykových jevů či oblastí, ve kterých spočívají největší odlišnosti slovní zásoby mluveného a psaného jazyka.

V poslední kapitole teoretické části představím lingvistické práce či výzkumy, které se věnují podobnému tématu jako tato diplomová práce. Představeny jsou vybrané práce jak z oblasti české lingvistiky, tak i některé zahraniční práce.

V praktické části se věnuji vlastnímu výzkumu. Nejprve podrobněji představím koncept výzkumu – na jaká lemmata se zaměřuji, kde a jakým způsobem jsem získala potřebná data, jaké metodologické problémy bylo potřeba řešit apod. Následně analyzuji jednotlivá lemmata. V rámci analýzy hledám odpovědi na otázky, které byly představeny v jednom z předchozích odstavců. V závěru praktické části shrnu zásadní poznatky a zjištění, ke kterým jsem na základě analýzy zkoumaného materiálu dospěla.

Aby bylo od začátku zřejmé, kterým lemmatům se chci ve své práci věnovat, rozhodla jsem se do úvodu této práce zařadit tabulku s 20 nejfrekventovanějšími lemmaty v korpusech SYN2020 a ORAL. Levý sloupec tabulky zobrazuje pořadí konkrétního lemmatu v rámci frekvenční špičky daného korpusu. Další dva sloupce představují konkrétní lemmata. Tabulka tak například ukazuje, že čtvrtým nejfrekventovanějším lemmatem je v korpusu SYN2020 lemma *v* a v korpusu ORAL lemma *no*. Lemmaty zobrazenými v tabulce č. 1 se bude podrobněji zabývat tato práce.

| 20 nejfrekventovanějších lemmat v korpusu SYN2020 a ORAL | | |
|---|---|---|
| **pořadí** | **korpus SYN2020** | **korpus ORAL** |
| 1 | *být* | *ten* |
| 2 | *a* | *být* |
| 3 | *se* | *a* |
| 4 | *v* | *no* |
| 5 | *na* | *se* |
| 6 | *ten* | *já* |
| 7 | *on* | *on* |
| 8 | *s* | *tak* |
| 9 | *že* | *že* |
| 10 | *z* | *jo* |
| 11 | *který* | *mít* |
| 12 | *mít* | *jako* |
| 13 | *do* | *tam* |
| 14 | *o* | *na* |
| 15 | *i* | *ale* |
| 16 | *ale* | *v* |
| 17 | *k* | *hmm* |
| 18 | *já* | *vědět* |
| 19 | *jako* | *ne* |
| 20 | *moci* | *co* |

*Tabulka 1: 20 nejfrekventovanějších lemmat v korpusu SYN2020 a ORAL*

# Teoretická část

## 1 Frekvence v lingvistice

Jelikož se tato práce zabývá nejfrekventovanějšími lemmaty, pojem *frekvence* je pro ni klíčový. Z toho důvodu se mu věnuji hned v počátečních stranách tohoto textu. V následující části postupně představím, co je to frekvence, a především uvedu její využití, díky kterému jí bývá v lingvistice (zejména korpusové) připisován tak velký význam.

Následně se zaměřím na to, jak může být frekvence při zkoumání jazyka pojímána a jakým způsobem se upravuje. V rámci toho představím základní pojmy, které budu ve své práci používat, např. absolutní frekvence, relativní frekvence či rank.

Další část této kapitoly tvoří popis tří významných českých frekvenčních slovníků. Jedná se o díla, která jsou založena primárně na frekvenci slov. Svým způsobem tak ovlivnila (a do určité míry i posunula) vnímání frekvence v české lingvistice.

V závěru této kapitoly se věnuji frekvenčním seznamům slov. Tuto část jsem do své práce zahrnula z toho důvodu, že ve svém výzkumu (v praktické části) pracuji se seznamy slov. Postupně tedy uvedu, co jsou to seznamy slov, k čemu se používají a jaká kritéria by měly splňovat (respektive které faktory by měly být zváženy při jejich tvorbě). V rámci toho představím významné anglické seznamy slov.

Na tomto místě je důležité zmínit jednu terminologickou poznámku. V této práci jsou pojmy „mluvená a psaná komunikace“ a „mluvený a psaný jazyk“ chápány jako synonyma, v následujícím textu je volně zaměňuji. Jsem si však vědoma toho, že v některých lingvistických pracích se jedná o dva různé termíny.[2]

### 1.1 Význam a využití frekvence

Frekvencí se v korpusové lingvistice označuje počet výskytů jevu v korpusu, četnost formy nebo jevu. Stanovit frekvenci slova či jevu na základě vlastní introspekce bývá obvykle nepřesné, proto jako jeden z hlavních zdrojů informace o frekvenci slouží jazykové korpusy. V rámci korpusové lingvistiky je frekvence považována za základní ukazatel, který má vliv na popis a zkoumání jazyka (Cvrček, 2017a).

[2] Například *Mluvnice současné češtiny* (Cvrček a kol., 2015, s. 18) charakterizuje jazyk jako prostředek komunikace, komunikaci pak popisuje jako přenos informací, který předpokládá na jedné straně produktora a na druhé straně recipienta.

Jedním ze základních přínosů informace o frekvenci je to, že uživateli daného jazyka pomáhá určit, co se v daném jazyce používá, co je důležité a který jev, slovo, konstrukce apod. je preferovanější. Řada autorů, kteří se ve svých pracích věnují frekvenci, uvádí, že frekvence funguje jako jeden z ukazatelů centra a periferie jazykových jevů. Například V. Cvrček a D. Kováříková (2011) píší, že frekvence může nasměrovat lingvisty ke zkoumání důležitých jazykových jevů, které je třeba popsat, místo aby byla pozornost zaměřována na jevy okrajové. Podobně F. Čermák a M. Křen (2004) zmiňují, že frekvence je jeden ze způsobů, jak zjistit, co je pro jazyk zvlášť důležité. Ve své další práci F. Čermák (2017) píše, že centrum jazyka a vysoká frekvence se do jisté míry podmiňují – co je frekventované a časté, je v jazyce centrální. Vysoká frekvence často naznačuje známost, typičnost, obvyklost a pravidelnost (což nemusí platit u jazykových anomálií, např. frazémů).

Podobně autoři P. Nation a R. Waring (1997) uvažují o frekvenci jako o nástroji, pomocí kterého je možné určit užitečnost[3] slova (či jazykových jevů). Jako příklad velmi užitečného slova v angličtině uvádějí výraz *the*, který je velmi frekventovaný jak v psané, tak v mluvené angličtině.

Informace o frekvenci v jazyce je dále velmi prospěšná pro nerodilé mluvčí při studiu cizího jazyka. Díky frekvenčním údajům může mluvčí posoudit, jak moc se daný výraz v jazyce užívá a jak moc je pro něj potřebný. Zároveň informace o frekvenci slov slouží pro vyučující cizích jazyků jako východisko, pomocí kterého mohou stanovit obsah a cíl vzdělávání. Za tímto účelem se vytvářejí frekvenční seznamy slov, např. *General Service List* nebo *New General Service List* (podrobněji se těmto seznamům slov věnuji v oddílu 1.4).

Frekvence je jedním z hlavních kritérií pro lexikografy při tvorbě a uspořádání hesláře slovníku. Lexikograf často přihlíží k frekvenci výskytu při rozhodování, které jednotky do hesláře zařadit. Například materiálový základ pro vytvoření hesláře *Akademického slovníku současné češtiny* tvoří soubor sestavený ze tří korpusů (SYN2000, SYN2005 a SYN2010). Jako hraniční pro zařazení výrazu do hesláře se považuje frekvence pěti výskytů v tomto souboru. Frekvence také hraje roli v případě pořadí homonym v hesláři slovníku. Na prvním místě se zpravidla uvádí homonymum frekventovanější. Obdobně se postupuje i u polysémních slov, kdy se při určování základního významu přihlíží k frekvenci daného výrazu. V případě slov s více možnými variantami (např. dubletami) se rovněž často přihlíží k variantě s vyšší frekvencí (Kochová – Opavská, 2016).

[3] Autoři používají výraz *usefulness.*

V předchozích odstavcích jsem představila význam frekvence v jazyce. Na základě rozličných způsobů využití informace o frekvenci je její přínos pro lingvistické zkoumání jazykových jevů zřejmý. Nicméně je velmi důležité podotknout, že při zkoumání a popisu jazyka je sice důležité znát jeho kvantitativní charakteristiky, avšak bez vhodné intepretace či kvalitativního rozboru jsou tato čísla často jen hrubým vodítkem. Jak již bylo řečeno, frekvence je pro tuto práci klíčová, nicméně nejedná se pouze o kvantitativní rozbor, ale také kvalitativní. Ve své práci se přikláním k následujícímu tvrzení autorů díla *Frekvence slov, slovních druhů a tvarů*: „Bylo by velká chyba, kdybychom chtěli přisuzovat číslům získaným kvantitativním rozborem konečnou platnost a kdybychom se chtěli jimi spokojovat. Číselné údaje, mají-li míti cenu pro poznání jazyka, musí být spojeny s kvalitativním rozborem […]“ (Jelínek – Bečka – Těšitelová, 1961, s. 9).

## 1.2 Způsoby zachycení frekvence

Frekvence se obvykle uvádí buď v absolutních hodnotách, nebo relativních. Absolutní frekvencí se rozumí prosté spočtení všech výskytů jednotlivého slova, formy nebo i všech slov v textu nebo v korpusu (Čermák, 2017). Dané zjištění o frekvenci se ale vztahuje pouze k jednomu konkrétnímu textu, se kterým pracujeme, nikoli komplexně k jazyku. Hodnota absolutní frekvence výrazu (či jevu) závisí na dalších faktorech textu, např. na jeho délce, autorovi, stylu apod. Abychom mohli frekvenci dvou či více slov (popřípadě jevů) z různých textů srovnávat a interpretovat, je vhodné ji určitým způsobem upravit. Příkladem toho může být relativní frekvence, tedy frekvence vztažená k celkové velikosti textu, případně korpusu. Relativní frekvence se počítá jako poměr absolutní frekvence a celkové délky textu (Čermák, 2017). V korpusové lingvistice se relativní frekvence často uvádí v i.p.m. (*instances per million*), tedy v počtu výskytů slova v textu o délce 1 milionu slov. Relativní frekvence tak slouží jako odhad pravděpodobnosti, s jakou se daná forma v textu (korpusu) vyskytne (Cvrček, 2017a).

Dalším způsobem úpravy frekvence je tzv. průměrná redukovaná frekvence (ARF, *average reduced frequency*). ARF kompenzuje nevyvážené stoupání absolutní frekvence určitého slova v důsledku jeho vysoké frekvence v případech, kde jsou zdroje omezené na jeden či několik málo pramenů a frekvence na nich založená není v průměru srovnatelná s ostatními slovy. Frekvence se tedy koriguje přepočítáním s ohledem na počet různých zdrojů, resp. textů (Čermák, 2017, s.110). ARF tedy zabraňuje tomu, aby slovo, které je velmi frekventované pouze v malém množství textu (případně pouze v textech jednoho autora), zkreslilo celkový

obraz frekvence nejčastějších slov. Tento způsob úpravy frekvence byl použit např. při vytváření *Frekvenčního slovníku češtiny* (Čermák – Křen, 2004).

Frekvence může být dále popisována pomocí tzv. ranku. Jedná se o způsob prezentace informací o frekvenci v seznamech slov, kdy každému slovu je na základě jeho pořadí v seznamu udělena určitá hodnota. Prvnímu (tedy obvykle nejfrekventovanějšímu) slovu či jevu je přiřazen rank 1, druhému slovu či jevu rank 2 atd. (Baker – Hardie – McEnery, 2006, s. 136). Vztahem mezi frekvencí a rankem se zabývá první Zipfův zákon, jehož autorem je americký sociolog G. K. Zipf. Tento zákon vyjadřuje, že frekvence slova je nepřímo úměrná jeho ranku. Odráží fakt, že každý text obsahuje malý počet velmi frekventovaných slov, ale většinu slov málo frekventovaných (Čermák, 2017, s. 106).

## 1.3 Frekvenční slovníky češtiny

Jedním ze základních děl, které zpracovává frekvenci slov v jazyce, je frekvenční slovník. Obvykle je sestaven na základě textového korpusu a jednotlivé lexikální jednotky, které slovník obsahuje, jsou opatřeny frekvenčními a dalšími kvantitativními údaji. Hesla ve slovníku bývají primárně řazena podle frekvence, dále mohou být řazena např. abecedně či na základě své slovnědruhové platnosti (Křen, 2017). Využití frekvenčních slovníků se v podstatě shoduje s údaji uvedenými v oddílu 1.1, proto jej již podrobněji nepopisuji.

Prvním frekvenčním slovníkem češtiny je dílo J. Jelínka, J. V. Bečky a M. Těšitelové s názvem *Frekvence slov, slovních druhů a tvarů v českém jazyce* (1961). V rámci tvorby tohoto díla autoři analyzovali celkem 75 literárních děl a tiskovin různého stylu a žánru. Autoři podotýkají, že slovníku by prospělo, kdyby v něm byla zachycena i mluvená komunikace, nicméně sehnat vhodný materiál bylo příliš obtížné. Textový základ slovníku tak dohromady tvoří přibližně 1,6 milionu excerpovaných slov. Autoři práce si jsou vědomi, že tento materiál není dostatečně velký, aby vylučoval možnosti nahodilosti a zkreslení. F. Čermák a M. Křen (2004, s. 7) například upozorňují na to, že příliš malá výchozí báze způsobila to, že okrajová slova jako *píst*, *ventil* a *válec* se chybně dostala mezi deset tisíc nejčastějších slov. Nicméně zde je důležité přihlédnout k době vzniku slovníku a faktu, že veškerý materiál byl analyzován ručně bez pomoci automatického počítačového zpracování.

Pokračovatelem tohoto díla je *Frekvenční slovník češtiny* (2004), který je zpracován kolektivem autorů pod vedením F. Čermáka a M. Křena. Zásadní změnou oproti předchozímu frekvenčnímu slovníku je to, že základnu pro *Frekvenční slovník češtiny* tvoří rozsáhlé jazykové korpusy (vytvořeny v Ústavu Českého národního korpusu), díky kterým je minimalizováno

nebezpečí málo reprezentativního a zkresleného výběru textů. Tyto texty jsou navíc počítačově zpracovány, takže za kratší čas je získán mnohem větší objem dat. Nicméně toto počítačové zpracování přináší i určité problémy – programy pro automatickou lemmatizaci a morfologickou analýzu textu vykazují určitou míru chybovosti. Z toho důvodu musí být získaný materiál různě korigován a podle autorů lze předpokládat, že se jim nepodařilo vyhnout úplně všem chybám (Čermák – Křen, 2004, s. 8).

Jako jeden z nedostatků *Frekvenčního slovníku češtiny* uvádí jeho autoři (Čermák – Křen, 2004, s. 8) nepřítomnost autentických mluvených textů, které byly v době vzniku slovníku daleko za finančními i časovými možnostmi současných korpusových projektů. V návaznosti na tento fakt byl o několik let později vydán *Frekvenční slovník mluvené češtiny* (2007), který vznikl rovněž pod vedením F. Čermáka. Tento slovník je založen na datech z Pražského mluveného korpusu, získaných na základě magnetofonových nahrávek v terénu (pořízených převážně v první polovině devadesátých let minulého století). Podle autora slovníku se jedná o první zcela ojedinělý obraz současné české mluvené situace (Čermák a kol., 2007, s. 7).

U *Frekvenčního slovníku mluvené češtiny* (2007) je zajímavým způsobem řešena lemmatizace hesel. Jako základní slovníkový tvar byla zvolena taková podoba lemmatu, na kterou ukazuje většinový úzus (v analyzovaném materiálu). Z toho důvodu jsou ve slovníku zaznamenána taková lemmata jako *bejt*, *cejtit*, *von*, *malej* apod. Zde je ale důležité upozornit na fakt, že jako materiálová základna pro tvorbu tohoto slovníku sloužily nahrávky zachycené v Praze a jejím okolí. Zůstává tedy otázkou, nakolik slovník opravdu reflektuje většinový úzus a jestli některé základní podoby slov nepramení právě v tom, že do slovníku nebyla zahrnuta např. některá nářečí.

Především pro studenty češtiny jako cizího jazyka je důležité dílo *A Frequency Dictionary of Czech: Core Vocabulary for Learners* (2011), jehož autory jsou opět F. Čermák a M. Křen. Slovník předkládá 5000 nejfrekventovanějších slov v češtině. Od předchozích frekvenčních slovníků se liší jednak tím, že je určen především nerodilým mluvčím češtiny, ale také tím, že autoři při jeho tvorbě vycházeli jak z korpusů psaného jazyka, tak i z korpusů jazyka mluveného. Na rozdíl od předchozích slovníků tak neobsahuje slova, která jsou frekventovaná buď pouze v psané, nebo v mluvené komunikaci, ale představuje nejfrekventovanější slova v komunikaci obecně.

## 1.4 Frekvenční seznamy slov

Seznamem slov se v korpusové lingvistice myslí takový seznam, který obsahuje všechna slova z konkrétního textu či korpusu a u každého slova uvádí jeho frekvenci (Baker – Hardie – McEnery, 2006, s. 169). Obvykle slova v těchto seznamech bývají řazena na základě své frekvence, případně mohou být řazena abecedně. Svým způsobem se tak seznamy slov velmi podobají frekvenčním slovníkům, často také tvoří výchozí bázi pro tvorbu těchto slovníků.

Na rozdíl od frekvenčních slovníků, které mají rozmanité funkce a jejich předpokládaným uživatelem může být prakticky kdokoliv, bývají podle P. Bakera a kol. (2006) frekvenční seznamy slov určeny primárně pro studenty či vyučující cizích jazyků. Podle autorů P. Nationa a R. Waringa (1997) mohou být seznamy slov užitečné pro vyučující cizích jazyků, pokud potřebují zjistit, zda studenty učí všechna potřebná slova (případně zda všechna slova, která studenty učí, jsou potřebná). V. Březina a D. Gablasová (2015) podotýkají, že seznamy slov jsou vhodné také pro začínající studenty, kteří si při studiu cizího jazyka kladou otázku, kde vůbec začít.

Seznamy slov obvykle vznikají na základě jazykového korpusu. Tvorba těchto seznamů by však neměla probíhat pouze automaticky (strojově), ale měla by být při ní dodržena určitá pravidla. Autoři P. Nation a R. Waring (1997) popisují 6 následujících faktorů, které by měly být zváženy při tvorbě seznamů slov:

1. reprezentativnost – seznam slov by měl být založen na různých typech textů s různými žánry, včetně textů mluvené komunikace;

2. frekvence a rozsah – slova s vysokou frekvencí by se měla vyskytovat v různých typech textů, nikoli v malém množství tematicky zaměřených textů;

3. etymologicky příbuzná slova – při tvorbě frekvenčního seznamu by měli autoři zvážit, zda některá příbuzná slova neřadit pod jedno heslo;

4. idiomy a víceslovné jednotky – do seznamu slov by měly být zahrnuty některé frekventované víceslovné jednotky (např. *good morning* či *never mind*);

5. informace o jednotce – u každé jednotky (slova) by měly být uvedeny i další informace, např. slovnědruhová příslušnost, frekvence, základní význam slova apod.;

6. ostatní kritéria – v rámci vytváření seznamu slov by měla být zvážena i další kritéria, která se týkají konkrétního seznamu, jeho zaměření a předpokládaného uživatele. Příkladem

takového kritéria je užitečnost či nezbytnost slova – do seznamu by měly být řazeny takové jednotky, které nemohou být synonymně nahrazeny jinými (Nation – Waring, 1997, s. 18–19).

Zároveň je důležité podotknout, že tato kritéria platí především pro tvorbu všeobecného seznamu slov, tzn. ne pro specifické seznamy, které se týkají pouze některých jazykových oblastí. Při tvorbě specifických seznamů slov, např. seznamu akademických slov[4], by některé body nemusely být relevantní (např. kritérium č. 1 a 2).

Jedním z nejznámějších a nejvlivnějších seznamů slov je anglický *General Service List* (GSL) z roku 1953, jehož autorem je M. West. Jedná se o seznam přibližně 2000 nejfrekventovanějších slov v angličtině. Tento seznam byl vytvořen pomocí anglického psaného korpusu a teoreticky byl podložen pracemi z první poloviny 20. století, které se věnují frekvenci. Slova do tohoto seznamu byla zařazována také na základě jejich zvládnutelnosti, potřebnosti či stylistického a emocionálního hodnocení (Březina – Gablasová, 2015).

I přes značný přínos GSL poukazuje řada lingvistů na některé problémy, které jsou s tímto seznamem spojeny. Jedním ze základních problémů je stáří tohoto seznamu. Jelikož byl vytvářen během 40. let minulého století, slovní zásoba, kterou seznam zpracovává, je poměrně zastaralá. Zároveň seznam nezachycuje aktuální slovní zásobu. Jako příklad slov, která zachycuje tento seznam, ale aktuálně se nepoužívají tak často, uvádí V. Březina a D. Gablasová (2015, s. 2) slova jako *shilling* (*šilink*), *servant* (*sluha* či *služebná*) a *telegraph* (*telegraf*). Podle těchto autorů by naopak do seznamu měla být zařazena následující slova, která v současnosti patří mezi velmi frekventované výrazy: *television* (*televize*), *computer* (*počítač*) a *internet*. Dalším argumentem pro nahrazení GSL novým seznamem slov je podle P. Nationa a R. Waringa (1997) fakt, že texty, které tvořily podkladovou bázi pro tento seznam, nebyly dostatečně žánrově a tematicky diverzifikované.

Na základě těchto argumentů začaly vznikat nové podoby GSL (jedná se v podstatě o jakési aktualizace starého seznamu). Jedním z nejznámějších seznamů, který navazuje na GSL, je *New General Service List* z roku 2013, jehož autory jsou Ch. Browne, B. Culligan a J. Phillips. Jeho materiálovou základnu tvoří rozsáhlý Cambridge English Corpus. Oproti předchozímu GSL je NGSL založen na žánrově a tematicky vyvážených textech a zároveň odráží aktuální slovní zásobu angličtiny (Stoeckel – Bennett, 2015).

[4] Seznam nejčastějších akademických slov a víceslovných jednotek v češtině poskytuje webová aplikace Akalex (Kováříková – Kovářík, 2021). Aplikace je dostupná na adrese https://www.korpus.cz/akalex/.

Autory další významné aktualizace GSL jsou V. Březina a D. Gablasová. Jejich seznam se rovněž nazývá *New General Service List* a byl zveřejněn rovněž v roce 2013. Vzhledem ke shodnému názvu a datu zveřejnění bývají obě nové podoby GSL často zaměňovány. Aby ke zmatení z názvů nedocházelo tak často, bývá seznam od autorů V. Březiny a D. Gablasové zkracován jako new-GSL, zatímco seznam autorů Ch. Browneho, B. Culligana a J. Phillipse jako NGSL.

New-GSL byl vytvořen ze 4 velkých korpusů angličtiny (včetně známého Britského národního korpusu). Slova pro new-GSL byla vybrána na základě tří kritérií: frekvence, disperze a distribuce. Tato opatření zaručují, že slova vybraná pro nový seznam slov jsou často používána ve velkém počtu textů a že seznam slov je sestaven transparentním a replikovatelným způsobem. Oproti původnímu seznamu slov se do jeho aktualizované podoby dostala slova jako *internet*, *e-mail* či *online* (Březina – Gablasová, 2015).

Při zběžném porovnání 100 nejfrekventovanějších výrazů v seznamech NGSL a new-GSL jsem našla jen drobné rozdíly. Většina z nich vychází z toho, že new-GSL při lemmatizaci zohledňuje na rozdíl od NGSL slovní druh konkrétního výrazu. Například výraz *right*[5] je v NGSL zaznamenán jako 92. nejfrekventovanější slovo, ale v new-GSL jako 306. (v platnosti substantiva), 386. (v platnosti adjektiva) a 412. (v platnosti adverbia). Nejedná se však o chyby, ale o odlišné pojetí způsobu lemmatizace a morfologického značkování. Při rozhodování se, se kterým seznamem pracovat, tedy záleží především na záměru uživatele a na tom, čeho chce pomocí seznamu dosáhnout.[6]

[5] Do češtiny lze tento výraz přeložit různě, např. právo, dobro, správný, pravý, vpravo, správně, přímo, ihned aj.
[6] Například při tvorbě slovníku, kdy lexikograf rozlišuje jednotlivé významy slov, se jeví jako užitečnější new-GSL. Naopak NGSL by mohl být užitečnější pro začínající studenty angličtiny, kteří si chtějí být jistí, že se naučí všechna základní a potřebná slova.

# 2 Vztah mluvenosti a psanosti

V následující části práce stručně charakterizuji vztah mezi mluvenou a psanou komunikací. Hned v úvodu této kapitoly je důležité zmínit, že se nejedná o komplexní, vyčerpávající popis tohoto vztahu, ale o základní charakteristiku mluvené a psané komunikace. Aby byl popis tohoto vztahu komplexnější, bylo by vhodné jej reflektovat z různých hledisek, např. z hlediska stylistiky, syntaxe či sémiotického hlediska. Vzhledem k omezenému rozsahu této práce se zabývám pouze těmi aspekty vztahu mezi psanou a mluvenou komunikací, které jsou pro tuto práci zásadní.

V první části této kapitoly se věnuji hranici mezi psanou a mluvenou komunikací. Postupně představím vlastnosti psaného a mluveného jazyka. V souvislosti s tím se zabývám prototypy psanosti a mluvenosti. Toto téma je pro mou práci podstatné, jelikož se ve vlastním výzkumu zaměřuji na psanou a mluvenou komunikaci. Je tedy důležité tyto pojmy popsat a představit, jak jsou v rámci této práce chápány.

Druhá část kapitoly se zabývá paradoxem reprezentace mluveného jazyka prostřednictvím grafického systému. Toto téma je do práce zahrnuto, protože ve svém výzkumu vycházím z transkripce mluveného textu, která s sebou nese určitá omezení. Je důležité tato omezení zmínit a být si jich vědomi. V souvislosti s tím zmiňuji problematiku adekvátního metajazyka pro popis mluvené a psané komunikace. Jak podrobněji představím níže, někteří autoři (např. Carter – McCarthy, 2017) zmiňují, že neexistuje vhodný a uznávaný metajazyk, pomocí kterého by měla být popisována mluvená komunikace. Je tedy důležité upozornit na to, že pro určité jevy neexistuje vhodné pojmenování – lingvisté je buď označují pojmy používanými pro popis psaného jazyka, které ale ne vždy dokáží vhodně popsat jazyk mluvený, nebo zavádějí vlastní pojmy, které ale nejsou ustálené či uznávané větší vědeckou komunitou.

## 2.1 Hranice mezi mluveností a psaností

V řadě lingvistických prací se uvádí, že běžná komunikace probíhá dvěma způsoby – pomocí psaného média a pomocí zvukového. Bylo by poměrně zavádějící jednoduše stanovit, že psaná komunikace je realizována pouze skrze písmo a mluvená pouze skrze zvuk. Místo toho bývá mluvená a psaná komunikace popisována nejen skrze médium, kterým je realizována (produkována či přijímána), ale hlavně souhrnem vlastností, které oba typy komunikace charakterizují.

Příkladem takového pojetí mluvené a psané komunikace je práce P. Kocha a W. Oesterreichera (1985). Podle autorů lze oba typy komunikace charakterizovat mediálně (příp. substanciálně),

kdy psaný text je produkován graficko-verbálně a mluvený vokálně-akusticky. Zároveň lze oba typy komunikace popisovat konceptuálně, tedy na základě charakteristik typických pro psaný či mluvený jazyk. Podle autorů se mluvené texty vyznačují např. dialogičností, střídáním mluvčích či komunikací tváří v tvář. Naopak psané texty jsou obvykle monologické, nedochází u nich ke střídání mluvčích a autor a recipient jsou obvykle časoprostorově odděleni. Kombinací obou rysů (mediálnosti a konceptuálnosti) je podle P. Kocha a W. Oesterreichera (1985) možné popsat různé druhy textů, např:

- mediální mluvenost a konceptuální mluvenost: neformální rozhovor s přáteli
- mediální mluvenost a konceptuální psanost: slavnostní projev
- mediální psanost a konceptuální mluvenost: webový chat
- mediální psanost a konceptuální psanost: text zákona

Podobným způsobem charakterizuje psanou a mluvenou komunikaci J. Hoffmannová (1992). Autorka chápe psanou a mluvenou komunikaci jako škálu, na jednom jejím konci se ocitá mluvenost a na druhém psanost. Mluvenost je podle autorky spjata např. s nepřipraveností, spontánností, neformálností, neuspořádaností apod. Na druhé straně škály je psanost, která se vyznačuje připraveností, plánovitostí, formálností, organizovaností atd. Jednotlivé texty se na škále pohybují na základě kombinace těchto vlastností. Autorka zmiňuje, že jen málo textů je možné umístit na póly škály, většina textů se pohybuje někde mezi nimi. Texty, které se nacházejí na obou koncích škály, bychom mohli považovat za prototypy psanosti, resp. mluvenosti.

Prototypickému mluvenému textu se věnuje také F. Čermák (2009). Ve své práci popisuje, jaké parametry by měl mít reprezentativní mluvený korpus s jasným jádrem prototypického mluveného jazyka. Podle F. Čermáka by měl prototypický mluvený text splňovat 12 následujících rysů: (1) mluvenost (jedná se o text, který je původně mluvený, nejedná se např. o psaný text, který je čten nahlas), (2) dialogičnost, (3) blízkost (účastníci komunikace by se měli navzájem dobře znát), (4) rovnost (účastníci komunikace by si měli být rovni), (5) neveřejnost, (6) neformálnost, (7) interaktivnost (účastníci by se měli aktivně podílet na komunikaci), (8) přítomnost, (9) nenásobnost (jeden účastník by měl hovořit pouze k jednomu účastníkovi, ne např. ke skupině), (10) spontánnost, (11) běžnost a (12) nevědomost[7] (účastníci by si neměli být vědomi, že jsou nahráváni). Na základě přítomnosti (+) či nepřítomnosti (-)

[7] Tento rys souvisí především se získáváním materiálu pro výzkumy mluveného jazyka. Jedná se o to, že pokud mluvčí ví, že je jeho projev zaznamenáván, může být tímto faktem ovlivněn (např. mluví spisovněji, omezuje užívání vulgarismů apod.).

těchto rysů je možné popsat různé druhy textů, například rozhovor v rádiu by mohl být podle F. Čermáka (2009) popsán následovně (u parametrů označených jako „?“ záleží na konkrétní situaci): ?+1, +2, -3, -4, -5, -6, +7, +8, ?-9, ?+10, ?-11, +12[8]. Prototypem mluveného textu by byla podle autora konverzace mezi přáteli, což uvádí i v jedné ze svých starších prací (Čermák – Sgall, 1997).

K otázce konverzace jako prototypu mluveného jazyka se ve své práci vyjádřili R. Carter a M. McCarthy (2017). Upozorňují na to, že konverzace je pokládána za prototyp mluveného jazyka, přesto bychom neměli zapomínat, že existují i jeho jiné formy, které by měly být do výzkumů mluveného jazyka také zahrnovány. Zároveň poukazují na fakt, že chápání pojmu *konverzace* může být bez bližší specifikace v některých ohledech poměrně problematické. Tímto pojmem je často označována pouze neformální společenská konverzace mezi přáteli. Ve skutečnosti ale může označovat také jiné typy konverzace, např. akademickou a obchodní mluvenou komunikaci. Při užití pojmu *konverzace* (bez bližší specifikace) tak není v některých kontextech jasné, jaký typ konverzace je popisován (Carter – McCarthy, 2017).

V předchozích odstavcích jsem stručně popsala vztah mezi psanou a mluvenou komunikací. Z tohoto popisu vyplývá, že stanovit hranici mezi psaností a mluveností je velmi komplikované. V této práci se přikláním k výše představenému pojetí (srov. Hoffmannová, 1992), v rámci kterého je komunikace popisována jako škála. Na jednom konci škály je mluvenost, na druhém psanost. Mezi těmito póly je přechodové pásmo, kam patří většina textů.

Vzhledem k nejasné hranici mezi mluveností a psaností je velmi složité (a prakticky nemožné) vytvořit ideální vzorek, který by celkově odrážel psanou či mluvenou komunikaci. Lingvista obvykle pracuje s jazykovým korpusem, který ale reflektuje pouze část mluvené či psané komunikace. Výsledky svého výzkumu pak lingvista často generalizuje na psaný či mluvený jazyk celkově, ve skutečnosti se ale vztahují pouze k určité části mluveného či psaného jazyka. Tento problém se týká i mé diplomové práce. Ve svém výzkumu vycházím z dat korpusu ORAL, který obsahuje pouze jeden typ mluveného textu, a to neformální konverzaci s přáteli. Některé závěry ohledně mluvené komunikace, ke kterým ve svém výzkumu docházím, jsou zapříčiněny především tím, že vycházím z materiálu dialogické povahy (a ne z materiálu, který zachycuje např. mluvenou přednášku, tedy formální text monologického charakteru).

[8] Myslím si, že hodnotit rys nevědomosti je zde poměrně sporné. V tomto případě je to pravděpodobně myšleno tak, že účastníci rozhovoru nebyli předem informováni o tom, že jejich projevy budou později sloužit jako materiál pro lingvistický výzkum. Účastníci si jsou ale s největší pravděpodobností vědomi toho, že jsou jejich projevy nahrávány.

## 2.2 Jazykový paradox a problém neexistence metajazyka

V souvislosti s výzkumem a popisem mluveného jazyka zmiňuje řada autorů určitý paradox – mluvená komunikace je reprezentována grafickým systémem, což je médium užívané primárně pro jazyk psaný. Obvykle se jedná o „základní“ transkripce, které ale mají řadu omezení, např. postrádají fonologické či kontextové informace. Jedná se tedy spíše o selekci jazykových rysů z konkrétní promluvy (Kirk – Andersen, 2016). Z toho důvodu samotnou transkripci často doplňují a upřesňují doprovodné poznámky či specifická anotace (např. značení překryvů dvou či více promluv; speciální znaky pro hezitace; informace o tom, zda byla promluva pronesena šeptem, se smíchem, s pláčem apod.).

Ačkoli transkripce není dokonalým systémem, pomocí kterého by měla být zkoumána mluvená komunikace, aktuálně se jedná o jeden z mála způsobů, jak mluvený text uchovat, reprodukovat, analyzovat či v něm vyhledávat (Kirk – Andersen, 2016). Lingvista, který zkoumá mluvený jazyk skrze transkripci, by si však měl být vědom omezení, která s sebou transkripce přináší.

S tímto paradoxem reprezentace mluveného jazyka prostřednictvím grafického systému souvisí také otázka adekvátního metajazyka pro popis mluvené a psané komunikace. Podle některých lingvistů (viz např. Carter – McCarthy, 2017) je jedním ze zásadních problémů popisu mluveného jazyka fakt, že neexistuje vhodný a uznávaný metajazyk, který by měl být pro popis mluveného jazyka používán. Tito autoři navrhují, abychom se při popisu mluveného jazyka oprostili od metajazyka používaného pro popis psané komunikace a vytvořili metajazyk nový. Hlavním důvodem je to, že některé pojmy, které se užívají pro psaný jazyk, nejsou schopny zachytit a adekvátně popsat jazyk mluvený. Primárním cílem badatelů v oblasti mluveného jazyka by tak podle R. Cartera a M. McCarthyho (2017) mělo být vytvoření vhodného repertoáru (metajazyka), pomocí kterého potom můžeme mluvený jazyk popisovat.

Příkladem nedostatečného metajazyka může být problematický pojem pro úsek, který by měl při popisu mluvené komunikace nahradit termín *věta* (a obecně její koncept, který pro popis mluvené komunikace není vyhovující). Podle G. Leeche (2000) je možné věty segmentovat na základě koncové interpunkce. Použít ale toto kritérium pro segmentaci mluveného projevu není spolehlivé. Pokud se v transkripci řeči objeví koncová interpunkce, je to pouze „artefakt“ procesu přepisu, jelikož v řeči žádná interpunkce jako taková neexistuje (Leech, 2000, s. 704).

V souvislosti s vytvářením adekvátního metajazyka začala řada lingvistů představovat své koncepty adekvátní jednotky pro popis mluveného jazyka. Například G. Leech (2000) navrhuje

koncept tzv. C-units (*communication units*)[9]. Autoři R. Carter a M. McCarthy (1995) navrhují pro popis mluveného jazyka využívat tzv. prodlouženou klauzi (*extended clause structure*)[10]. V oblasti české lingvistiky se např. J. Hoffmannová a J. Zeman (2017) přiklánějí k užívání pojmu *výpověď* (který chápou jako strukturně, významově a prozodicky uzavřený celek, užitý v konkrétní komunikační situaci).

[9] Jedná se o maximální syntakticky analyzovatelnou jednotku, jejíž průměrná délka je (pro angličtinu) přibližně pět až šest slov. Tato jednotka může být gramaticky analyzována, ale nemůže být dále gramaticky spojena s další jednotkou za účelem vytvoření rozsáhlejší syntaktické jednotky (podrobněji o *C-units* viz Leech, 2000, s. 703).
[10] Struktura prodloužené klauze vypadá následovně: pre-clause + clause + post-clause. Pozice *pre-clause* obsahuje tzv. topic (multifunkční prostředek, který vyjadřuje téma výpovědi a zároveň usnadňuje recipientovi její zpracování), pozice *clause* vyjadřuje jádro výpovědi a pozice *post-clause* obsahuje tzv. tail (multifunkční prostředek, pomocí kterého může být retrospektivně potvrzeno to, co bylo dříve řečeno) (podrobněji o prodloužené klauzi viz Carter – McCarthy, 1995, s. 152).

# 3 Korpusový výzkum mluveného a psaného jazyka

Během 80. let začaly být publikovány první korpusy mluveného jazyka. Snadná dostupnost velkého množství lehce analyzovatelných dat podnítila mnoho lingvistů, aby se začali soustředit na mluvený jazyk, kterému do té doby nebyla věnována taková pozornost jako jazyku psanému. V souvislosti s tímto vzrůstajícím zájmem o mluvený jazyk vyvstala otázka, jakým způsobem k němu přistupovat, jak jej zkoumat a popisovat. V této kapitole představím dva přístupy, které se věnují otázce (korpusového) výzkumu mluveného a psaného jazyka. Hlavním rozdílem obou přístupů je názor na to, jak moc se gramatika mluvené komunikace liší od gramatiky psané komunikace a jak moc má být mluvený jazyk popisován a zkoumán ve vztahu k psanému jazyku.

Představiteli prvního přístupu jsou autoři mluvnice *Longman Grammar of Spoken and Written English* (1999), tedy zejména G. Leech, D. Biber, S. Johansson, S. Conrad a E. Finegan. Základní tezí tohoto přístupu je to, že gramatiky mluveného a psaného jazyka jsou v podstatě stejné, obě využívají shodný gramatický repertoár. To, co je odlišuje, jsou frekvence výskytu jednotlivých jevů, které se v gramatikách objevují, a implementace těchto jevů v jazyce (Leech, 2000, s. 675). Mluvená a psaná gramatika by měla být vnímána holisticky a k jejím popisům by se měl využívat stejný popisný rámec (tedy kategorie, pravidla a struktury, které se používají za účelem popisu mluveného a psaného jazyka).

Tento popisný rámec (užívaný pro popis mluvené i psané gramatiky) je podle G. Leeche (2000, s. 691) tvořen třemi následujícími rovinami:

> (1) Deskriptivní gramatikou, která specifikuje systém či repertoár pravidel, struktur a kategorií;
>
> (2) Gramatikou performance, která popisuje, jak je systém či repertoár pravidel, struktur a kategorií užíván v rámci různých variet jazyka[11] (např. v beletrii, publicistice, akademických textech a konverzaci);
>
> (3) Informacemi o relativních frekvencích výskytu tohoto repertoáru (relativní frekvence je stanovena podle dat obsažených v korpusu).

G. Leech (2000) upozorňuje, že v rámci tohoto modelu je velmi pravděpodobné, že výskyty některých gramatických rysů budou mít v určitých varietách až nulovou frekvenci. Jinak řečeno

[11] Zde je důležité upozornit, že autor používá pojem *variety jazyka* pro něco, co bychom v oblasti české lingvistiky nazývali spíše jako *textové makrotypy*.

– některé části tohoto repertoáru nemusí být použity ve všech varietách jazyka. Díky výše představenému modelu je možno na základě korpusových dat stanovit, že např. rys A je omezen pouze na varietu X a je prakticky vyloučen z variety Y. Tento model tedy předpokládá, že ačkoliv mluvený a psaný jazyk sdílejí stejný repertoár (mají stejnou deskriptivní gramatiku), implementace tohoto repertoáru je velmi odlišná (Leech, 2000, s. 692).

Podstatnou výhodou tohoto modelu je podle G. Leeche (2000) především to, že zohledňuje škálovitý vztah mezi psaným a mluveným jazykem, tedy že psané texty se do určité míry podobají mluveným textům a mluvené texty se naopak v různé míře podobají textům psaným.

Představiteli druhého přístupu jsou především autoři R. Carter, R. Hughes a M. McCarthy, kteří bývají na základě své působnosti na Nottinghamské univerzitě někdy souhrnně označováni jako „nottinghamská škola“ (viz např. Leech, 2000, či Čermáková – Kopřivová, 2018). Tento přístup vychází z práce D. Brazila (1995), ve které píše, že abychom porozuměli mluvené gramatice, musíme přijmout zcela odlišný model od těch, které se tradičně používají pro jazyk psaný. Hlavním důvodem, proč odmítnout modely psaného jazyka pro popis mluvené komunikace, je to, že mluvený jazyk vzniká a je interpretován lineárně (Brazil, 1995, s. 21). Na toto pojetí navazují představitelé nottinghamské školy, jejichž hlavní tezí je to, že gramatika mluveného a psaného jazyka se inherentně liší.

R. Carter a M. McCarthy ve své práci píší, že popisy gramatiky mluveného jazyka jsou stále zatíženy břemenem popisů jazyka psaného (Carter – McCarthy, 2017, s. 2). Popisný rámec, který se užívá pro popis psaného jazyka, nemůže být užíván pro popis mluveného jazyka. Tím představitelé druhého přístupu odmítají koncept stejného popisného rámce, který ve své práci popsal G. Leech.

Pro oba přístupy je stěžejní otázka frekvence výskytu v jazyce, v tomto ohledu panuje podle R. Cartera a M. McCarthyho (2017) mezi oběma přístupy shoda. Oba přístupy se primárně zajímají o jevy, které mají v mluvené komunikaci vysokou frekvenci, či o jevy, kterým nebyla při popisu mluvené komunikace dosud věnována dostatečná pozornost. Odlišný je ale způsob, jakým s údaji o frekvenci zacházejí. Představitelé prvního přístupu se běžně snaží tyto údaje porovnat s psanou komunikací a na základě tohoto srovnání pak vyvozovat závěry ohledně mluvené komunikace (příkladem tohoto postupu je již výše zmiňovaná mluvnice *Grammar of Spoken and Written English* z roku 1999). Naopak představitelé nottinghamské školy vyvozují závěry o mluvené komunikaci, aniž by ji srovnávali s komunikací psanou. Popis mluveného

jazyka by podle nich měl být založen výhradně na datech mluveného jazyka a jeho charakteristika by měla probíhat bez ohledu na jazyk psaný.

Na otázku, který přístup je „lepší“ (přesnější, efektivnější, srozumitelnější apod.) nejspíš neexistuje jednoznačná odpověď. Osobně si myslím, že se nelze rozhodnout pouze pro jeden přístup a aplikovat jej na všechny budoucí výzkumy. Spíše podle povahy konkrétního výzkumu volíme mezi těmito přístupy. Tato diplomová práce se svým pojetím přibližuje spíše prvnímu přístupu, a to především zkoumáním vztahu mezi mluvenou a psanou komunikací. Představitelé druhého přístupu popisují komunikaci v některých ohledech příliš binárně – tedy že na jedné straně existuje psaná komunikace a na druhé mluvená, ale příliš se nezmiňují o nějakém přechodovém pásmu mezi nimi. Oproti tomu první přístup vnímá komunikaci více škálovitě, tedy že v rámci mluvenosti a psanosti se pohybujeme na určité škále těchto rysů, což odpovídá pojetí mluvené a psané komunikace v této diplomové práci.

# 4 Slovní zásoba psané a mluvené komunikace

V jedné z předchozích kapitol jsem popsala, jakým způsobem lze nahlížet na psanou a mluvenou komunikaci. V souvislosti s tím jsem zmínila, jak je charakterizována psanost a mluvenost, jaké mají vlastnosti a kde se mezi nimi nachází hranice. V této kapitole se zabývám slovní zásobou psané a mluvené komunikace. Jelikož se ve svém vlastním výzkumu věnuji slovní zásobě mluvené a psané komunikace, je důležité shromáždit a stručně popsat dosavadní poznatky, na které můžu ve své práci navázat (případně je rozšířit, potvrdit či vyvrátit). Z toho důvodu jsem na základě odborné literatury vymezila pět jazykových jevů či oblastí, ve kterých spočívají největší odlišnosti slovní zásoby mluveného a psaného jazyka. Tato kapitola se zaměřuje na slovní zásobu, často se ale jedná o jevy, které jsou na pomezí lexikologie a jiných jazykovědných disciplín (převážně morfologie a syntaxe).

Většina autorů lingvistických prací, ze kterých v této kapitole vycházím, výsledky svých výzkumů zobecňuje na mluvenou komunikaci celkově. Ve skutečnosti ale obvykle při výzkumu vychází pouze ze spontánní konverzace. Některé poznatky se tedy vztahují spíše jen k tomuto typu mluveného textu.[12]

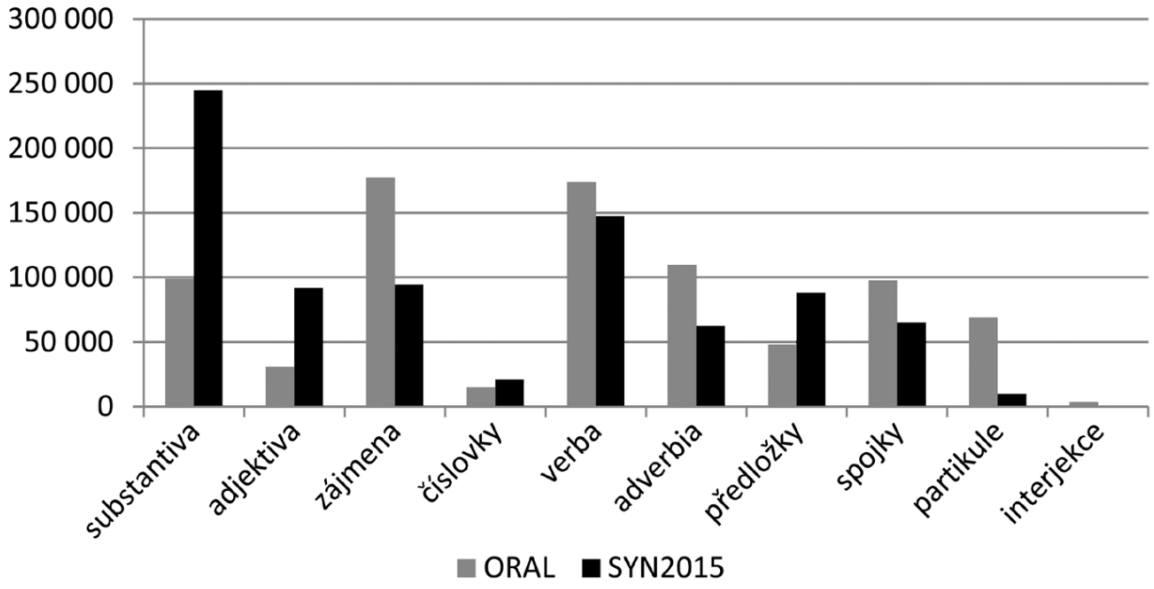
## 4.1 Distribuce slovních druhů

Řada autorů, kteří se zabývají distribucí slovních druhů v psané a mluvené komunikaci (srov. Čermáková – Kopřivová, 2018; Hoffmannová – Zeman, 2017), upozorňuje na nejasnou slovnědruhovou platnost mnoha lexémů v mluvené komunikaci. A. Čermáková a M. Kopřivová (2018, s. 234) píší, že tradiční slovnědruhová kategorizace se v mnoha ohledech zdá být pro mluvený jazyk nedostatečná. Zároveň při automatické anotaci korpusů (především mluvených) nedochází k rozpoznání různých významů a funkcí u některých slovnědruhově homonymních výrazů, takže informace o jejich slovnědruhové platnosti mohou být chybné. Vzhledem k tomu nemusí být slovnědruhové popisy psané a především mluvené komunikace, které vychází z automaticky anotovaných korpusů, vždy přesné, ale spíše jen orientační.

Distribuce slovních druhů se v psané a mluvené komunikaci poměrně liší. A. Čermáková a M. Kopřivová (2018) ve své práci představily graf distribuce slovních druhů v psané a mluvené češtině (viz obrázek 1), při jehož tvorbě vycházely z psaného korpusu SYN2015 a mluveného korpusu ORAL. Svislá osa grafu uvádí frekvenci slovního druhu

[12] O tomto problému pojednávám podrobněji v oddílech 2.1 a 6.1.1.

(udávanou v i.p.m.), vodorovná osa zobrazuje slovní druhy. Z grafu je patrné, že distribuce číslovek a sloves je podobná, distribuce ostatních slovních druhů se poměrně liší.




*Obrázek 1: Distribuce slovních druhů v psané a mluvené češtině (Čermáková – Kopřivová, 2018, s. 234)*

Jedním z největších rozdílů je distribuce substantiv, která jsou v psané komunikaci mnohem frekventovanější než v mluvené. Podle A. Adamovičové (2017) to může být způsobeno tím, že v psané komunikaci jsou velmi častá verbální substantiva, v mluvené komunikaci se ale více používají slovesa, od jejichž kmene byla tato substantiva odvozena. Podle F. Čermáka (1993) může být nízké zastoupení substantiv v mluvené komunikaci vysvětleno tím, že mluvčí při prvním pojmenování určitého předmětu, jevu, činnosti apod. použije substantivum, dále jej ale nahrazuje zájmenem či gestem. Tento předpoklad může zároveň vysvětlovat vysokou frekvenci ukazovacích zájmen v mluvené komunikaci, jimž se podrobněji věnuji v oddílu 4.2.

Zajímavá je distribuce předložek, které jsou na základě obrázku 1 téměř dvakrát frekventovanější v psané komunikaci. Podle A. Adamovičové (2017) se na velkém podílu předložek v psaném korpusu podílejí především víceslovné předložky vyskytující se zejména v nominálních konstrukcích, které pro mluvený projev nejsou typické.

Frekvenčně se velmi liší distribuce částic. A. Adamovičová (2017) tento rozdíl vysvětluje tím, že částice v mluvené komunikaci přebírají hodnotící funkci adjektiv. Z toho důvodu jsou pak adjektiva frekventovanější v psaném jazyce. Podle A. Čermákové a M. Kopřivové (2018) může být vysoká frekvence částic způsobena tím, že mezi částice bývají při výzkumech mluvené komunikace řazeny i některé výrazy, které se vymykají tradiční slovnědruhové kategorizaci, např. výplňková slova (*prostě*, *jakoby*, *jakože*) a kontaktové výrazy (*viď*, *pozor*). K takovému

tvrzení se přiklání i F. Čermák (2008), který zmiňuje, že je potřeba otázku slovnědruhové homonymie částic prozkoumat hlouběji a komplexněji.

Jak již bylo řečeno, tradiční slovnědruhová kategorizace se zdá být především pro mluvený jazyk v mnoha ohledech nedostatečná. Z toho důvodu charakterizuje řada autorů určité výrazy (především v mluvené komunikaci) jako diskurzní markery. Těm se ve své práci věnují A. Čermáková, L. Jílková, Z. Komrsková, M. Kopřivová a P. Poukarová (2019). Diskurzní markery[13] charakterizují jako krátká slova vyskytující se v mluvené komunikaci, která nejsou primárně nositeli sémantické informace, ale zprostředkovávají především informace pragmatické. Nemají lexikální význam (případně je jejich původní význam silně oslaben), nejedná se však o slova bezvýznamná, nepotřebná. Mohou mít následující funkce: pomáhají mluvčímu získat nebo předat slovo; slouží k vyplnění nebo zpomalení promluvy tehdy, když si mluvčí chce udržet slovo; potvrzují posluchačovu pozornost; ohlašují nové téma či posun v tématu; označují novou či starou informaci; jsou signálem opravy sebe sama nebo jiného mluvčího; vyjadřují posluchačovu reakci na to, co bylo řečeno; mluvčímu a posluchači pomáhají při vzájemné spolupráci (Čermáková a kol., 2019, s. 244–246).[14]

## 4.2 Výskyt osobních a (pseudo)ukazovacích zájmen

Jak již bylo zmíněno, zájmena jsou mnohem frekventovanější v mluvené komunikaci než v psané. V konverzaci se často hojně vyskytují osobní zájmena (především *já* a *ty*), která referují přímo k jejím participantům. Podle A. Čermákové a M. Kopřivové (2018) mluvčí často používají zájmeno *já*, aby si udrželi slovo, zdůraznili vlastní mínění či pomohli komunikačnímu partnerovi při orientaci v reprodukci cizí a vlastní řeči (pokud komunikačnímu partnerovi popisují dříve proběhlý rozhovor s jiným mluvčím a potřebují dát najevo, které argumenty byly jiného mluvčího a které byly jeho vlastní).

Kromě osobních zájmen upozorňují J. Hoffmannová a J. Zeman (2017) na vysokou frekvenci zájmen ukazovacích, resp. v některých případech spíše pseudoukazovacích. Tato zájmena mohou mít funkci gramatickou (uplatňují se při mimotextové deixi a vnitrotextové koreferenci) či pragmatickou (mají expresivní platnost, vyjadřují emoce a často fungují jako výplňkové slovo). V případě pragmatické funkce autoři uvažují o přechodu demonstrativ k částicím.

[13] Autorky zároveň upozorňují, že dosud není stanovena jednoznačná definice diskurzních markerů a užívaná terminologie je značně rozkolísaná, tyto výrazy jsou dále označovány např. jako *discourse particles*, *pragmatic markers* či *fillers* (Čermáková a kol., 2019, s. 244, 246). Ve své práci budu užívat termín *diskurzní marker*.
[14] Ve své práci autorky vychází z charakteristiky diskurzních markerů podle L. J. Brinton (1996).

Výrazu *to* se ve své práci věnují autoři P. Klimešová, Z. Komrsková, M. Kopřivová a D. Lukeš (2015). V úvodu zmiňují, že slovníkové zpracování tohoto výrazu není dostatečné, proto navrhují jeho vlastní klasifikaci. Mimo kategorii odkazování a kategorii kombinace (kam spadají výrazy *to*, které tvoří ustálenou jednotku s dalším komponentem, např. *to jo*) vyčleňují kategorii mluvenosti. Ta sdružuje užití výrazu *to* v jevech, které jsou typické pouze pro mluvený projev a neobjevují se v psaném jazyce. Jedná se především o falešné starty, nedokončené výpovědi a o užití výrazu z důvodu neschopnosti mluvčího vybavit si konkrétní slovo. V těchto případech bychom mohli uvažovat o přechodu výrazu *to* k částicím, příp. k diskurzním markerům.

## 4.3 Interaktivní prostředky

Další odlišnosti ve slovní zásobě psané a mluvené komunikace jsou zapříčiněny tím, že mluvený jazyk bývá interaktivní – mluvčí spolu vzájemně spolupracují při výstavbě textu. Jedním z nejčastějších prostředků interakce jsou již zmíněné diskurzní markery. Mezi nejčastější interaktivní prostředky patří výrazy *no*, *ano*, *jo*, *jasně*, *hmm* či *ne*. Podle autorek A. Čermákové a kol. (2019) pomocí těchto výrazů mluvčí vyjadřuje vztah k posluchači, vybízí jej k reakci, dává posluchači najevo, že hovor sleduje, případně že s něčím ne/souhlasí. Na konkrétní funkci těchto výrazů se často podílejí doprovodné suprasegmentální rysy.

Dalšími interaktivními prostředky mohou být gesta, pomocí kterých dává posluchač najevo, že hovor sleduje a ne/souhlasí s řečeným. Při použití gest v mluvené komunikaci opět narážíme na problém zachycení mluveného jazyka skrze grafický systém – základní transkripce obvykle neposkytují informace o gestech (či mimice) mluvčích (podrobněji viz oddíl 2.2).

## 4.4 Repetitivnost a omezenost prostředků

D. Biber a kol. (1999) v rámci výsledků svého výzkumu mluvené a psané komunikace zmiňují, že mluvený jazyk je více omezený a repetitivní než psaný. Mluvčí často opakuje to, co řekl další účastník komunikace těsně před ním. To D. Biber a kol. (1999) označují jako lokální opakování (*local repetition*). Autoři zároveň zjistili, že mluvená komunikace má nižší *token-type ratio*[15] než psaná. Znamená to, že v mluvené komunikaci je méně typů slov, která ale mají vyšší frekvenci, naopak v psané komunikaci je více typů slov s obecně nižší frekvencí. Toto zjištění dokládá, že psaný jazyk je celkově lexikálně bohatší, oproti tomu je mluvený jazyk poměrně omezený.

[15] Jedná se o poměr počtu různých slov (typů) k celkovému počtu všech slov v korpusu (Cvrček, 2017b).

Omezenější a repetitivní lexikální repertoár mluveného jazyka se podle A. Čermákové a M. Kopřivové (2018) projevuje různými rutinními konverzačními obraty. Ty vznikají podle autorek především kombinací výrazů *no*, *jo* a *to*. Mezi nejčastější čtyřslovné kombinace v mluveném projevu (na základě dat z korpusu ORAL) patří *no tak to je*, *to víš že jo*, *to je to je* a *no no no no* (s různou intonací), zatímco v psaném textu (na základě dat z korpusu SYN2015) jsou to následující kombinace: *bez ohledu na to*, *že se jedná o, a od té doby*, *a v neposlední řadě* (Čermáková – Kopřivová, 2018, s. 228). Hodnota relativní frekvence nejčastějších čtyřslovných kombinací je v mluveném korpusu několikanásobně vyšší než v psaném. Toto frekvenční srovnání podle autorek poukazuje na rutinní užívání těchto spojení a lexikální omezenost mluveného jazyka.

## 4.5 Jevy související s promluvou v reálném čase a sdíleným kontextem

Mluvená komunikace se od psané liší tím, že vzniká lineárně v reálném čase. Mluvčí obvykle nemá čas si promluvu předem rozvrhnout, opravit, přeformulovat apod. S tím souvisí řada jazykových jevů, které jsou typické pro mluvený projev, např. pauzy, restarty, elipsy či hezitace. Z lexikálního hlediska jsou zajímavé případy, kdy mluvčí z určitých příčin promluvu nedokončí. Podle J. Hoffmannové a J. Zemana (2017, s. 56) jsou příznačnými výrazovými exponenty takových neurčitých zakončení spojení jako *a tak (ňák)*, *nebo tak něco*, *nebo něco takovýho* aj.

Mluvená konverzace se zároveň odehrává v rámci určitého sdíleného kontextu (fyzického i mentálního). Tento aspekt je podle G. Leeche (2000) reflektován v užívání některých výrazů (či jevů), které mají v rámci konverzace velmi vysokou frekvenci (např. již zmiňované časté užívání osobních a ukazovacích zájmen, elipsy a gesta). Díky sdílenému kontextu tak mluvčí nemusí být při jazykovém projevu vždy explicitní a často používá výrazy s vágním významem (srov. Čermák, 1993).

# 5 Podobné práce a výzkumy

V následujících odstavcích představím některé práce či výzkumy lingvistů, které se věnují podobnému tématu jako tato diplomová práce. Publikací, které by se věnovaly stejnému tématu, tedy zkoumaly nejfrekventovanější lemmata v psané a mluvené komunikaci, popsaly je a srovnaly prostřednictvím psaného a mluveného korpusu, jsem nalezla jen velmi málo. Z toho důvodu se věnuji pracím, které se té mé podobají pouze částečně (např. zpracovávají podobné téma, ale jiný způsobem). Představeny budou vybrané práce jak z oblasti české lingvistiky, tak i některé zahraniční práce.

Základními publikacemi, které se věnují frekvenci slov a částečně se tak zabývají podobným tématem jako tato diplomová práce, jsou frekvenční slovníky. Výše byly představeny hlavní frekvenční slovníky češtiny (viz oddíl 1.3). Podobným dílem pro angličtinu je publikace autorů G. Leeche, P. Raysona a P. Wilsona s názvem *Word Frequencies in Written and Spoken English* (2001). Na základě Britského národního korpusu autoři vytvořili frekvenční slovník mluvené a psané angličtiny. Frekvenční slovníky se od ostatních děl, která budou dále představena, liší zejména tím, že se zpravidla zabývají pouze kvantitativními vlastnostmi jazyka, nikoliv kvalitativními. Obvykle tak slouží jako materiálová základna pro další (kvalitativní či kvantitativně-kvalitativní) výzkumy jazyka.

Rozdílným frekvencím slov v psané a mluvené komunikaci se ve své práci věnuje B. Dejmek (2002). Jako materiál pro svůj výzkum autor použil údaje z publikace *Frekvence slov, slovních druhů a tvarů v českém jazyce* (Jelínek – Bečka – Těšitelová, 1961)[16] a vlastní záznamy mluvených projevů, které byly pořízeny během 60. let v Přelouči. Analýzou 20 nejčastějších slov v těchto materiálech autor dochází ke zjištění, že oproti psané komunikaci se v mluvené velmi často užívají výrazy, které fungují jako důležité kontaktové prostředky (např. *no*, *jo*, *tak*) (Dejmek, 2002). Potenciálním nedostatkem toho výzkumu by mohla být skutečnost, že závěry o mluvené komunikaci jsou vyvozovány na základě analýzy projevu pouze sedmi mluvčích, přičemž všichni pochází z okolí Přelouče a jejich věk se pohybuje v rozmezí 32–40 let. Nízký počet a nedostatečná demografická a geografická rozmanitost dat by mohly ovlivnit výsledky výzkumu.

Srovnáním nejfrekventovanějších slov v psané a mluvené komunikaci se ve své práci zabývá A. Adamovičová (2017). Autorka srovnává 50 nejfrekventovanějších substantiv, adjektiv, verb a adverbií v psaném korpusu SYN2015 a v Pražském mluveném korpusu. Toto srovnání je ale

[16] Tato publikace je podrobněji představena v oddílu 1.3.

velmi stručné. Nejedná se o komplexnější výzkum, cílem příspěvku je spíše naznačit, jakým způsobem a k jakému typu výzkumů je možné užívat jazykové korpusy. Autorka zároveň podotýká, že porovnáním psaných a mluvených korpusů lze získat ucelenější přehled o rozdílech mezi psaným a mluveným diskursem (Adamovičová, 2017, s. 59).

Podobně A. Čermáková a M. Kopřivová (2018) srovnávají v rámci svého výzkumu 50 nejfrekventovanějších slov v psaném korpusu SYN2015 a v mluveném korpusu ORAL. Toto srovnání však není hlavním cílem jejich práce, jedná se pouze o nástin lexikálního repertoáru v psaném a především mluveném jazyce. Autorky dochází k několika následujícím zjištěním: mezi nejčastější slova v obou korpusech patří převážně synsémantika; ve frekvenční špičce slovní zásoby mluvené češtiny se nacházejí slova výrazně interaktivního charakteru (např. *jo*, *no*, *hmm*), výrazy poukazující na participaci více mluvčích (např. *já*, *ty*, *von*), vycpávková slova a výrazy s deiktickou funkcí zdůrazňující důležitost celého mimojazykového kontextu (např. *tam*, *sem*, *ten*) (Čermáková – Kopřivová, 2018, s. 229).

V oblasti české jazykovědy je především v posledních letech věnována pozornost velmi frekventovaným výrazům s různými funkcemi (a to jak při srovnání tohoto výrazu v mluvené a psané komunikaci, tak i v rámci jazyka obecně). Zdrojem pro tyto výzkumy jsou především jazykové korpusy. Příkladem takového výzkumu je již zmiňovaná práce P. Klimešové a kol. (2015), ve které autoři pomocí korpusových dat zkoumají jednotlivé funkce výrazu *to*. Další podobnou prací je článek od J. Hoffmannové s názvem *„Malá slova“ a jejich podíl na stylu mluvených projevů* (2017), ve kterém se věnuje synsémantikům s různými funkcemi (např. výrazům *no*, *jo*, *tak*, *jako*). Funkce těchto výrazů vyděluje do dvou skupin. První skupinu tvoří slova, která mají racionální funkci, to znamená, že často uvozují, upřesňují a vysvětlují, jsou často užívána při opravách v promluvách a vyvozují důsledek či závěr promluvy. Druhou skupinu funkcí autorka pojmenovává jako neracionální. Výrazy v této skupině jsou často používány jako vycpávková slova, exponenty neurčitosti, vyjadřují váhání mluvčího, podílejí se na postupném formulování či hledání výrazu (Hoffmannová, 2017). Výrazy *no* a *jo* se ve svém výzkumu zabývá L. Jílková (2017). Autorka zkoumá, zda existuje nějaká korelace mezi významem a funkcí slov *jo* a *no* a jejich přízvukováním. V závěru dochází ke zjištění, že zcela jednoznačné, obecně platné korelace mezi ne/přízvukováním a jistou funkcí či významem slov *no* a *jo* vysledovat nelze. Zároveň ale upozorňuje, že ne/přízvukování těchto výrazů může být často individuální.

Autorky A. Čermáková, L. Jílková, Z. Komrsková, M. Kopřivová a P. Poukarová (2019) se ve své práci (která byla zmíněna v oddílu 4.1) věnují diskurzním markerům v mluvené komunikaci. Na základě analýzy náhodného vzorku 100 konkordančních řádků z korpusů mluvené češtiny autorky klasifikovaly vybrané výrazy do několika skupin a popsaly, jaké mají zkoumané výrazy vlastnosti a jak se užívají v jazyce. Podle funkce jednotlivých diskurzních markerů vyčlenily tři skupiny: interakční diskurzní markery (vztahují se k adresátovi), modální diskurzní markery (vztahují se ke sdělení) a konektivní diskurzní markery (vztahují se k organizaci projevu). Autorky se ve své práci zabývají výrazy, které jsou v mluveném jazyce velmi frekventované, některé z nich budou zkoumány i v této diplomové práci. Zjištění autorek, která se vztahují ke konkrétním výrazům, jsou podobněji popsána v části věnující se vlastnímu výzkumu (u popisu jednotlivých lemmat).

Podobnému tématu, které zpracovává tato diplomová práce, se věnuje J. Allwood (1998), který srovnává 10 nejfrekventovanějších slov v psaném a mluveném korpusu švédštiny. V závěru dochází k několika následujícím zjištěním: 95 % zkoumaných slov jsou synsémantika (jak v psaném, tak v mluveném korpusu); nejčastějšími slovními druhy v psaném i mluveném korpusu jsou slovesa a zájmena; při mluvené komunikaci používáme přibližně polovinu lemmat z těch, která používáme při psané komunikaci; zájmena, spojky, adverbia a citoslovce se častěji užívají v mluvené komunikaci, naopak předložky, substantiva a adjektiva v psané komunikaci (Allwood, 1998, s. 28–29).

Slovní zásobou mluvené komunikace se zabývá M. McCarthy (2001). Ve svém výzkumu se zaměřuje na 2000 nejfrekventovanějších lemmat v korpusu CANCODE (Cambridge and Nottingham Corpus of Discourse in English), tato lemmata považuje za jádro mluvené slovní zásoby. Na základě jejich analýzy konstatuje, že lemmata tvořící jádro slovní zásoby mluvené komunikace je možné rozdělit do devíti následujících skupin: 1. modální slovní zásoba (obsahuje výrazy, které vyjadřují určitý stupeň jistoty či nutnosti, např. *can*, *must*, *maybe*); 2. slovesa s oslabeným lexikálním významem (tato slovesa nabývají významu se substantivy, předložkami či částicemi, např. *do*, *make*, *get*); 3. interaktivní slova (vyjadřují postoje a stavy mluvčích, např. *a bit*, *actually*, *really*); 4. diskurzní markery (např. *well*, *right*, *anyway*); 5. základní nominální pojmy (konkrétní a abstraktní substantiva, např. *school*, *life*, *car*); 6. deiktické výrazy (spojují mluvčího s okolním světem, např. *this*, *now*, *here*); 7. základní adjektiva (používaná k hodnocení lidí, situací, událostí či věcí, např. *bad*, *nice*, *lovely*);

8. základní adverbia (často vyjadřují čas, způsob či zvyklost, např. *today*, *quickly*, *usually*);
9. základní slovesa (pojmenovávají aktivity a činnosti, např. *eat*, *sit*, *listen*) (McCarthy, 2001, s. 31–35).

V předchozích odstavcích jsem uvedla výzkumy lingvistů, kteří se zabývají podobným tématem jako tato diplomová práce. Z některých výše představených prací vycházím ve vlastním výzkumu, který je popsán v následující části.

# Praktická část

# 6 Vlastní výzkum

V této části se věnuji vlastnímu výzkumu. Nejprve představím koncept výzkumu – na jaká lemmata se zaměřuji, jakým způsobem jsem získala potřebná data, jaké metodologické problémy bylo potřeba řešit apod. V další části analyzuji jednotlivá lemmata, soustředím se na jejich význam a zkoumám, zda se významy v mluvené a psané komunikaci shodují, či liší. V závěru shrnu zásadní poznatky a zjištění, ke kterým jsem dospěla na základě analýzy zkoumaného materiálu.

Na tomto místě je důležité zmínit jednu terminologickou poznámku. Jak již bylo řečeno, tato práce se zabývá 20 nejfrekventovanějšími lemmaty v korpusech SYN2020 a ORAL. Tato lemmata ve své práci nazývám jako *frekvenční špička*. Toto pojmenování jsem se rozhodla používat jednak z ekonomického důvodu (jde o kratší spojení než *výrazy které spadají mezi 20 nejfrekventovanějších lemmat*), jednak je tento termín z oblasti korpusové lingvistiky snadno a intuitivně pochopitelný.

## 6.1 Popis výzkumu

### 6.1.1 Zkoumaná lemmata

Ve svém výzkumu se zaměřuji na 20 nejfrekventovanějších lemmat v psaném korpusu SYN2020 a na 20 nejfrekventovanějších lemmat v mluveném korpusu ORAL. V rámci svého výzkumu chci podrobněji prozkoumat frekvenční špičku (tedy malé množství velmi frekventovaných lemmat) mluvené a psané komunikace. Z toho důvodu jsem se rozhodla analyzovat 20 nejfrekventovanějších lemmat v obou zmíněných korpusech. Zároveň se jedná o dostatečné množství lemmat, ze kterého, jak dále ukážu, lze vyvozovat obecnější závěry o mluvené a psané komunikaci. Počet zkoumaných lemmat byl stanoven také s přihlédnutím k rozsahu diplomové práce.

Další variantou, jak zkoumat frekvenční špičku zmíněných korpusů, by bylo zaměřit se ne na lemmata, ale na slovní tvary. V takovém případě by začátek seznamu nejfrekventovanějších slovních tvarů tvořila slova neohebná (která mají obvykle pouze jeden tvar)[17]. Dále by v seznamu byly nejfrekventovanější tvary ohebných slov (v podstatě by se tedy jednalo o nejčastější tvary nejfrekventovanějších lemmat). Seznam 20 nejfrekventovanějších slovních

[17] Zde je důležité upozornit, že i některá neohebná slova mají v korpusech SYN2020 a ORAL více tvarů, např. předložka *v* má tvary *v* a *ve*.

tvarů by se však příliš nelišil od výrazů, kterými se zabývá tato práce. Ve svém výzkumu jsem se rozhodla zaměřit se na nejfrekventovanější lemmata (a ne slovní tvary) především z toho důvodu, že v rámci analýzy jednotlivých lemmat jsou reflektovány všechny jejich tvary (včetně nejčastějších). Domnívám se, že při zkoumání frekvenční špičky dokáže analýza lemmat přinést širší a komplexnější poznatky než analýza slovních tvarů.

Potřebná data k výzkumu jsem získala z databáze Českého národního korpusu. V rámci popisu psané komunikace jsem vycházela z korpusu SYN2020. Jedná se o synchronní reprezentativní a referenční korpus současné psané češtiny, který obsahuje více než 100 milionů textových slov. V korpusu SYN2020 jsou ve stejném poměru zastoupeny tři následující textové makrotypy – beletrie, oborová literatura a publicistika. Značnou výhodou tohoto korpusu je, že se jedná o lemmatizovaný a morfologicky anotovaný korpus, navíc obsahuje také syntaktické značkování.[18]

V rámci zkoumání mluvené komunikace jsem vycházela z korpusu ORAL. Jedná se o korpus transkriptů nahrávek převážně neformálních rozhovorů rodilých mluvčích češtiny. Nahrávky byly pořízeny během let 2002–2011. Korpus obsahuje přibližně 5,4 milionů slov a je lemmatizovaný a morfologicky značkovaný, oproti korpusu SYN2020 ale neobsahuje syntaktické značkování. Jedním z nedostatků tohoto korpusu je to, že není vyvážený, převažují v něm data z české části České republiky.[19]

Oba tyto korpusy jsou vhodné ke zkoumání frekvenčních špiček mluvené a psané komunikace. Zachycují současné a reálné užívání českého jazyka a obsahují velké množství snadno dostupných dat. Podávají informace o frekvenci jazykových jevů, což je pro tento výzkum zásadní.

V souvislosti s použitým materiálem je důležité upozornit na problém, který jsem již zmiňovala v oddílu 2.1. Vytvořit ideální vzorek, který by celkově reflektoval psanou a mluvenou komunikaci, je prakticky nemožné (například už jenom kvůli tomu, že vymezit hranice psanosti a mluvenosti je poměrně složité). Materiál, se kterým pracuji, odráží pouze určitou část psané a mluvené komunikace. To se týká především popisu mluvené komunikace, jelikož korpus ORAL zachycuje pouze neformální konverzaci s přáteli, která je některými autory považována za prototyp mluvenosti (srov. Čermák, 2009; Hoffmannová, 1992). Některé závěry ohledně

[18] Informace ke korpusu SYN2020 byly čerpány z Příručky ČNK, dostupné na adrese https://wiki.korpus.cz/doku.php/cnk:syn2020 [Poslední přístup 24. 10. 2022].
[19] Informace ke korpusu ORAL byly čerpány z Příručky ČNK, dostupné na adrese https://wiki.korpus.cz/doku.php/cnk:oral [Poslední přístup 24. 10. 2022].

mluvené komunikace, ke kterým ve svém výzkumu docházím, jsou zapříčiněny především tím, že vycházím z materiálu dialogické povahy (a ne z materiálu, který zachycuje např. mluvenou přednášku, tedy formální text monologického charakteru).

### 6.1.2 Získání dat

Potřebná data jsem získala pomocí funkce Seznam slov webového rozhraní KonText. Použití této funkce je oproti běžnému dotazování v korpusu (jehož výsledkem je konkordance) výhodné v tom, že celý proces vyhledávání je rychlejší a výpočetně méně náročnější. Pomocí této funkce jsem získala seznam nejfrekventovanějších lemmat v korpusech SYN2020 a ORAL. Oba seznamy zároveň obsahovaly informace o absolutní frekvenci a ranku (pořadí podle frekvence) každého lemmatu.

Takto získané seznamy lemmat byly následně manuálně upraveny. Ke každému lemmatu jsem doplnila informaci o jeho relativní frekvenci (uvedenou v i.p.m., tedy v počtu výskytů slova v textu o délce 1 milionu slov). Relativní frekvence je zaokrouhlena na celá čísla. Díky této frekvenci je možné srovnávat lemmata z obou korpusů, které mají odlišnou velikost.[20]

Vzhledem k odlišnému způsobu lemmatizace některých jednotek v obou použitých korpusech bylo potřeba jednotlivá lemmata sjednotit. Tato úprava se týkala tří následujících lemmat: *být*, *on* a *co*. Pod lemma *být* spadají v korpusu SYN2020 i výrazy jako *aby, abychom*, *kdyby*, *kdybychom*, *jakoby*, *ses*, *sis* apod. V korpusu ORAL jsou tyto výrazy lemmatizovány odlišným způsobem (ve většině případů tvoří samostatná lemmata). Z důvodu sjednocení těchto lemmat byly z analyzovaného materiálu z korpusu SYN2020 vyřazeny výše zmíněné výrazy, které v korpusu ORAL spadají pod jiná lemmata. Touto úpravou se v rámci seznamu 20 nejfrekventovanějších lemmat v korpusu SYN2020 snížila frekvence lemmatu *být*, nicméně jeho pořadí (rank) v tomto seznamu zůstalo stejné.[21]

Pod lemmatem *on* jsou v korpusu SYN2020 přiřazeny pouze tvary singuláru (*on*, *ona*, *ono* a jejich příslušné pádové tvary). Pod stejným lemmatem jsou v korpusu ORAL přiřazeny jak tvary singuláru, tak plurálové tvary (*oni*, *ony*, *ona* a jejich příslušné pádové tvary). Z důvodu sjednocení těchto lemmat byly pod lemma *on* přiřazeny zmíněné plurálové tvary. Tím se

[20] Podrobněji o relativní frekvenci viz oddíl 1.2.

[21] Pro tuto úpravu jsem použila dotaz [lemma="být"&word!="[Aa].*|[Kk]dyb.*|[Ss]es|[Ss]is|[Jj]ako.*"], který zajistil vyřazení zmíněných výrazů spadajících pod lemma *být* v korpusu SYN2020.

v rámci seznamu 20 nejfrekventovanějších lemmat v korpusu SYN2020 zvýšila frekvence tohoto lemmatu a změnil se jeho rank (z hodnoty 8 na hodnotu 7).[22]

Poslední úprava se týkala lemmatu *co*. V korpusu SYN2020 spadá pod lemma *co* i výraz *což* (a jeho další tvary, např. *čehož*, *čemuž*). Jedná se však o odlišná lemmata, ve výkladových slovnících (např. v *Akademickém slovníku současné češtiny*) jsou charakterizována jako dvě samostatná hesla. Z toho důvodu byl ze získaných dat z korpusu SYN2020 odstraněn výraz *což* (a jeho další tvary). Tím se snížila frekvence lemmatu *co* a jeho rank byl změněn z hodnoty 20 na hodnotu 24. Mezi 20 nejfrekventovanějších lemmat korpusu SYN2020 se tak dostalo lemma *moci*, které bylo před touto úpravou na 21. místě. V korpusu ORAL je výraz *což* anotován jako samostatné lemma, tato úprava se tedy dat získaných z korpusu ORAL netýkala.[23]

Takto získaná a upravená data byla použita pro tento výzkum. Před představením samotné analýzy dat je důležité zmínit problém lemmatizace víceslovných spojení (např. některých frazémů, víceslovných předložek či víceslovných spojek) v korpusech SYN2020 a ORAL. Víceslovné jednotky jsou rozděleny na jednotlivé komponenty, ty jsou pak přiřazeny ke konkrétním lemmatům. Například víceslovná předložka *v rámci* je rozdělena na komponenty *v* a *rámci*, které jsou přiřazeny pod lemmata *v* a *rámec*. Víceslovné jednotky tak netvoří víceslovná lemmata.

## 6.2 Analýza materiálu

V této kapitole představím frekvenční špičky korpusů SYN2020 a ORAL. Následně obě tyto frekvenční špičky porovnám. Při srovnání se soustředím především na sémantické vlastnosti zkoumaných lemmat (tedy na jednotlivé významy a funkce), ale zároveň se zabývám i některými jevy syntaktickými či morfologickými, které často se sémantickými vlastnostmi souvisí.

U každého lemmatu uvedu jeho základní významy, při tomto popisu vycházím ze slovníků ASSČ, SSČ a SSJČ. Následně představím, jak se dané lemma užívá v korpusech SYN2020 a ORAL. Za tímto účelem u většiny zkoumaných lemmat provádím analýzu na náhodném vzorku z korpusu SYN2020 a ORAL. Velikost vzorku se pohybuje v rozmezí 100–300 konkordančních řádků ve zmíněných korpusech. K ověření dostatečné velikosti jednotlivých vzorků byla použita webová aplikace ČNK Korpusová kalkulačka[24] (Cvrček, 2019). Při analýze

[22] Pro tuto úpravu jsem použila dotaz [lemma="on|oni"], který zajistil, že pod výraz *on* spadaly i plurálové tvary, které jsou v korpusu SYN2020 anotovány jako samostatné lemma *oni*.
[23] Pro tuto úpravu jsem použila dotaz [lemma="co"&word!="[cC]ož|[čČ]ehož|[čČ]emuž|[čČ]emž|[Čč]ímž"], který zajistil, že pod lemma *co* nespadal výraz *což* (a jeho ostatní tvary).
[24] Jedná se o aplikaci, která uživatelům korpusů poskytuje pomoc při počítání základních statistických úloh. Aplikace je dostupná na adrese https://www.korpus.cz/calc/.

významů jednotlivých lemmat vycházím především z významů popsaných ve výkladových slovnících, případně z odborných prací, které se zkoumaným výrazům věnují.

Cílem této analýzy není detailně popsat všechny významy, které mohou daná lemmata vyjadřovat, ani určit slovnědruhovou platnost těchto lemmat, ale představit základní významy a především zjistit, v jakých významech se daná lemmata nejčastěji užívají v psané a mluvené komunikaci a zda se jejich užití v psané a mluvené komunikaci liší.

Při analýze materiálu často vycházím z morfologické anotace použitých korpusů. Jedná se o automatickou anotaci, která nemusí být vždy správná. Především slovnědruhová anotace zkoumaných lemmat není dostatečná. V případě korpusu ORAL může být chybná anotace způsobena nedodržením pravopisných norem při transkribování mluveného textu. Na chybnou anotaci, která by zásadním způsobem mohla ovlivnit analýzu, upozorňuji u konkrétních výrazů.

V rámci popisu lemmat často uvádím příklady z korpusů SYN2020 a ORAL. Tyto příklady nechávám v původní podobě, tzn. včetně pravopisných či typografických chyb, v případě korpusu ORAL také včetně doprovodných poznámek. V použitých příkladech je tučně označen výraz, ke kterému se příklad vztahuje.

### 6.2.1 Frekvenční špička korpusu SYN2020

V tabulce č. 2 je zobrazena frekvenční špička korpusu SYN2020. V prvním sloupci je uvedeno lemma, v druhém rank (pořadí), ve třetím absolutní frekvence (AF) lemmatu, ve čtvrtém relativní frekvence (RF, uvedeno v i.p.m.). Poslední sloupec uvádí rank konkrétního lemmatu v korpusu ORAL. Tučně jsou vyznačena ta lemmata, která se nacházejí jak ve frekvenční špičce korpusu SYN2020, tak i ve frekvenční špičce korpusu ORAL. Lemmata, která jsou vyznačena kurzivou, se nacházejí pouze ve frekvenční špičce korpusu SYN2020. Jednotlivá lemmata (jejich funkce a významy) jsou popsána v části 6.2.3.

Jak již bylo řečeno, v korpusu SYN2020 jsou ve stejném poměru zastoupeny tři následující textové makrotypy – beletrie, oborová literatura a publicistika. Abych zjistila, zda vysoká frekvence nějakého lemmatu není způsobena tím, že výrazně převažuje v jednom textovém makrotypu, srovnala jsem frekvenční špičky všech tří textových makrotypů zastoupených v korpusu SYN2020. Toto srovnání zobrazuje tabulka č. 3. První dva sloupce tabulky představují celkovou frekvenční špičku korpusu SYN2020, další sloupce pak frekvenční špičky konkrétních textových makrotypů. Lemmata, která jsou vyznačena tučně, se nacházejí ve všech třech frekvenčních špičkách textových makrotypů. Lemmata, která jsou vyznačena kurzivou, se nacházejí ve dvou frekvenčních špičkách, zbylá lemmata (která nejsou vyznačena tučně či kurzivou) se nacházejí pouze v jedné frekvenční špičce.

| 20 nejfrekventovanějších lemmat v korpusu SYN2020 | | | | |
|---|---|---|---|---|
| **rank** | **lemma** | **AF** | **RF** | **rank v ORAL** |
| 1 | **být** | 4 012 251 | 32 934 | 2 |
| 2 | **a** | 3 171 687 | 26 034 | 3 |
| 3 | **se** | 3 099 618 | 25 443 | 5 |
| 4 | **v** | 2 310 188 | 18 963 | 16 |
| 5 | **na** | 1 874 140 | 15 385 | 14 |
| 6 | **ten** | 1 736 732 | 14 256 | 1 |
| 7 | **on** | 1 308 493 | 10 741 | 7 |
| 8 | *s* | 982 890 | 8 068 | 27 |
| 9 | **že** | 965 418 | 7 925 | 9 |
| 10 | *z* | 854 388 | 7 013 | 31 |
| 11 | *který* | 784 796 | 6 442 | 80 |
| 12 | **mít** | 745 384 | 6 118 | 11 |
| 13 | *do* | 690 059 | 5 664 | 28 |
| 14 | *o* | 584 782 | 4 800 | 60 |
| 15 | *i* | 576 218 | 4 730 | 70 |
| 16 | **ale** | 571 713 | 4 693 | 15 |
| 17 | *k* | 548 068 | 4 499 | 76 |
| 18 | **já** | 508 431 | 4 173 | 6 |
| 19 | **jako** | 427 921 | 3 513 | 12 |
| 20 | *moci* | 411 117 | 3 375 | 38 |

*Tabulka 2: 20 nejfrekventovanějších lemmat v korpusu SYN2020*

V tabulce č. 3 můžeme vidět, že frekvenční špičky jednotlivých textových makrotypů jsou velmi podobné. Prvních 14 nejfrekventovanějších lemmat korpusu SYN2020 se nachází ve všech frekvenčních špičkách textových makrotypů zastoupených v tomto korpusu. Dalším výrazem, který se nachází ve všech těchto frekvenčních špičkách, je lemma *k*. Lemmata *i*, *ale* a *jako* se nacházejí ve dvou frekvenčních špičkách. Pouze v jedné frekvenční špičce se nachází lemma *já* (konkrétně v beletrii) a lemma *moci* (konkrétně v oborové literatuře). Lemmata, která tvoří celkovou frekvenční špičku korpusu SYN2020, se zároveň nacházejí mezi 34 nejfrekventovanějšími lemmaty v beletrii, oborové literatuře a publicistice. Jinak řečeno – lemmata, jejichž rank se pohybuje v korpusu SYN2020 v rozmezí 1–20, mají v jednotlivých textových makrotypech zastoupených v tomto korpusu hodnotu ranku nejvýše 34. To znamená, že celková frekvenční špička korpusu SYN2020 neobsahuje žádné lemma, které by mělo výrazně vyšší či výrazně nižší frekvenci v jednom z těchto textových makrotypů. V této

frekvenční špičce se tak nenachází lemma, které by bylo typické pouze pro jeden textový makrotyp.

| 20 nejfrekventovanějších lemmat v textových makrotypech zastoupených v korpusu SYN2020 | | | | | | | |
|---|---|---|---|---|---|---|---|
| **SYN2020** | | **SYN2020 beletrie** | | **SYN2020 oborová literatura** | | **SYN2020 publicistika** | |
| **rank** | **lemma** | **rank** | **lemma** | **rank** | **lemma** | **rank** | **lemma** |
| 1 | **být** | 1 | **být** | 1 | **být** | 1 | **být** |
| 2 | **a** | 2 | **se** | 2 | **a** | 2 | **a** |
| 3 | **se** | 3 | **a** | 3 | **v** | 3 | **v** |
| 4 | **v** | 4 | **ten** | 4 | **se** | 4 | **se** |
| 5 | **na** | 5 | **on** | 5 | **na** | 5 | **na** |
| 6 | **ten** | 6 | **na** | 6 | **ten** | 6 | **ten** |
| 7 | **on** | 7 | **v** | 7 | **s** | 7 | **s** |
| 8 | **s** | 8 | **že** | 8 | **který** | 8 | **on** |
| 9 | **že** | 9 | já | 9 | **z** | 9 | **z** |
| 10 | **z** | 10 | **s** | 10 | **že** | 10 | **který** |
| 11 | **který** | 11 | **mít** | 11 | **on** | 11 | **že** |
| 12 | **mít** | 12 | **z** | 12 | *i* | 12 | **mít** |
| 13 | **do** | 13 | **do** | 13 | **k** | 13 | *i* |
| 14 | **o** | 14 | *ale* | 14 | **o** | 14 | **do** |
| 15 | *i* | 15 | **který** | 15 | **mít** | 15 | **o** |
| 16 | *ale* | 16 | co | 16 | **do** | 16 | *ale* |
| 17 | **k** | 17 | **k** | 17 | tento | 17 | rok |
| 18 | já | 18 | *jako* | 18 | *pro* | 18 | **k** |
| 19 | *jako* | 19 | **o** | 19 | moci | 19 | *pro* |
| 20 | moci | 20 | když | 20 | *jako* | 20 | za |

*Tabulka 3: 20 nejfrekventovanějších lemmat v textových makrotypech zastoupených v korpusu SYN2020*

### 6.2.2 Frekvenční špička korpusu ORAL

V tabulce č. 4 je zobrazena frekvenční špička korpusu ORAL. V prvním sloupci je uvedeno lemma, v druhém rank (pořadí), ve třetím absolutní frekvence (AF) lemmatu, ve čtvrtém relativní frekvence (RF, uvedeno v i.p.m.). Poslední sloupec uvádí rank konkrétního lemmatu v korpusu SYN2020. Tučně jsou vyznačena ta lemmata, která se nacházejí jak ve frekvenční špičce korpusu ORAL, tak i ve frekvenční špičce korpusu SYN2020. Lemmata, která jsou vyznačena kurzivou, se nacházejí pouze ve frekvenční špičce korpusu ORAL. Jednotlivá lemmata (jejich funkce a významy) jsou popsána v části 6.2.3.

| 20 nejfrekventovanějších lemmat v korpusu ORAL | | | | |
|---|---|---|---|---|
| **rank** | **lemma** | **AF** | **RF** | **rank v SYN2020** |
| 1 | **ten** | 484 824 | 76 210 | 6 |
| 2 | **být** | 376 383 | 59 164 | 1 |
| 3 | **a** | 181 411 | 28 516 | 2 |
| 4 | *no* | 154 513 | 24 288 | 212 |
| 5 | **se** | 142 116 | 22 339 | 3 |
| 6 | **já** | 124 772 | 19 613 | 18 |
| 7 | **on** | 124 693 | 19 601 | 7 |
| 8 | *tak* | 116 219 | 18 269 | 26 |
| 9 | **že** | 112 707 | 17 717 | 9 |
| 10 | *jo* | 98 988 | 15 560 | 579 |
| 11 | **mít** | 87 086 | 13 689 | 12 |
| 12 | **jako** | 87 062 | 13 685 | 19 |
| 13 | *tam* | 83 238 | 13 084 | 80 |
| 14 | **na** | 75 162 | 11 815 | 5 |
| 15 | **ale** | 62 406 | 9 810 | 16 |
| 16 | **v** | 57 072 | 8 971 | 4 |
| 17 | *hmm* | 50 095 | 7 875 | 4 104 |
| 18 | *vědět* | 45 001 | 7 074 | 57 |
| 19 | *ne* | 44 343 | 6 970 | 90 |
| 20 | *co* | 38 687 | 6 081 | 24 |

*Tabulka 4: 20 nejfrekventovanějších lemmat v korpusu ORAL*

Jak již bylo řečeno, korpus ORAL je tvořen přepisy nahrávek neformální konverzace mezi přáteli. Jedná se tak o mluvený korpus, který nereflektuje mluvenou komunikaci obecně, ale pouze její část, která by mohla být považována za prototyp mluvenosti (o prototypu mluvenosti podrobněji viz oddíl 2.1). Zařazení některých lemmat mezi frekvenční špičku korpusu ORAL může být způsobeno právě dialogickou povahou tohoto korpusu. Zajímavé je v tomto ohledu srovnání frekvenčních špiček korpusů ORAL a ORATOR. Korpus ORATOR obsahuje monologické projevy, na které se mluvčí mohli připravit, např. přednášky, přípitky, úvody společenských akcí apod. Jedná se tak o poměrně formální, až oficiální situace. Korpus ORATOR není vyvažován podle žádného z demografických kritérií. Obsahuje přibližně 1,2 milionů slov, je tedy výrazně menší než korpus ORAL.[25]

[25] Informace ke korpusu ORATOR byly čerpány z Příručky ČNK, dostupné na adrese https://wiki.korpus.cz/doku.php/:cnk:orator [Poslední přístup 29. 10. 2022].

Srovnání frekvenčních špiček obou těchto mluvených korpusů zobrazuje tabulka č. 5. Tyto frekvenční špičky tvoří 15 shodných lemmat (v tabulce jsou vyznačena tučně) a 5 lemmat se liší (v tabulce jsou vyznačena kurzivou). Společná lemmata by mohla být považována za nejfrekventovanější lemmata v mluvené komunikaci obecně. Odlišná lemmata jsou typická pro konkrétní mluvenou komunikační situaci, která je v korpusu reflektována. Přítomnost lemmat *no*, *jo*, *hmm*, *vědět* a *ne* ve frekvenční špičce korpusu ORAL je zřejmě způsobena dialogickou povahou tohoto korpusu.

| **20 nejfrekventovanějších lemmat v korpusech ORAL a ORATOR** | | | |
|---|---|---|---|
| **ORAL** | | **ORATOR** | |
| **rank** | **lemma** | **rank** | **lemma** |
| 1 | **ten** | 1 | **ten** |
| 2 | **být** | 2 | **být** |
| 3 | **a** | 3 | **a** |
| 4 | *no* | 4 | **se** |
| 5 | **se** | 5 | **v** |
| 6 | **já** | 6 | **že** |
| 7 | **on** | 7 | **tak** |
| 8 | **tak** | 8 | **na** |
| 9 | **že** | 9 | *který* |
| 10 | *jo* | 10 | **mít** |
| 11 | **mít** | 11 | **on** |
| 12 | **jako** | 12 | **jako** |
| 13 | **tam** | 13 | **ale** |
| 14 | **na** | 14 | *z* |
| 15 | **ale** | 15 | **tam** |
| 16 | **v** | 16 | **já** |
| 17 | *hmm* | 17 | *tady* |
| 18 | *vědět* | 18 | **co** |
| 19 | *ne* | 19 | *vlastně* |
| 20 | **co** | 20 | *s* |

*Tabulka 5: 20 nejfrekventovanějších lemmat v korpusech ORAL a ORATOR*

### 6.2.3 Srovnání frekvenčních špiček korpusu SYN2020 a korpusu ORAL

V úvodu této části srovnám frekvenční špičky korpusů SYN2020 a ORAL. Následně podrobněji popíšu jednotlivá lemmata, která tyto špičky tvoří. Z důvodu lepší přehlednosti jsem

zkoumaná lemmata rozdělila do tří následujících skupin: (1) Lemmata nacházející se ve frekvenční špičce obou korpusů; (2) Lemmata nacházející se pouze ve frekvenční špičce korpusu SYN2020; (3) Lemmata nacházející se pouze ve frekvenční špičce korpusu ORAL.

Tabulka č. 6 zobrazuje frekvenční špičky korpusů SYN2020 a ORAL. Levá část tabulky představuje data získaná z korpusu SYN2020, pravá část tabulky zobrazuje data získaná z korpusu ORAL. Tučně jsou vyznačena ta lemmata, která se nacházejí jak ve frekvenční špičce korpusu SYN2020, tak i ve frekvenční špičce korpusu ORAL. Lemmata, která jsou vyznačena kurzivou, se nacházejí pouze v jedné frekvenční špičce.

| **20 nejfrekventovanějších lemmat v korpusech SYN2020 a ORAL** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **korpus SYN2020** | | | | | **korpus ORAL** | | | | |
| **rank** | **lemma** | **AF** | **RF** | **rank v ORAL** | **rank** | **lemma** | **AF** | **RF** | **rank v SYN2020** |
| 1 | **být** | 4 012 251 | 32 934 | 2 | 1 | **ten** | 484 824 | 76 210 | 6 |
| 2 | **a** | 3 171 687 | 26 034 | 3 | 2 | **být** | 376 383 | 59 164 | 1 |
| 3 | **se** | 3 099 618 | 25 443 | 5 | 3 | **a** | 181 411 | 28 516 | 2 |
| 4 | **v** | 2 310 188 | 18 963 | 16 | 4 | *no* | 154 513 | 24 288 | 212 |
| 5 | **na** | 1 874 140 | 15 385 | 14 | 5 | **se** | 142 116 | 22 339 | 3 |
| 6 | **ten** | 1 736 732 | 14 256 | 1 | 6 | **já** | 124 772 | 19 613 | 18 |
| 7 | **on** | 1 308 493 | 10 741 | 7 | 7 | **on** | 124 693 | 19 601 | 7 |
| 8 | *s* | 982 890 | 8 068 | 27 | 8 | *tak* | 116 219 | 18 269 | 26 |
| 9 | **že** | 965 418 | 7 925 | 9 | 9 | **že** | 112 707 | 17 717 | 9 |
| 10 | *z* | 854 388 | 7 013 | 31 | 10 | *jo* | 98 988 | 15 560 | 579 |
| 11 | *který* | 784 796 | 6 442 | 80 | 11 | **mít** | 87 086 | 13 689 | 12 |
| 12 | **mít** | 745 384 | 6 118 | 11 | 12 | **jako** | 87 062 | 13 685 | 19 |
| 13 | *do* | 690 059 | 5 664 | 28 | 13 | *tam* | 83 238 | 13 084 | 80 |
| 14 | *o* | 584 782 | 4 800 | 60 | 14 | **na** | 75 162 | 11 815 | 5 |
| 15 | *i* | 576 218 | 4 730 | 70 | 15 | **ale** | 62 406 | 9 810 | 16 |
| 16 | **ale** | 571 713 | 4 693 | 15 | 16 | **v** | 57 072 | 8 971 | 4 |
| 17 | *k* | 548 068 | 4 499 | 76 | 17 | *hmm* | 50 095 | 7 875 | 4 104 |
| 18 | **já** | 508 431 | 4 173 | 6 | 18 | *vědět* | 45 001 | 7 074 | 57 |
| 19 | **jako** | 427 921 | 3 513 | 12 | 19 | *ne* | 44 343 | 6 970 | 90 |
| 20 | *moci* | 411 117 | 3 375 | 38 | 20 | *co* | 38 687 | 6 081 | 24 |

*Tabulka 6: 20 nejfrekventovanějších lemmat v korpusech SYN2020 a ORAL*

V tabulce č. 6 můžeme vidět, že frekvenční špičky korpusů SYN2020 a ORAL tvoří 12 shodných lemmat. Zbylá lemmata se vyskytují pouze v jedné frekvenční špičce. Zajímavé je, že všechna lemmata tvořící celkovou frekvenční špičku korpusu SYN2020 se zároveň nacházejí mezi 80 nejfrekventovanějšími lemmaty v korpusu ORAL (viz pátý sloupec). To znamená, že všechna lemmata tvořící frekvenční špičku korpusu SYN2020 jsou zároveň poměrně frekventovaná i v korpusu ORAL. Naopak ve frekvenční špičce korpusu ORAL jsou tři lemmata (*hmm*, *no* a *jo*), která mají v korpusu SYN2020 poměrně nízkou frekvenci (viz desátý sloupec).[26]

Informace o relativní frekvenci lemmat (sloupce nadepsané *RF*) potvrzují výše představené tvrzení (viz oddíl 4.4), že mluvený jazyk je více repetitivní a omezený než psaný. Relativní frekvence lemmat v korpusu ORAL je mnohem vyšší než relativní frekvence lemmat v korpusu SYN2020. To dokládá tezi D. Bibera a kol. (1999), že v mluvené komunikaci je méně typů slov, která ale mají vyšší frekvenci, naopak v psané komunikaci je více typů slov s obecně nižší frekvencí.

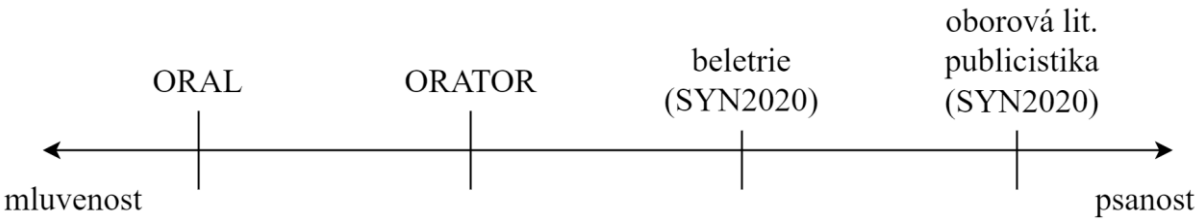
V teoretické části (v oddílu 2.1) jsem zmínila, že někteří autoři popisují komunikaci jako škálu. Na jednom konci škály je mluvenost a na druhém psanost. Mezi těmito póly se nenachází jasná hranice, ale přechodové pásmo. Na základě této teze jsem srovnala frekvenční špičky korpusů ORAL, ORATOR a textových makrotypů zastoupených v korpusu SYN2020 (tedy beletrie, oborové literatury a publicistiky). Výsledky tohoto srovnání zobrazuje tabulka č. 7. První sloupec a druhý řádek uvádí konkrétní korpus či textový makrotyp korpusu SYN2020. Čísla v tabulce vyjadřují počet shodných lemmat tvořících frekvenční špičku příslušných korpusů či makrotypů. Například frekvenční špičku korpusu ORAL a frekvenční špičku beletrie tvoří 13 shodných (a tedy 7 odlišných) lemmat, frekvenční špičku korpusu ORAL a oborové literatury tvoří 10 shodných (a tedy 10 odlišných) lemmat apod.

| **Počet shodných lemmat ve frekvenční špičce** | | | | | |
|---|---|---|---|---|---|
| | ORAL | ORATOR | BELETRIE | OBOROVÁ LIT. | PUBLICISTIKA |
| ORAL | X | 15 | 13 | 10 | 10 |
| ORATOR | 15 | X | 16 | 13 | 13 |
| BELETRIE | 13 | 16 | X | 16 | 16 |
| OBOROVÁ LIT. | 10 | 13 | 16 | X | 17 |
| PUBLICISTIKA | 10 | 13 | 16 | 17 | X |

*Tabulka 7: Počet shodných lemmat ve frekvenční špičce korpusů ORAL, ORATOR a textových makrotypů zastoupených v korpusu SYN2020*

[26] Mezi lemmata *jo*, *no* a *hmm* v korpusu SYN2020 byly zařazeny i jejich příslušné varianty, které se určitým způsobem snaží napodobit jejich zvukovou stránku, např. *no*, *nó*, *noo*, *hm*, *hmm*, *hmmm* apod. Tyto tvary jsou v korpusu SYN2020 lemmatizovány samostatně, v korpusu ORAL jsou ale přiřazeny pouze pod jedno lemma. Z důvodu sjednocení lemmat byly k těmto lemmatům v korpusu SYN2020 přiřazeny i jejich zmíněné varianty. Informace o ranku těchto lemmat v korpusu SYN2020 je spíše orientační.

Na základě údajů v tabulce č. 7 by jednotlivé korpusy a textové makrotypy zastoupené v korpusu SYN2020 mohly být na zmíněnou škálu komunikace umístěny následovně:




*Obrázek 2: Umístění korpusů ORAL, ORATOR a textových makrotypů zastoupených v korpusu SYN2020 na komunikační škálu*

Čím blíže jsou k sobě na škále umístěny jednotlivé korpusy či textové makrotypy, tím více se shodují jejich frekvenční špičky[27]. Nejvíce se shodují např. frekvenční špičky oborové literatury a publicistiky či frekvenční špičky korpusu ORATOR a beletrie. Frekvenční špička korpusu ORAL se nejvíce podobá frekvenční špičce korpusu ORATOR a nejméně frekvenční špičce oborové literatury a publicistiky.

#### 6.2.3.1 Lemmata nacházející se ve frekvenční špičce obou korpusů

V této části představím 12 následujících lemmat, která tvoří jak frekvenční špičku korpusu SYN2020, tak i frekvenční špičku korpusu ORAL: *a*, *ale*, *jako*, *že*, *ten*, *já*, *on*, *být*, *mít*, *na*, *v*, *se*. Některá z těchto lemmat se v jazyce chovají podobně, mají podobné funkce či významy. Z toho důvodu budou některá lemmata představena společně, čímž zamezím zbytečnému opakování, ke kterému by došlo, pokud bych jednotlivá lemmata představovala zvlášť.

##### 6.2.3.1.1 Lemmata *a*, *ale*, *jako*, *že*

Lemmata *a*, *ale*, *jako* a *že* se vyskytují ve frekvenční špičce obou korpusů. Kromě vysoké frekvence spojuje tato lemmata jejich podobná funkce. Všechna tato lemmata fungují v jazyce jako spojky (ať už souřadicí, či podřadicí), zároveň ale mohou fungovat jako částice, citoslovce či diskurzní markery[28] (zejména v mluvené komunikaci).

V následujících odstavcích představím jednotlivá lemmata. U každého lemmatu uvedu jeho základní významy, které jsou popsány v českých výkladových slovnících (vycházím zejména z ASSČ, SSČ a SSJČ). Následně představím analýzu těchto lemmat, kterou jsem provedla na náhodném vzorku o velikosti 200 konkordančních řádků z korpusů SYN2020 a ORAL. V této analýze vycházím především z významů popsaných ve výkladových slovnících, případně z odborných prací, které se zkoumaným výrazům věnují. Cílem této analýzy je zjistit odpovědi

[27] Pravděpodobně se budou více shodovat i jejich další rysy.
[28] Termín *diskurzní marker* je představen v oddílu 4.1.

na následující dotazy: V jakých významech se tato lemmata užívají v psané a mluvené komunikaci? Mají tato lemmata v psané a mluvené komunikaci stejné významy, nebo se liší (například jejich frekvenční distribucí)? Je nějaký význam typický pouze pro mluvenou (případně pro psanou) komunikaci?

**Lemma *a***

ASSČ charakterizuje výraz *a* jako souřadicí spojku (příp. součást spojkového výrazu), která vyjadřuje následující poměry mezi větami či větnými členy: slučovací, odporovací (*a přece*, *a přesto*, *a přitom*, synonymum výrazu *ale*), důsledkový (*a proto*, *a tak*, *a tedy*), stupňovací (*a navíc*, *a dokonce*) a vylučovací (*a/nebo*). Kromě funkce spojky uvádí ASSČ také funkci částice (*A co teď? Tak a dost!*) a citoslovce (*A, to jste vy?*).

Na základě analýzy vzorku o velikosti 200 konkordančních řádků z korpusů SYN2020 a ORAL vyplývá, že se lemma *a* nejčastěji užívá ve funkci spojky. V korpusu SYN2020 má tuto funkci 97 % analyzovaných výrazů, v korpusu ORAL 80 %. Ukázalo se, že poměrně velký počet výskytů lemmatu *a* v korpusu SYN2020 je součástí víceslovných spojek (např. *a proto*, *a přesto*, *a tedy*, *a tudíž*). Naopak v korpusu ORAL tvoří výraz *a* víceslovnou spojku jen zřídka. Spojka *a* se tedy v korpusu ORAL užívá především pro vyjádření slučovacího poměru, v korpusu SYN2020 i pro vyjádření dalších poměrů.

Zbylé případy je možné řadit mezi částice a citoslovce. Tuto skupinu tvoří 3 % analyzovaného vzorku v korpusu SYN2020 a 20 % v korpusu ORAL. U většiny případů není jednoznačné, zda se jedná o částici, či citoslovce. Tyto případy odpovídají charakteristice diskurzních markerů (představené v oddílu 4.1). Mluvčí tyto výrazy užívají především z pragmatických důvodů. V takových případech stojí výraz *a* obvykle na začátku promluvy. Mluvčí chce často navázat kontakt s komunikačním partnerem, upoutat jeho pozornost, případně mu dát najevo, že přebírá slovo (příklad 1). Pomocí výrazu *a* na začátku promluvy mluvčí navazuje na předchozí kontext. K tomuto užití často dochází v otázkách (příklad 2). Výraz *a* může dále signalizovat změnu tématu, jako je tomu v příkladu 3. Někdy se může podílet na víceslovných diskurzních markerech, např. *a jo*, *a tak* (ne ve významu důsledkového poměru), *no a (co)*, *jo a*, *hele a*.

(1) *A: furt dělali blbce jak sem je nepozvala na promoci ne . a já vůbec nevim že dokončila školu no tak*

*B:* ***a*** *to .* ***a*** *teď je magistra nebo co je ? .. magistra ? . vona je fakt magistra ? .. to bych do ni neřek* (ORAL)

(2) *A: když Honza nechce posunout termín Rumunska*

*B:* ***a*** *proč jako by to kvůli nám neposunuli ?* (ORAL)

(3) *A: pojedem jenom s Čendem*

*B: jo jo jo .* ***a*** *co zítra budeš doma ?* (ORAL)

**Lemma *ale***

ASSČ charakterizuje výraz *ale* jako souřadicí spojku (příp. součást spojkového výrazu), která vyjadřuje odporovací a stupňovací poměr mezi větnými členy či větami. Kromě spojovací funkce uvádí ASSČ také funkci částice (*To je ale hezké! Ale přijď brzy! Ale ne, tak to není!*) a citoslovce (*Ale, tak vy ji znáte? Ale ale, to se nedělá!*).

Výrazu *ale* se ve své práci o diskurzních markerech věnovaly autorky A. Čermáková, L. Jílková, Z. Komrsková, M. Kopřivová a P. Poukarová (2019). Na náhodném vzorku 100 konkordančních řádků z korpusu ORAL provedly analýzu výrazu *ale*. V této analýze rozlišovaly, zda je výraz *ale* užit v původním významu (za původní význam považovaly spojovací výraz s odporovacím významem), nebo jako diskurzní marker (kam byly řazeny výrazy s oslabenou konektivitou). Z této analýzy vyplynulo, že v 17 % dokladů bylo *ale* užito jako spojovací výraz (v původním významu) a 83 % dokladů bylo užito jako diskurzní marker (Čermáková a kol., 2019, s. 315).

Stejné třídění, které použily autorky práce o diskurzních markerech (2019), jsem použila ve své analýze lemmatu *ale*. K původnímu významu řadím i spojovací výrazy se stupňovacím významem, který autorky ve své práci nezmiňují (pravděpodobně se nevyskytoval v analyzovaném vzorku). V úvodu této analýzy je důležité zmínit, že v některých případech není zcela jednoznačné, zda výraz *ale* vyjadřuje původní význam, nebo zda dochází k jeho oslabení. V těchto případech je klasifikace výrazu *ale* poměrně subjektivní.

Z analýzy vzorku o velikosti 200 konkordančních řádků z korpusů SYN2020 a ORAL vyplývá, že v korpusu SYN2020 převažuje původní význam (71 % zkoumaných dokladů). Naopak v korpusu ORAL se výraz *ale* v původním významu užívá poměrně málo (21 % zkoumaných dokladů). Zatímco výraz *ale* v korpusu SYN2020 často vyjadřuje i stupňovací poměr (nalezeno 16 dokladů), ve zkoumaném vzorku z korpusu ORAL jsem pro výraz *ale* vyjadřující stupňovací poměr našla pouze jeden doklad (uveden v příkladu 4).

(4) *(...) některý uměj nahrávat jenom přímo na ty dývýdýčka . některý mají ještě jakoby v sobě disk takže uměj jednak nahrávat* ***ale*** *zároveň ty se můžeš třeba* - (ORAL)

Ve funkci diskurzního markeru je použito 79 % výrazů v analyzovaném vzorku z korpusu ORAL a 29 % z korpusu SYN2020. Oproti ostatním zkoumaným lemmatům (*a*, *jako*, *že*) spadá poměrně velké množství výrazu *ale* v korpusu SYN2020 mezi diskurzní markery. Značná část se jich objevuje v přímé řeči postav v beletrii (příklad 5) a v doslovných přepisech rozhovorů v publicistice (příklad 6). V korpusu ORAL mluvčí často užívají výraz *ale* na začátku promluvy. V takových případech může tento výraz vyjadřovat nesouhlas s obsahem předchozího výroku jiného mluvčího (příklad 7 a 8).

(5) *„Copak to je, co mi chceš povědět, Arture?“ opakovala paní Weasleyová nebezpečným hlasem. „****Ale*** *nic, Molly,“ zamumlal její manžel.* (SYN2020)

(6) *To znamená, že stejně jako my rozumíme slovenštině, tak u vás zase skoro každý rozumí češtině.* ***Ale*** *k tvému přání – Minecraftu jsme se v ábíčku věnovali už několikrát (...)* (SYN2020)

(7) *A: to už mě jenom tak jako kritizuje za mý kouření a -*

*B:* ***ale*** *kdepak já tě nekritizuju* (ORAL)

(8) *A: já si myslim . já sem teďkon nad tim špekuloval von ty děti choděj do školy tak voni vona přeci to nemůže pustit prvního června to dá tak na prvního července dyť ty děti musej fu* chodit do školy*

*B:* ***ale*** *ta holka řikala . řikala že devadesát dva je ročník tak ta už nemůže chodit na základní školu* (ORAL)

**Lemma *jako***

SSČ výraz *jako* charakterizuje jako podřadicí spojku, která má srovnávací význam (*zdravý jako ryba*, *velký jako já*) a uvádí doplňkové výrazy s identifikačním významem (*dojel jako první*). SSJČ výraz *jako* se srovnávacím významem charakterizuje jako příslovce, doplňkové výrazy s identifikačním významem popisuje jako spojku souřadicí. Oba slovníky uvádějí význam souřadicí (slučovací) spojky (*mladí jako staří*). SSJČ navíc u spojky souřadicí uvádí význam výčtu (*některé květiny, jako chrpa, koukol a vlčí mák, rostou v obilí*) a význam vysvětlení či doplnění (výraz *jako* je synonymní s výrazy *tedy* a *totiž*) se stylistickým kvalifikátorem „obecně s pokleslým významem“.

Výrazu *jako* se ve své práci věnuje J. Hoffmannová (2010), která zkoumá jeho fungování a distribuci v mluveném projevu. V práci uvádí, že tento výraz může mít v mluvené komunikaci následující funkce: signalizuje reprodukci cizí či vlastní řeči; vyjadřuje nejistotu a pochybnost; vysvětluje; uvozuje; doplňuje; funguje jako vycpávka. Výrazu *jako* se J. Hoffmannová věnuje i ve své pozdější práci (2017). V tomto článku popisuje již zmíněné funkce výrazu *jako*. V závěru dodává, že jednotlivé významy a funkce tohoto výrazu (a dalších analyzovaných výrazů, např. *teda*, *jakoby*, *jakože*) se v každodenních mluvených projevech velmi těžko rozeznávají, protože jsou často propleteny a vzájemně splývají (Hoffmannová, 2017, s. 168).

Dále se tomuto výrazu ve své práci o diskurzních markerech věnovaly autorky A. Čermáková, L. Jílková, Z. Komrsková, M. Kopřivová a P. Poukarová (2019). Na náhodném vzorku 100 konkordančních řádků z korpusu ORTOFON provedly analýzu výrazu *jako*. Z této analýzy vyplynulo, že v 10 % dokladů bylo *jako* užito v původním významu (srovnání či vysvětlení) a 90 % dokladů bylo užito jako diskurzní marker s členicí funkcí (Čermáková a kol., 2019, s. 318).

Podobné třídění, které použily autorky práce o diskurzních markerech (Čermáková a kol., 2019), jsem použila ve své analýze lemmatu *jako*. Za původní význam považuji význam srovnání, doplňkové výrazy s identifikačním významem a výrazy, které uvádí výčet.

Z analýzy vzorku o velikosti 200 konkordančních řádků z korpusů SYN2020 a ORAL vyplývá, že v korpusu SYN2020 výrazně převažují výrazy *jako* s původním významem (99 % zkoumaných dokladů). Nejčastěji vyjadřují srovnání (134 dokladů). Výrazů, které uvádějí výčet, jsem ve zkoumaném vzorku v korpusu SYN2020 našla pouze 10.

Naopak v korpusu ORAL se výraz *jako* v původním významu užívá velmi málo (7 % zkoumaných dokladů). V 11 dokladech se jedná o význam srovnání, 2 jsou užity jako doplňkové výrazy s identifikačním významem (příklad 9) a jeden uvádí výčet (příklad 10, ve kterém je elipsa nadřazeného pojmu).

(9) *A: do . kam pojede do Belgie*

*(...)*

*B: do Bruselu .* ***jako*** *voják* (ORAL)

(10) *A: tam že je ňáká skříňka na věci . mužou tam bejt takový* ***jako*** *stany spacáky . lyžáky .. lyže tam můžou ležet uplně v pohodě* (ORAL)

Mezi diskurzní markery je možné řadit pouhé 2 doklady z analyzovaného vzorku v korpusu SYN2020 a 93 % analyzovaných dokladů v korpusu ORAL. Jedná se tak o největší významový rozdíl, co se týče zkoumaných lemmat *a*, *ale*, *jako* a *že*. Oba doklady v korpusu SYN2020 jsou užity v přímé řeči a autor pomocí nich patrně napodobuje funkci vycpávky (příklad 11, kde jako vycpávkové výrazy fungují všechna čtyři *jako*[29]). Jak zmínila J. Hoffmannová (2017, s. 168), jednotlivé funkce výrazu *jako* se od sebe v mluvených projevech velmi těžko rozeznávají. V korpusu ORAL má většina zkoumaných výrazů funkci vycpávky (příklad 12, kde první a druhé *jako* funguje jako vycpávka, třetí *jako* vyjadřuje původní význam). Jak zmiňuje řada autorů (např. Hoffmannová – Komrsková – Poukarová, 2017; Hoffmannová, 2017), výrazy *jako* často signalizují či doprovázejí reprodukci cizí či vlastní řeči (příklad 13). V některých případech může výraz *jako* vyjadřovat emoce (překvapení, rozhořčení) mluvčího (příklad 14).

(11) *A když Morgain dojde do ložnice, tak se ji Ezech ptá:* ***Jako*** *líbím se ti* ***jako*** *takhle,* ***jako, jako****?* (SYN2020)

(12) *kov tam asi být* ***jako*** *může ale ne něco co by šlo* ***jako*** *použít* ***jako*** *zbraň nebo tak* (ORAL)

(13) *sem to tam jedla tak a von . no ne* ***jako*** *jestli sem* ***jako*** *neřikala že náhodou (se smíchem)* ***jako*** *že budu hubnout* (ORAL)

(14) *proboha milýho a proč ste volili proti* ***jako*** *?* (ORAL)

**Lemma *že***

V SSČ a SSJČ je výraz *že* charakterizován jako podřadicí spojka, která připojuje vedlejší větu obsahovou, prostředkovou, příčinnou, přípustkovou apod. SSJČ navíc zmiňuje, že podřadicí spojka *že* uvádí větu závislou na příslovečném výraze, obyčejně s oslabenou větnou platností, např. *jistě že už to víš*; *možná že budeš čekat*. Oba slovníky uvádí funkci částice, která vyjadřuje emoce (výčitku, zklamání, lítost), vyzývá k odpovědi (*je tam zima, že?*) a navazuje na předchozí sdělení (*kolik že vás bylo?*).

V rámci vlastní analýzy výrazu *že* jsem zkoumala, jak často tento výraz v korpusech SYN2020 a ORAL funguje jako spojka a jak často funguje jako částice (příp. diskurzní marker). Ve zkoumaném vzorku o velikosti 200 konkordančních řádků z korpusu SYN2020 jsem nenašla žádný doklad užití výrazu *že* ve funkci částice. Všechny zkoumané výrazy fungují jako spojka.

[29] Přestože je v tomto příkladu výraz *jako* užit čtyřikrát, zkoumaný vzorek obsahoval pouze první *jako*. Totéž platí i pro další příklady, ve kterých je tučně označeno více výrazů (zkoumaný vzorek obsahoval pouze jeden z označených výrazů).

Nejčastěji připojují vedlejší větu obsahovou (178 dokladů), ve 20 případech šlo o připojení věty účinkové či příčinné (obvykle pomocí spojení *proto, že*), ve 2 případech se jednalo o spojení *jasně že* a *samozřejmě že*.

V korpusu ORAL má funkci spojky 81 % zkoumaných výrazů. Podobně jako v korpusu SYN2020 se nejčastěji jedná o připojení vedlejší věty obsahové (128 dokladů), ve 25 případech výraz *že* připojuje účinkovou či příčinnou větu, v 7 případech se jednalo o spojení *možná že* a *škoda že*. Zbylé případy v korpusu ORAL (19 %) lze považovat za diskurzní markery. Velmi často se nacházejí na konci repliky (příp. na konci významového či intonačního celku), kde vyjadřují vybídnutí k souhlasu a fungují jako tázací dovětky, podobně jako např. *viď*, *rozumíš*, *myslíš*[30] (příklad 15). Většina těchto výrazů tvoří ustálené spojení s výrazem *jo*, pomocí kterého se mluvčí obrací ke komunikačnímu partnerovi (příklad 16), případně může toto spojení propojovat jednotlivé části replik (příklad 17).

(15) *A: to bylo ono* ***že*** *? .. (kašel)*

*B: to bylo v Podjalubí nebo jak . (...)* (ORAL)

(16) *A: já sem ho viděla to máš v pokojíčku* ***že jo***

*B: no mně ho dala dvakrát já sem ho jednou ztratila (...)* (ORAL)

(17) *no @ tam to bylo* ***že jo*** *tak já sem tam jako šla* ***že jo*** *. tak jako jak dycky mám jako ten čas tak sem si tam vzala jako knížku (...)* (ORAL)

**Výsledky analýzy lemmat *a*, *ale*, *jako, že***

Cílem analýzy lemmat *a*, *ale*, *jako* a *že* bylo zjistit, v jakých významech se tato lemmata užívají v psané a mluvené komunikaci, zda se užití těchto lemmat v psané a mluvené komunikaci liší a zda je nějaký význam typický pouze pro mluvenou, příp. psanou komunikaci. Na základě výše představené analýzy je možné stanovit, že zkoumané výrazy se v korpusu SYN2020 užívají převážně ve funkci spojky (souřadicí či podřadicí). U lemmat *a*, *jako* a *že* bylo nalezeno málo dokladů, ve kterých je možné tyto výrazy považovat za diskurzní marker. U lemmatu *ale* je za diskurzní marker možné pokládat 29 % analyzovaných výrazů. Většina výrazů v korpusu SYN2020, které je možné považovat za diskurzní markery, se objevuje v beletrii, autoři je často užívají s cílem napodobit mluvený jazyk.

[30] Těmto tázacím dovětkům se ve své práci věnuje např. Z. Komrsková, 2018.

Také analyzované výrazy *a* a *že* v korpusu ORAL mají častěji funkci spojky. Spojka *a* se v korpusu ORAL užívá především pro vyjádření slučovacího poměru, další poměry vyjadřuje zřídka (na rozdíl od užívání spojky *a* v korpusu SYN2020). Spojka *že* připojuje v korpusu ORAL nejčastěji vedlejší věty obsahové. Přibližně pětina analyzovaných dokladů výrazů *a* a *že* je užita ve funkci diskurzních markerů.

Funkci spojky v korpusu ORAL má 21 % zkoumaných výrazů *ale* a 7 % zkoumaných výrazů *jako*. Zbylé doklady těchto výrazů je možné pokládat za diskurzní marker. Mluvčí je užívají z různých důvodů, např. pro propojení jednotlivých replik, jako výplňku, pro udržení plynulosti rozhovoru. Tyto výrazy se velmi často nacházejí na začátku repliky (příp. na začátku významového celku), mluvčí pomocí nich upozorňují, že přebírají slovo, případně navazují na něco, co bylo již řečeno.

#### 6.2.3.1.2 Lemma *ten*

SSČ a SSJČ charakterizují výraz *ten* jako zájmeno, pomocí kterého se odkazuje k osobě nebo věci, o které byla nebo bude řeč. Ustrnulý tvar *to* řadí mezi částice s významem zdůraznění (*to prší!*) či navázání (*to máš pravdu*).

Lemma *ten*[31] se vyskytuje ve frekvenční špičce obou zkoumaných korpusů, přesto je v korpusu ORAL mnohem frekventovanější (viz relativní frekvence lemmatu *ten* v tabulce č. 6, oddíl 6.2.3). Jedná se o nejfrekventovanější lemma v korpusu ORAL. Nejfrekventovanějším tvarem je v obou korpusech tvar *to*. V korpusu SYN2020 tvoří 55 % celého lemmatu, v korpusu ORAL 61 % celého lemmatu.

Řada autorů poukazuje na rozmanité funkce výrazu *to* v mluvené komunikaci. J. Hoffmannová a J. Zeman (2017) uvádí, že výraz *to* může mít jednak gramatickou funkci (uplatňuje se při mimotextové deixi a vnitrotextové koreferenci), ale také pragmatickou (má expresivní platnost, vyjadřuje emoce a často funguje jako výplňkové slovo). V případě pragmatické funkce autoři uvažují o přechodu demonstrativa *to* k částicím.

Podrobnější analýzu výrazu *to* ve své práci provedli autoři P. Klimešová, Z. Komrsková, M. Kopřivová a D. Lukeš (2015). V rámci vlastní klasifikace tohoto výrazu vyčlenili kategorii odkazování, kategorii kombinace (kam spadají výrazy *to*, které tvoří ustálenou jednotku s dalším komponentem, např. *to jo*) a kategorii mluvenosti. Kategorie mluvenosti sdružuje užití výrazu *to* v jevech, které jsou typické pouze pro mluvený projev a neobjevují se v psaném

[31] Pod lemma *ten* spadají i rodové varianty *ta* a *to* (a jejich příslušné tvary).

jazyce. Jedná se především o falešné starty, nedokončené výpovědi a o užití výrazu z důvodu neschopnosti mluvčího vybavit si konkrétní slovo.

Na základě výše zmíněných poznatků jsem na náhodném vzorku o velikosti 300 konkordančních řádků z obou korpusů provedla vlastní analýzu lemmatu *ten* (nikoliv pouze tvaru *to*). Cílem této analýzy je zjistit, zda se užívání lemmatu *ten* liší v mluvené a psané komunikaci. Při vlastní analýze jsem částečně vycházela z kategorizace použité ve výzkumu P. Klimešové a kol. (2015).

V obou korpusech má lemma *ten* nejčastěji gramatickou funkci (uplatňuje se při mimotextové deixi a vnitrotextové koreferenci). V korpusu SYN2020 má tuto funkci 82 % analyzovaných výrazů, v korpusu ORAL 54 %.

Druhou nejčastější kategorii (v obou korpusech) tvoří případy, kdy je lemma *ten* užito v přívlastkové pozici. Ve výzkumu P. Klimešové a kol. (2015) není tato kategorie zmíněna. Důvodem může být to, že do této kategorie spadá velmi málo výrazů *to*, pravděpodobně se autoři při analýze dat s případy spadajícími do této kategorie nesetkali. Případy, kdy je lemma *ten* užito v přívlastkové pozici, tvoří samostatnou kategorii. Tyto výrazy nevyjadřují gramatickou funkci, zároveň se však nejedná ani o kategorii mluvenosti, což potvrzuje to, že do této kategorie spadá 15 % zkoumaných výrazů z korpusu SYN2020. Pomocí těchto výrazů může autor či mluvčí připomínat něco, co už bylo zmíněno. V korpusu ORAL spadá do této kategorie čtvrtina analyzovaných výrazů. Příkladem výrazu spadajícího do této kategorie je výraz *ty* v příkladu 18. Na rozdíl od výrazů *to* a *tim*, které mají v tomto příkladu gramatickou funkci, může být výraz *ty* z promluvy vypuštěn. V souvislosti s užíváním zájmena *ten* v přívlastkové pozici upozorňují někteří autoři (např. Schneiderová, 1993) na to, že zájmeno *ten* v těchto případech nabývá povahy určitého členu.

(18) *no a voni byly **ty** kroupy že jo tak že by **to** bylo **tim** ?* (ORAL)

Do kategorie mluvenosti, kterou ve svém výzkumu popsali P. Klimešová a kol. (2015), spadá 12 % analyzovaných výrazů v korpusu ORAL. V příkladu 19 spadá první výraz *to* do kategorie mluvenosti, druhý výraz má gramatickou funkci (odkazuje k něčemu, co bylo řečeno dříve). V analyzovaném vzorku v korpusu SYN2020 do kategorie mluvenosti spadá pouze 5 dokladů (tedy necelá 2 %).

(19) *chcu jet **to** .. až **to** roztaje prostě (...)* (ORAL)

Překvapivě málo častá je kategorie kombinace (vymezená ve zmíněném výzkumu P. Klimešové a kol., 2015), do které v korpusu ORAL spadá pouze 5 % analyzovaných dokladů, v korpusu SYN2020 pouhé 1 %. Překvapivé je, že ve výzkumu P. Klimešové a kol. (2015) do této kategorie spadalo více analyzovaných výrazů než do kategorie mluvenosti. Tento rozdíl může být způsoben tím, že ve zmíněném výzkumu autoři analyzovali pouze tvar *to*, většinu ustálených spojení tvoří právě tento tvar (např. *to jo*, *to ne*, *to teda jo*).

Zbylé výrazy (12 v korpusu ORAL a 1 v korpusu SYN2020) nebylo možné analyzovat kvůli chybné lemmatizaci (nejednalo se o lemma *ten*).

Na základě této analýzy vyplývá, že většina zkoumaných výrazů (82 %) si v korpusu SYN2020 udržuje gramatickou funkci, ve zbylých případech dochází k oslabení gramatické funkce. V korpusu ORAL má gramatickou funkci přibližně polovina zkoumaných výrazů. Čtvrtina zkoumaných výrazů v korpusu ORAL je užita v přívlastkové pozici a má vycpávkovou či připomínací funkci. Do kategorie mluvenosti spadá 12 % zkoumaných výrazů a 5 % zkoumaných výrazů tvoří ustálená spojení.

#### 6.2.3.1.3 Lemmata *já*, *on*

SSČ a SSJČ popisují výraz *já* jako osobní zájmeno, které označuje mluvčího dané promluvy, případně může vyjadřovat zájem mluvčího na ději (tzv. dativ sdílnosti, např. *ty jsi mi pěkný hrdina*). Výraz *on* je v těchto slovnících rovněž popsán jako osobní zájmeno, které označuje osobu (mimo mluvčího i adresáta) nebo věc, o které už byla řeč nebo která se vyrozumívá ze souvislosti. SSJČ navíc uvádí částici *ono*, která vyjadřuje náladové zabarvení věty (např. *však ono to půjde*).

Lemmata *já* a *on*[32] se vyskytují ve frekvenční špičce obou zkoumaných korpusů, přesto jsou v korpusu ORAL mnohem frekventovanější (viz jejich relativní frekvence v tabulce č. 6, oddíl 6.2.3). Vysoká frekvence lemmatu *já* v korpusu ORAL je pravděpodobně způsobena dialogickou povahou tohoto korpusu. Podle autorek A. Čermákové a M. Kopřivové (2018) mluvčí často užívají zájmeno *já*, aby si udrželi slovo, zdůraznili vlastní mínění či pomohli komunikačnímu partnerovi při orientaci v reprodukci řeči. Podobnými důvody může být zapříčiněna vysoká frekvence výrazu *on* v korpusu ORAL. Mluvčí v rozhovoru potřebuje dát najevo, které názory, argumenty či myšlenky jsou jeho vlastní a které náleží někomu jinému.

[32] Pod lemma *on* spadají i rodové varianty *ona* a *ono* (a jejich příslušné tvary).

Na základě výše zmíněných poznatků o lemmatech *já* a *on* není překvapivé, že nejčastějšími tvary těchto lemmat v korpusu ORAL jsou tvary nominativu (nominativní tvar *já* tvoří 68 % celého lemmatu, nominativ lemmatu *on* tvoří 52 % celého lemmatu). J. Hoffmannová a J. Zeman (2017, s. 56) upozorňují, že častější uplatňování některých osobních zájmen v pozici podmětu v nepřipravených mluvených projevech umožňuje redukci složených tvarů préterita, např. *já to věděl*. Nominativní tvary lemmat *já* a *on* mluvčí užívají také při reprodukci cizí či vlastní řeči. V takových případech často dochází k vypouštění celého predikátu (srov. Hoffmannová – Komrsková – Poukarová, 2017).

V korpusu SYN2020 jsou tyto nominativní tvary méně frekventované (nominativní tvar *já* tvoří 23 % celého lemmatu, nominativ lemmatu *on* tvoří pouze 7 % celého lemmatu). Nižší frekvence těchto tvarů v korpusu SYN2020 může být způsobena tím, že na rozdíl od mluvené komunikace není potřeba explicitně zdůraznit, která mínění jsou vlastní a která cizí. Naopak podmět často nemusí být vyjádřen vůbec, protože obvykle vyplývá z kontextu.

V korpusu SYN2020 je nejčastějším tvarem lemmatu *já* dativní tvar *mi* (tvoří 34 % celého lemmatu v korpusu SYN2020). Zajímavé je, že pro vyjádření dativu se v korpusu SYN2020 příliš nepoužívá delší dativní tvar *mně*. Naopak v korpusu ORAL se oba dativní tvary užívají přibližně stejně často (ze všech dativních tvarů lemmatu *já* tvoří tvar *mi* 57 % a tvar *mně* 43 %). Nabízí se několik vysvětlení tohoto jevu. Zaprvé je důležité zmínit, že při transkripci tvarů *mě* a *mně* není v korpusu ORAL vždy dodržena pravopisná norma, z toho důvodu může být při automatické anotaci chybně určen pád těchto výrazů. Některé dativní tvary tedy ve skutečnosti mohou být tvary genitivu či akuzativu (a naopak), anotace pádů je spíše orientační. Důvodem, proč se v psané komunikaci pro vyjádření dativu příliš nepoužívá tvar *mně*, může být to, že tvar *mně* v příklonné pozici byl původně považován za chybný, k jeho pravopisné kodifikaci došlo až později (srov. Váhala, 1960). Vysokou frekvenci dativních tvarů *mně* v mluvené komunikaci způsobují některé dialekty. Podle *Českého jazykového atlasu 4* (Balhar a kol., 2002, s. 340–341) jsou dativní tvary *mně* v nepřízvučné pozici typické pro severovýchodočeskou nářeční oblast, středomoravskou a východomoravskou nářeční oblast.

Zajímavý je výraz *ono* (resp. *vono*), který má v korpusu ORAL mnohem větší frekvenci než v korpusu SYN2020. Podle J. Hoffmannové a J. Zemana (2017) dochází při spontánním projevu k tvarovému ustrnutí tohoto výrazu a lze jej považovat za částici, příp. za pomocný a výplňkový výraz (jako je tomu v příkladu 20 z korpusu ORAL).

(20) *je pěkná no . **vono** vůbec vopravdu tendleten film byl dobrej no* (ORAL)

#### 6.2.3.1.4 Lemmata *být*, *mít*

Lemmata *být* a *mít* spojuje to, že mají plný lexikální význam, ale na rozdíl od většiny ostatních sloves fungují také jako slovesa pomocná. Tato pomocná funkce je jeden z důvodů, proč jsou lemmata *být* a *mít* velmi frekventovaná jak v mluvené, tak i v psané komunikaci. Ve vlastní analýze jsem se zaměřila především na to, jak často mají tato slovesa plný lexikální význam, jaký je poměr mezi užitím sloves s lexikálním významem a užitím s pomocnou funkcí a zda se tento poměr liší v mluvené a psané komunikaci (resp. v korpusech ORAL a SYN2020). Nejprve představím analýzu lemmatu *být*, poté analýzu lemmatu *mít*.

Na náhodném vzorku o velikosti 300 konkordančních řádků z obou korpusů jsem provedla analýzu lemmatu *být*. První kategorii tvoří výrazy s plným lexikálním významem (existenciálním či zástupným). V korpusu SYN2020 do této kategorie spadá 12 % zkoumaných výrazů, v korpusu ORAL 21 % zkoumaných výrazů. V korpusu ORAL jsou časté především výrazy se zástupným významem. Mluvčí často užijí výraz *být*, místo aby použili specifické sloveso pro konkrétní děj či činnost. Takový případ je uveden v příkladu 21, kde by výraz *je* mohl být nahrazen jiným plnovýznamovým slovesem (např. *nachází se*, *žije*, *pracuje*, *bydlí* apod.). Naopak v psané komunikaci, kde se autor snaží vyjadřovat explicitně a popsat danou situaci co nejpřesněji, není užití výrazu *být* se zástupným významem příliš časté.

(21) *teď už **je** v Americe* (ORAL)

Druhou kategorii tvoří výrazy, které mají funkci gramatického auxiliáru[33]. Jedná se o préteritový auxiliár (např. *jsem*, *jsi*), kondicionálový auxiliár (např. *bych*, *by*) a pomocné futurální tvary (např. *budu*, *budou*). Z analyzovaných výrazů v korpusu SYN2020 do této kategorie spadá 31 %, v korpusu ORAL 41 % (tyto výrazy tvoří především préteritové auxiliáry).

Třetí kategorii tvoří výrazy, které mají funkci lexikálního auxiliáru, tedy spony (např. *ona je doktorka*, *on je malý*) a opisného pasiva (např. *cena byla udělena*). Z analyzovaných výrazů v korpusu SYN2020 do této kategorie spadá 56 %, v korpusu ORAL 31 %. Zajímavé je, že ve zkoumaných výrazech v korpusu SYN2020 se nachází 27 výskytů neadjektivizovaného tvaru opisného pasiva, v korpusu ORAL pouze 1.

[33] Informace o auxiliárních tvarech byly čerpány z Karlík, P. (2017): Auxiliár. In Karlík P. – Nekula M. – Pleskalová, J. (eds.): *CzechEncy – Nový encyklopedický slovník češtiny*. Dostupné z https://www.czechency.org/slovnik/AUXILIÁR [Poslední přístup 20. 11. 2022].

Zbylé výrazy (20 v korpusu ORAL a 1 v korpusu SYN2020) nebylo možné analyzovat (kvůli chybné lemmatizaci, případně se jednalo o nedořečenou výpověď).

Podobně byla na náhodném vzorku o velikosti 300 konkordančních řádků v obou korpusech provedena analýza lemmatu *mít*. První skupinu tvoří výrazy s plným lexikálním významem, tedy s významem vlastnění, přináležitosti, posesivity apod. Do této skupiny spadá 40 % analyzovaných výrazů v korpusu SYN2020 a 58 % v korpusu ORAL.

Druhou skupinu tvoří frazémy a ustálená spojení (např. *mít pocit*, *mít smysl*, *mít cenu*, *mít v úmyslu*, *mít kliku*, *mít šanci*, *mít na svědomí* apod.). Tato spojení tvoří samostatnou skupinu, nejedná se ani o pomocnou funkci, ale ani o lexikální význam vlastnění či přináležitosti. Ze zkoumaných výrazů v korpusu SYN2020 do této skupiny spadá 35 %, v korpusu ORAL pouze 12 %.

Třetí skupinu tvoří výrazy *mít*, které mají funkci pomocného slovesa. Do této skupiny jsou zařazeny jak modální významy (např. *mám říct*, *mám udělat*), tak významy rezultativní (např. *mají zavřeno*, *kabát mám pověšený*). Spadá sem čtvrtina zkoumaných výrazů v korpusu SYN2020, v korpusu ORAL 19 %. V obou korpusech je častější modální konstrukce než konstrukce rezultativní.

Výrazy spadající do poslední (čtvrté) skupiny mají kontaktovou funkci. Tuto skupinu tvoří pouze 8 výrazů z korpusu ORAL, ve zkoumaných výrazech v korpusu SYN2020 nebyl nalezen žádný výraz, který by spadal do této skupiny. Jedná se tedy o funkci, která je charakteristická především pro mluvenou komunikaci. Mluvčí výraz *mít* s touto funkcí užívají za účelem upoutání dalšího účastníka komunikace a k udržení kontaktu s ním. Pro tuto funkci jsou příznačné tvary druhé osoby, jako je tomu v příkladu 21.

(21) *jo a dyž přídeme se psama ven tak už je . to* ***máš*** *třeba já nevim . půl osmé* (ORAL)

V případě zbylých 24 výrazů v korpusu ORAL se jedná o nedořečené výpovědi. Z toho důvodu nebylo možné tyto výrazy zařadit do zmíněných skupin.

Cílem této analýzy bylo zjistit, jak často mají slovesa *být* a *mít* plný lexikální význam, jak často fungují jako pomocná slovesa a zda se tyto funkce liší v psané a mluvené komunikaci. V obou korpusech je sloveso *být* užíváno převážně ve funkci pomocného slovesa (v korpusu SYN2020 87 % zkoumaných výrazů, v korpusu ORAL 72 %). Výrazy s plným lexikálním významem (především se zástupným významem) jsou častější v korpusu ORAL (21 %). Naopak u slovesa *mít* převažují v obou korpusech užití s lexikálním významem vlastnění či přináležitosti (40 %

zkoumaných výrazů v korpusu SYN2020, 58 % v korpusu ORAL). Funkci pomocného slovesa má 25 % analyzovaných tvarů slovesa *mít* v korpusu SYN2020 a 19 % v korpusu ORAL. Na základě analýzy vyplývá, že slovesa *být* a *mít* s lexikálním významem se v korpusu ORAL vyskytují častěji než v korpusu SYN2020. Naopak slovesa *být* a *mít* s pomocnou funkcí se vyskytují častěji v korpusu SYN2020 než v korpusu ORAL.

#### 6.2.3.1.5 Lemmata *na*, *v*

Lemmata *na* a *v* jsou v obou korpusech užívána převážně ve funkci předložek. Výkladové slovníky (SSČ a SSJČ) uvádějí u výrazu *na* kromě významu předložky také význam citoslovce, které vyjadřuje vybídnutí k přijetí něčeho nebo někoho. Slovnědruhová automatická anotace výrazu *na* není ani v jednom korpusu spolehlivá. Část těchto výrazů je sice označena za citoslovce, ve většině případů se ale jedná o chybnou anotaci. V korpusu SYN2020 jsou za citoslovce označeny případy, kdy věta končí výrazem *na*. Ve skutečnosti se ale jedná o elipsu (např. *podívejte se na.*). V korpusu ORAL jsou za citoslovce označeny čtyři výrazy, ve všech případech se však jedná o chybnou anotaci. Na náhodném vzorku o velikosti 200 konkordančních řádků z obou korpusů jsem provedla analýzu výrazu *na*. V obou korpusech se ve všech případech jednalo o předložku. Výraz *na* se ve funkci citoslovce užívá jen okrajově.[34]

Lemmata *v* a *na* se nacházejí ve frekvenční špičce obou zkoumaných korpusů, přesto jsou v korpusu SYN2020 mnohem frekventovanější. Ukázalo se, že poměrně velký počet výskytů lemmat *v* a *na* v korpusu SYN2020 je součástí víceslovných předložek[35], naopak v korpusu ORAL tvoří víceslovné předložky jen zřídka. To dokládá i tabulka č. 8, která zobrazuje frekvenci vybraných víceslovných předložek (jejichž součástí jsou lemmata *v* a *na*) v korpusech SYN2020 a ORAL. První sloupec tabulky zobrazuje víceslovnou předložku. Následující sloupce představují relativní frekvenci (RF, uvedeno v i.p.m.) těchto víceslovných předložek v korpusech SYN2020 a ORAL.

Na základě tabulky č. 8 je zřejmé, že ze zkoumaných víceslovných předložek se v korpusu ORAL používá pouze *v rámci*, zbylé víceslovné předložky nejsou příliš frekventované. Zajímavé je, že v korpusu ORAL není ani jeden doklad užití předložky *na rozdíl od*, v korpusu SYN2020 je ale velmi frekventovaná.

[34] V korpusu SYN2020 je možné výraz *na* ve funkci citoslovce najít například pomocí dotazu [word="Na"][lemma=","], počet výskytů je ale nízký.
[35] Víceslovné výrazy nejsou v použitých korpusech lemmatizovány samostatně, ale jsou rozděleny na jednotlivé komponenty, které jsou pak přiřazeny ke konkrétním lemmatům (podrobněji viz oddíl 6.1.2).

| Frekvence vybraných víceslovných předložek v korpusech SYN2020 a ORAL | | |
|---|---|---|
| víceslovná předložka | RF (SYN2020) | RF (ORAL) |
| *bez ohledu na* | 35 | 2 |
| *na rozdíl od* | 70 | 0 |
| *na základě* | 104 | 13 |
| *nehledě na* | 7 | 5 |
| *v důsledku* | 39 | < 1 |
| *v rámci* | 168 | 33 |
| *v souvislosti s* | 41 | 2 |
| *ve srovnání s* | 31 | 1 |
| *ve vztahu k* | 18 | < 1 |

*Tabulka 8: Frekvence vybraných víceslovných předložek v korpusech SYN2020 a ORAL*

Lemmata *na* a *v* udělují dva pády: akuzativ a lokál. Prepozice *na* se v korpusech SYN2020 a ORAL pojí o něco častěji s akuzativem (přibližně dvě třetiny výskytů). Naopak předložka *v* se v obou korpusech pojí častěji s lokálem. V korpusu ORAL se s lokálem pojí přibližně 86 % výskytů (z celého lemmatu), v korpusu SYN2020 přibližně 95 % výskytů (z celého lemmatu). Výrazná převaha lokálu v korpusu SYN2020 je pravděpodobně způsobena víceslovnými předložkami. Například všechny výrazy *v*, které zobrazuje tabulka č. 8, jsou v korpusu SYN2020 anotovány jako výrazy vyjadřující lokál.

Kromě zmíněných víceslovných předložek tvoří lemmata *v* a *na* v korpusu SYN2020 následující kolokace, které v korpusu ORAL nejsou příliš časté: *v případě*, *ve skutečnosti*, *v řadě* (*případů*, *oblastí*, *ohledů*), *v době*, *na jedné straně*, *na druhé straně*, *na první pohled*, *na okamžik*, *na úkor*. Naopak některá spojení s těmito předložkami jsou častější v korpusu ORAL: *v pohodě*, *v klidu*, *v prdeli*, *na Vánoce*, *na chatě*, *na zahradě*, *na internetu*, *na hovno*, *jít na záchod*, *jít na pivo*, *jít na kafe*.[36]

Na rozdíl od psané komunikace mohou mít lemmata *v* a *na* v mluvené komunikaci pragmatickou funkci. Stejně jako mnoho dalších výrazů může mluvčí tato lemmata použít v následujících situacích: pozdržení výpovědi; rozmýšlení, jak bude promluva pokračovat; udržení slova; hledání vhodného výrazu apod. V takových případech dochází ke kumulaci stejných výrazů, jako je tomu v příkladech 22 a 23.

(22) *tos neslyšela ? jo to nám vlastně říkali **na na na** tý **na** táborový škole . no prostě rybník oddílovej* (ORAL)

(23) *ne v Metru ne . asi . **v v v** . asi **v** tom .. asi **v** Blesku ? (...)* (ORAL)

[36] Řada těchto spojení je dána tématem rozhovoru (mluvčí se často baví např. o svých plánech).

#### 6.2.3.1.6 Lemma *se*

Vysoká frekvence lemmatu *se* v obou korpusech je způsobena tím, že výraz *se* je často součástí slovesného tvaru (např. *pohybovat se*, *zeptat se*, *líbit se*). Výraz *se* je však lemmatizován zvlášť. Pokud by víceslovné jednotky tvořily v těchto korpusech víceslovná lemmata, výraz *se* by frekvenční špičky těchto korpusů netvořil.

Na základě výše zmíněného poznatku není překvapivé, že nejčastějšími tvary tohoto lemmatu jsou v obou korpusech tvary *se* a *si*. V korpusu SYN2020 tvoří 95 % celého lemmatu, v korpusu ORAL 98 %. Zajímavé je, že tvar *si* je v korpusu ORAL přibližně dvakrát častější než v korpusu SYN2020. Může to být způsobeno vysokou frekvencí některých sloves v korpusu ORAL, která se s výrazem *si* často spojují, např. *myslet si*, *říkat si*, *vzít si*, *dát si*. Dalším důvodem vysoké frekvence tvaru *si* může být to, že tento tvar je často chybně lemmatizován. Ve skutečnosti se nejedná o lemma *se*, ale o lemma *být* (tvar *jsi*).

### 6.2.3.2 Lemmata nacházející se pouze ve frekvenční špičce korpusu SYN2020

V této části představím osm následujících lemmat, která tvoří pouze frekvenční špičku korpusu SYN2020, ale netvoří frekvenční špičku korpusu ORAL: *který*, *moci*, *do*, *k*, *o*, *s*, *z, i*. Některá z těchto lemmat se v jazyce chovají podobně, mají podobný význam či funkci. Tato lemmata budou (stejně jako u předchozí skupiny) popsána společně.

#### 6.2.3.2.1 Lemma *který*

Ve výkladových slovnících SSČ a SSJČ je výraz *který* popsán jako zájmeno tázací, které vyjadřuje otázku po výběru jedné ze souhrnu osob či věcí, dále jako zájmeno vztažné, které uvozuje vedlejší větu vztažnou, a jako zájmeno neurčité, které je synonymní s výrazy *některý* a *kterýkoliv*.

Lemma *který* je v korpusu SYN2020 přibližně pětkrát častější než v korpusu ORAL (kde se jedná o 80. nejfrekventovanější lemma). Nižší frekvence v korpusu ORAL je pravděpodobně způsobena tím, že mluvčí pro uvození vedlejších vět vztažných (nesoucích určitou zpřesňující informaci) často užívají také výrazy *co* a *jak*[37] (příklady 24 a 25 z korpusu ORAL). Ve větách (resp. výpovědích) tyto výrazy nahrazují výraz *který*. Toto je zároveň jeden z důvodů, proč jsou lemmata *co* a *jak* frekventovanější v korpusu ORAL.

(24) *všechno všechny ty potraviny* ***co*** *se tam kupujou tak sou v takovejch balíkách* (ORAL)

[37] Tomuto jevu se ve svých pracích věnují např. Fried (2011) či Sláma – Šimík (2021).

(25) *jo a s tou Helenou . Hela ňák ráno sme se potkaly já Helena a ta tadyta holka **jak** bydlí naproti ta z Budějic a to* (ORAL)

Nejčastější funkcí lemmatu *který* je v korpusu SYN2020 pravděpodobně funkce vztažného zájmena (uvozujícího vedlejší větu). Toto je možné soudit na základě toho, že před většinou lemmat *který* stojí v korpusu SYN2020 čárka (příp. jiný výraz, před kterým stojí čárka). Funkce tázacího zájmena není v korpusu SYN2020 příliš častá. Tázací zájmena stojí obvykle na začátku věty. Výrazů *který* stojících na začátku věty je v korpusu SYN2020 poměrně málo (přibližně 0,2 % z celého lemmatu *který*).

#### 6.2.3.2.2 Lemma *moci*

Lemma *moci* funguje v jazyce primárně jako pomocné (modální) sloveso, může mít ale i plný lexikální význam. Přestože se lemma *moci* nevyskytuje ve frekvenční špičce korpusu ORAL, jedná se o 38. nejčastější lemma v tomto korpusu. Ve vlastní analýze jsem se zaměřila především na to, jak často má sloveso *moci* plný lexikální význam, jaký je poměr mezi užitím slovesa s lexikálním významem a užitím s pomocnou funkcí a zda se tento poměr liší v mluvené a psané komunikaci.

Na náhodném vzorku o velikosti 200 konkordančních řádků z obou korpusů jsem provedla analýzu lemmatu *moci*. První kategorii tvoří výrazy, které fungují jako pomocná (modální) slovesa. Do této kategorie spadá naprostá většina zkoumaných výrazů (všechny výrazy v korpusu SYN2020 a 99 % v korpusu ORAL). U zkoumaných výrazů v korpusu ORAL dochází k zajímavému jevu – k elipse sloves v infinitivním tvaru (ve zkoumaném vzorku jsem pro tento jev našla v korpusu ORAL 35 dokladů, v korpusu SYN2020 pouze 2 doklady). Velmi často dochází k elipse sloves pohybu, jako je tomu v příkladu 26, kde dochází k elipse slovesa *jít*. Může se však jednat i o vypuštění jiných plnovýznamových sloves, díky předchozímu kontextu je výpověď srozumitelná (jako je tomu v příkladu 27, kde dochází k vypuštění výrazu *napsat*).

(26) *jo neska si s tym pohraju . stejnak **nemužem** ven ..* (ORAL)

(27) *A: no a tak proč sis to nenapsal ručně ?*

*B: sto adres ? . samozřejmě že **můžu** . **můžu** jistě* (ORAL)

V analyzovaném vzorku jsem našla pouze dva výskyty slovesa *moci* s plným lexikálním významem, oba se nachází v korpusu ORAL. V obou případech se jednalo o význam *být odpovědný* (*moci za něco*). Mimo analyzovaný vzorek jsem v korpusu ORAL našla dva další

významy, které nejsou uvedeny v SSČ ani SSJČ. Oba významy je možné považovat za lexikální. Výraz *moci* může být užíván pro vyjádření obliby, případně ve významu *mám rád*, jako je tomu v příkladu 28[38]. Dále je možné tento výraz použít ve významu *mít z něčeho legraci*, *smát se něčemu*, případně *podivovat se nad něčím*. Tento význam dokládá příklad 29.

(28) *ale jako musim říct že víc než to semifredo mi třeba chutná tiramisu já nevim tiramisu vždycky já* ***můžu*** (ORAL)

(29) *to je jako masakr . kolikrát co tam ukazujou za věci ty jo uplně z toho* ***nemůžu*** *.. teďka tam postavili (nesrozumitelné) ve městě . (se smíchem) normálně . z překližky (se smíchem) bankomat (...)* (ORAL)

Z této analýzy vyplývá, že se sloveso *moci* používá primárně jako pomocné (modální) sloveso. V takových případech může především v mluvené komunikaci docházet k vypuštění slovesa v infinitivním tvaru (obvykle s významem pohybu), se kterým se sloveso *moci* pojí. Sloveso *moci* s lexikálním významem se používá pouze zřídka.

#### 6.2.3.2.3 Lemmata *do*, *k*, *o*, *s*, *z*

Tato lemmata spojuje kromě jejich vysoké frekvence to, že jsou, až na několik výjimek, v korpusu SYN2020 užívána ve funkci předložek. Přestože se tato lemmata nevyskytují ve frekvenční špičce korpusu ORAL, jsou v tomto korpusu také poměrně frekventovaná (nejnižší frekvenci má lemma *k*, jehož rank je 76). V psané i v mluvené komunikaci se tyto výrazy užívají převážně ve funkci předložek, mohou se však lišit určité konstrukce, ustálená spojení či kolokace, které tyto výrazy tvoří.

Jak již bylo řečeno, velkou část předložek v korpusu SYN2020 tvoří víceslovné předložky. Frekvenci vybraných víceslovných předložek v korpusech SYN2020 a ORAL ukazuje tabulka č. 9. Do tabulky byly zařazeny víceslovné jednotky, které jsou tvořeny předložkami *s*, *z* a *k*. Víceslovné předložky, jejichž součástí jsou výrazy *na* a *v*, jsou představeny v tabulce č. 8. Levý sloupec tabulky č. 9 zobrazuje víceslovnou předložku. Následující sloupce představují relativní frekvenci (RF, uvedeno v i.p.m.) těchto víceslovných předložek v korpusech SYN2020 a ORAL. Z tabulky je patrné, že se víceslovné předložky v korpusu ORAL příliš neobjevují. Naopak v korpusu SYN2020 jsou poměrně časté. Tato skutečnost je jedním z důvodů, proč se

[38] Tento jev se do jisté míry podobá konstrukcím typu *Karla Gotta nemusím*, kterým se ve své práci věnují Banášová – Hansen – Nekula (2011).

lemmata *k*, *s* a *z* nacházejí mezi 20 nejfrekventovanějšími lemmaty v korpusu SYN2020, ale nikoliv v korpusu ORAL.

| **Frekvence vybraných víceslovných předložek v korpusech SYN2020 a ORAL** | | |
|---|---|---|
| víceslovná předložka | RF (SYN2020) | RF (ORAL) |
| *čelem k* | 4 | < 1 |
| *s přihlédnutím k* | 3 | 0 |
| *s výjimkou* | 19 | < 1 |
| *směrem k* | 60 | 5 |
| *spolu s* | 103 | 3 |
| *vzhledem k* | 105 | 24 |
| *z důvodu* | 20 | 1 |
| *z hlediska* | 56 | 7 |

*Tabulka 9: Frekvence vybraných víceslovných předložek v korpusech SYN2020 a ORAL*

Kromě představených víceslovných předložek tvoří lemmata *do*, *k*, *o*, *s* a *z* v korpusu SYN2020 následující kolokace, které v korpusu ORAL nejsou příliš časté: *do doby*, *do* (*značné*, *určité*, *jaké*) *míry*, *do budoucna*, *k dispozici*, *k dostání*, *zájem o*, *informace o*, *s odstupem*, *s pravděpodobností*, *s jistotou*, *jeden z mála*, *z pohledu* (*uživatele*, *zákazníka*, *technického*…). V korpusu SYN2020 jsou dále frekventované některé konstrukce s těmito předložkami, v korpusu ORAL se tak často neobjevují. Například předložka *z* se často spojuje se superlativem adjektiv: *největší z*, *nejlepší z*, *nejmladší z*, *jeden z nejstarších*, *jeden z nejvýznamnějších* apod. Dále se předložka *z* často objevuje v konstrukcích vyjadřujících přídomek: *Jiří z Poděbrad*, *Karel ze Žerotína*, *Albrecht z Valdštejna*. Tato spojení se objevují především v publicistice a oborové literatuře.

Naopak některá spojení s těmito předložkami jsou častější v korpusu ORAL. Například předložka *s* se v korpusu ORAL oproti korpusu SYN2020 výrazně častěji spojuje s určitou osobou: *s dědou*, *s tátou*, *s Martinem*[39] apod. Předložka *do* tvoří v mluvené komunikaci často kolokace s vulgárním významem či s významem nadávky a klení: *do prdele*, *do prčic*, *do prkýnka* apod.

Na rozdíl od psané komunikace mohou mít tato lemmata v mluvené komunikaci pragmatickou funkci. Stejně jako mnoho dalších výrazů může mluvčí tato lemmata použít v následujících situacích: pozdržení výpovědi; rozmýšlení, jak bude výpověď pokračovat; udržení slova; hledání vhodného výrazu apod. V takových případech dochází ke kumulaci stejných výrazů, jako je tomu v příkladech 30 a 31. Zajímavé je, že ke kumulaci stejných výrazů dochází častěji,

[39] Vzhledem k anonymizaci osobních údajů v korpusu ORAL jsou velmi frekventovaná spojení *s NJ* a *s NP* (zkratka *NJ* symbolizuje osobní jméno, *NP* příjmení).

pokud obsahují vokál (např. *od*, *o*, *na*, *se*, *ke*, *ze* apod.). To je pravděpodobně způsobeno jejich snadnější výslovností. Z toho důvodu také mluvčí vkládají mezi kumulované výrazy, které neobsahují vokál (např. *k*, *s*, *z*, *v*), různé hezitační zvuky (příklad 32, hezitace je značena symbolem @).

(30) *chodili **do do do** toho **do** školy se m* se mnou* (ORAL)

(31) *to není o c* . cenzuře to je **o o o** tom **o** té **o** té **o** tom způsobu* (ORAL)

(32) *takže já musím jít **k @ k k** tejto . **@ k** obvodní (...)* (ORAL)

#### 6.2.3.2.4 Lemma *i*

Lemma *i* funguje v jazyce primárně jako výraz, který spojuje, zdůrazňuje či vytýká větné členy, případně je součástí víceslovných spojek. V korpusu SYN2020 je toto lemma více než dvakrát frekventovanější než v korpusu ORAL (kde se jedná o 70. nejfrekventovanější lemma).

Ukázalo se, že poměrně velký počet výskytů lemmatu *i* v korpusu SYN2020 je součástí víceslovných spojek, naopak v korpusu ORAL tvoří víceslovné spojky jen zřídka. To dokládá tabulka č. 10, která zobrazuje frekvenci vybraných víceslovných spojek, jejichž součástí je výraz *i*. První sloupec tabulky zobrazuje víceslovnou spojku. Následující sloupce představují relativní frekvenci (RF, uvedeno v i.p.m.) těchto víceslovných spojek v korpusech SYN2020 a ORAL.

| **Frekvence vybraných víceslovných spojek v korpusech SYN2020 a ORAL** | | |
|---|---|---|
| víceslovná spojka | RF (SYN2020) | RF (ORAL) |
| *ale i* | 233 | 37 |
| *dokonce i* | 75 | 10 |
| *i kdyby* | 39 | 32 |
| *i když* | 289 | 215 |

*Tabulka 10: Frekvence vybraných víceslovných spojek v korpusech SYN2020 a ORAL*

Lemma *i* často tvoří v korpusu SYN2020 různé opozitní (binomiální) konstrukce, kdy lemma *i* spojuje dva výrazy stejného slovního druhu a stejné formy. Mezi nejčastější kombinace (příp. kolokace) v korpusu SYN2020 patří *malý i velký*, *dítě i dospělý*, *muž i žena*, *den i noc*, *tělo i duše*, *český i zahraniční*, *vnitřní i vnější*, *dobrý i zlý* apod. V korpusu ORAL se téměř nevyskytují (nejfrekventovanější je spojení *česky i německy*, které se v korpusu ORAL vyskytuje pouze třikrát).[40]

[40] Pro vyhledání těchto spojení jsem použila následující dotaz: 1:[lemma=".*"][lemma="i"]2:[lemma=".*"] & 1.pos=2.pos & 1.case=2.case.

#### 6.2.3.3 Lemmata nacházející se pouze ve frekvenční špičce korpusu ORAL

V této části představím 8 následujících lemmat, která tvoří pouze frekvenční špičku korpusu ORAL, ale netvoří frekvenční špičku korpusu SYN2020: *hmm*, *jo*, *no*, *ne*, *co*, *vědět*, *tak*, *tam*. Některá z těchto lemmat se v jazyce chovají podobně, mají podobný význam či funkci. Tato lemmata budou (stejně jako u předchozích dvou skupin) popsána společně.

##### 6.2.3.3.1 Lemmata *hmm*, *jo*, *no*, *ne*

Tato lemmata mají velmi podobné vlastnosti a funkce. Z toho důvodu nejprve obecně popíšu jejich společné vlastnosti. Následně představím každé lemma zvlášť, soustředím se především na jeho specifické funkce a významy, případně na vlastnosti, kterými se liší od ostatních.

Tyto výrazy popisují autorky A. Čermáková, L. Jílková, Z. Komrsková, M. Kopřivová a P. Poukarová (2019, s. 259) jako responzní diskurzní markery, pomocí kterých dává posluchač najevo, že hovor sleduje, orientuje se v něm a ne/souhlasí s řečeným. Na základě analýzy náhodného vzorku 100 konkordančních řádků ve vybraných mluvených korpusech autorky popsaly, jaké mají zkoumané výrazy vlastnosti a jak se užívají v jazyce. Kromě výrazů *no*, *jo*, *hmm* a *ne*, kterými se zabývá tato diplomová práce, řadí autorky mezi responzní diskurzní markery výrazy *jasně*, *ano* a neverbální projev nesouhlasu označovaný jako *emm*. Tyto výrazy nejsou považovány za diskurzní markery, pokud jsou užívány k vyjádření souhlasu či nesouhlasu po položené otázce. Všechna ostatní užití jsou řazena mezi diskurzní markery.

Autorky zmíněné kapitoly o diskurzních markerech (Čermáková a kol., 2019) prozkoumaly výrazy *jo*, *no*, *hmm* a *ne* poměrně podrobně. Při popisu jednotlivých lemmat stručně představím některá zjištění autorek. Ve své analýze jsem došla k velmi podobným závěrům. Z toho důvodu se dále zaměřím především jevy, kterým se autorky ve své práci nevěnovaly (případně se jim věnovaly jen velmi málo).

Kromě vlastností popsaných v práci o diskurzních markerech (Čermáková a kol., 2019) jsem na základě analýzy dat z použitých korpusů zjistila ještě další podobnosti těchto lemmat. Jak již bylo řečeno, jedná se o lemmata, která jsou velmi frekventovaná v korpusu ORAL, ale v korpusu SYN2020 nejsou příliš častá. Jejich vysoká frekvence je pravděpodobně způsobena dialogickou povahou korpusu ORAL. Dalším důvodem, proč mají tyto výrazy vysokou frekvenci, může být to, že je mluvčí často pronesou několikrát za sebou. Tyto responzní diskurzní markery se často vzájemně kombinují, případně se k nim přidávají i další frekventované výrazy v mluveném jazyce, např. *no jo no*, *no tak to jo*, *no to ne*. Jejich funkce je často spojena s intonací, kterou byl výraz pronesen.

V korpusu SYN2020 mají tyto výrazy různé tvary, pomocí kterých se autor určitým způsobem snaží napodobit jejich zvukovou stránku, např. *no*, *nó*, *nooo*, *nóó*. Výrazy *jo*, *no* a *hmm* nejsou v korpusu SYN2020 příliš frekventované. Nejčastěji se vyskytují v beletrii (přibližně 85 % dokladů každého lemmatu), kde se snaží napodobit mluvenou řeč. Tato lemmata jsou užita v přímé řeči postav, případně se mohou vyskytovat ve vnitřním monologu některé z postav. V publicistice a oborové literatuře se výrazy *jo*, *no* a *hmm* objevují nejčastěji v doslovných přepisech rozhovorů (často se sportovní tematikou). Je tedy otázkou, zda se jedná o psanou komunikaci, nebo už spíše o určitý typ komunikace mluvené.

**Lemma *hmm***

Výkladové slovníky SSJČ a SSČ popisují výraz *hmm* (resp. *hm*) jako prostředek, který vyjadřuje zdrženlivé přitakání, pochyby, nesouhlas, pohrdání, rozpaky či podivení. V souvislosti s lemmatem *hmm* v korpusu ORAL je důležité upozornit, že se nejedná o slova, ale spíše o zvuky se sémantickou složkou. Tyto výrazy mohou být v některých případech zaměněny s hezitačními zvuky, které jsou v korpusu ORAL značeny symbolem @, hranice mezi hezitací a souhlasným zvukem nemusí být ale vždy jasná.

Autorky výše představené práce o diskurzních markerech (Čermáková a kol., 2019) uvádějí kromě významů zmíněných ve slovnících další funkce výrazu *hmm*. Často vyjadřuje zpětnou vazbu posluchače, posluchač dává najevo, že řečenému rozumí, vnímá a povzbuzuje mluvčího. V některých případech může tento výraz signalizovat změnu tématu (či změnu mluvčího).

Ve vlastní analýze tohoto výrazu jsem se zaměřila především na jeho pozici v promluvě. Zajímavé je, že ze všech nalezených dokladů *hmm* v korpusu ORAL se 86 % *hmm* vyskytuje v iniciační pozici. V případech, kdy se tento výraz nachází uprostřed promluvy (příklad 33 a 34), často neplní žádnou z výše zmíněných funkcí, jedná se tedy spíše o hezitační zvuk (a tedy špatnou transkripci).

(33) *(...) a říká no tak sem v páťáku tak už umim* ***hmm*** *můžu si zajít večer i do kina* (ORAL)

(34) *a jako král jako král můžeš taky znova lovit* ***hmm*** *zdanit ten loveckej zámeček* (ORAL)

**Lemma *jo***

V SSJČ je tento výraz považován za příslovce (s významem souhlasu a přisvědčení) a za citoslovce, které vyjadřuje lítost, překvapení či nelibost. Řada autorů ale v souvislosti s tímto výrazem zmiňuje jeho sémantickou vyprázdněnost (srov. Hoffmannová, 2017; Jílková, 2017).

Autorky výše představené práce o diskurzních markerech (Čermáková a kol., 2019) zmiňují, že výraz *jo* funguje v jazyce jako diskurzní marker, pokud dochází k oslabení původního významu (tedy významu souhlasu). Diskurzní marker *jo* slouží k zavedení nového tématu či k uchopení slova, na konci promluvy může sloužit jako výzva k reakci.

Výraz *jo* tvoří velmi často kolokaci s výrazem *že*. V korpusu ORAL má spojení *že jo* absolutní frekvenci 16582, což je přibližně 17 % celkové frekvence výrazu *jo* v tomto korpusu. Většina těchto spojení má funkci diskurzního markeru (kromě případů typu *myslím, že jo*; *řekl, že jo* apod.).

**Lemma *no***

Podle SSČ má výraz *no* navazovací, vybízecí a přitakávací funkci. V SSJČ je navíc zmíněn význam pobídnutí, povzbuzení, chlácholení, rozmrzení, odporu, překvapení či vyjádření nejistoty a osobních pocitů.

V souvislosti se zkoumáním funkcí výrazu *no* O. Müllerová (1996) zmiňuje, že se od sebe odlišují funkce v monologickém a dialogickém užití. V monologu výraz *no* usnadňuje mluvčímu navázat na to, co už řekl, nebo vyjadřuje vztah k tomu, co již pronesl. V dialogu pomocí něj mluvčí vyjadřuje své postoje k partnerovi nebo jeho výpovědím.

Stejně jako u výrazu *jo* autorky výše uvedené práce o diskurzních markerech (Čermáková a kol., 2019) zmiňují, že výraz *no* funguje v jazyce jako diskurzní marker, pokud dochází k oslabení významu souhlasu. Diskurzní marker *no* slouží k zavedení nového tématu či k uchopená slova. Na začátku promluvy má často přípravnou funkci. Mluvčí pomocí něj může signalizovat reprodukci cizí či vlastní řeči.

Kromě výše zmíněných funkcí, které popisují A. Čermáková a kol. (2019), jsem v korpusu ORAL našla ještě jeden význam, který autorky nezmiňují. Mluvčí může použít výraz *no*, pokud chce pokárat či napomenout dalšího účastníka komunikace (příklad 35). V takových případech je tento výraz často spojen s vokativem (příklad 36). Tato užití je rovněž možné řadit mezi diskurzní markery.

(35) *ale **no** tak nedělejme si z toho srandičky* (ORAL)

(36) *A: **no** Kubo to ale nemůžeš takhle*

*B: co ?*

*A: no tak chlastat* (ORAL)

**Lemma *ne***

Výkladové slovníky (SSČ a SSJČ) označují *ne* za částici, která vyjadřuje nesouhlas s obsahem předchozího výroku, popírá část věty nebo jednotlivá slova. SSJČ považuje tento výraz za citoslovce, pokud vyjadřuje překvapení či podiv. V SSČ je výraz s tímto významem považován za modální částici.

Stejně jako u předchozích výrazů autorky výše uvedené práce o diskurzních markerech (Čermáková a kol., 2019) zmiňují, že výraz *ne* není považován za diskurzní marker, pokud je užíván k vyjádření nesouhlasu po položené otázce. Autorky řadí ostatní užití k diskurzním markerům.

Na rozdíl od výrazů *no*, *jo* a *hmm* je výraz *ne* v korpusu SYN2020 poměrně frekventovaný. Odlišná je i jeho frekvenční distribuce v jednotlivých textových makrotypech korpusu SYN2020. Předchozí výrazy se vyskytují převážně v beletrii, výraz *ne* se poměrně často užívá také v oborové literatuře a publicistice. Na základě těchto odlišností jsem se rozhodla prozkoumat funkce výrazu *ne* v psané komunikaci.

Na náhodném vzorku o velikosti 100 konkordančních řádků z korpusu SYN2020 jsem provedla analýzu výrazu *ne*. V původním významu (vyjádření nesouhlasu po výzvě, popření věty či jednotlivých slov) je užito 77 % zkoumaných výrazů. Zbylá užití výrazu *ne* mají funkci diskurzního markeru, většina z nich je užita v beletrii. V takových případech výraz *ne* vyjadřuje nesouhlas, kterému nepředcházela výzva k reakci (8 % analyzovaných dokladů, příklad 37), vyjadřuje výzvu k reakci (8 % analyzovaných dokladů, příklad 38), případně je užíván jako konektivní prostředek, který navazuje na předchozí kontext (4 % analyzovaných dokladů, příklad 39). Zbylé doklady (3 %) není možné analyzovat vzhledem k omezenému kontextu či chybné anotaci.

(37) *V Caseových očích však po veselí nebyla ani památka. „Sarah," začal ztěžka. „**Ne**," přerušila ho okamžitě* (SYN2020)

(38) *„A kolik měl tvůj otec v osmdesátým? Kolem padesátky, **ne**?" „Přesně padesát."* (SYN2020)

(39) *Teď šlo o to, kolik toho věděla Ráchel. Stihla jí Isabela všechno říct? Určitě už všechno ví. **Ne**, ženy musí zemřít. Není jiné východisko.* (SYN2020)

Z této analýzy vyplývá, že výraz *ne* je v korpusu SYN2020 užíván převážně v původním významu. Naopak výsledky analýzy výrazu *ne* v korpusu ORTOFON, kterou provedly

Čermáková a kol. (2019), ukazují, že výraz *ne* s v mluvené komunikaci užívá převážně ve významu diskurzního markeru (61 %). Vysoká frekvence lemmatu *ne* v korpusu SYN2020 (oproti lemmatům *jo*, *no* a *hmm*) je způsobena tím, že výraz *ne* neslouží pouze k vyjádření nesouhlasu po položené otázce, ale slouží především jako částice, která popírá část věty nebo jednotlivá slova.

#### 6.2.3.3.2 Lemma *co*

ASSČ řadí výraz *co* k pěti různým slovním druhům. Výraz *co* může mít jednak funkci zájmena (které uvozuje tázací větu, obsahovou větu nebo vztažnou větu nesoucí určitou zpřesňující informaci, příp. označuje neurčitou věc či jev), jednak zájmenného příslovce (vyjadřuje otázku po příčině nějakého jednání a uvozuje vedlejší větu vztažnou, zejména příslovečnou). Dále výraz *co* řadí mezi spojky (uvozuje vedlejší větu vyjadřující časový úsek či způsob děje věty hlavní), částice (zdůrazňuje nejvyšší možnou míru něčeho) či citoslovce (vyjadřuje údiv a překvapení, vyjadřuje vybídnutí k souhlasu s předchozím tvrzením).

Přestože se lemma *co* nevyskytuje ve frekvenční špičce korpusu SYN2020, jedná se o 24. nejčastější lemma v tomto korpusu. Nejfrekventovanějším tvarem v obou korpusech je tvar *co*. V korpusu SYN2020 tvoří 87 % celého lemmatu, v korpusu ORAL dokonce 94 %. J. Hoffmannová a J. Zeman (2017, s. 57) ve své práci upozorňují, že se původní zájmeno *co* uplatňuje v mluvených projevech jako univerzální a vysoce frekventovaný nesklonný vztažný, tedy spojovací výraz.

Na základě výše zmíněných poznatků jsem na náhodném vzorku o velikosti 200 konkordančních řádků z obou korpusů provedla vlastní analýzu lemmatu *co* (nikoliv pouze tvaru *co*). Při vlastní analýze jsem vycházela především z významů představených v ASSČ. Cílem této analýzy je zjistit, v jakém významu se lemma *co* užívá v mluvené a psané komunikaci a zda se v psané a mluvené komunikaci užívají všechny zmíněné významy.[41]

V obou korpusech se výraz *co* užívá nejčastěji ve funkci zájmena. V korpusu SYN2020 tvoří 86 % zkoumaného vzorku, v korpusu ORAL 71 %. V korpusu ORAL se častěji objevuje užití zájmena *co*, které uvozuje vedlejší větu vztažnou nesoucí určitou zpřesňující informaci (mohlo by být nahrazeno výrazy *který* a *jak*)[42]. V analyzovaném vzorku v korpusu ORAL jsem našla 26 výskytů tohoto užití (příklad 40), v korpusu SYN2020 pouze 9 (příklad 41).

[41] Použité korpusy sice obsahují slovnědruhovou anotaci výrazu *co*, tato anotace ale není spolehlivá (např. v korpusu ORAL jsou všechny výrazy *co* určeny jako zájmena).
[42] Tomuto jevu se ve svých pracích věnují např. Fried (2011) či Sláma – Šimík (2021).

(40) *všechny byly předražený ty věci **co** tam měli* (ORAL)

(41) *Dobrý den, tenhle dům patřil našim prarodičům, rádi bychom se mrkli do dřevníku, jestli tam nejsou pušky, **co** nám děda vyráběl*. (SYN2020)

Pro výraz *co* ve funkci příslovce jsem ve zkoumaném vzorku v korpusu SYN2020 našla 14 dokladů, v korpusu ORAL pouze 3 (příklad 42, kde mluvčí vyjadřuje otázku po příčině jednání).

(42) *(na psa) ale **co** mě voblizkuješ já nic nemám* (ORAL)

V analyzovaném vzorku v korpusu SYN2020 jsem pro výraz *co* ve funkci spojky nenašla žádný doklad, v korpusu ORAL pouze jeden. Výraz *co* ve funkci spojky se tedy užívá jen velmi okrajově.

Užití výrazu *co* ve funkci částice bylo nalezeno pouze ve vzorku v korpusu SYN2020 (konkrétně 8 dokladů, jeden z nich uvádí příklad 43). Ve všech případech se jednalo o užití *co*, které zdůrazňuje nejvyšší možnou míru něčeho. Zkoumaný vzorek z korpusu ORAL neobsahoval ani jeden doklad tohoto užití. V korpusu ORAL je možné slovo *co* s tímto významem vyhledat pomocí dotazu [lemma="co"] [word="nej.*" &pos="A|D"], který zajistí vyhledání lemmatu *co*, po kterém následuje adjektivum či adverbium v superlativu (např. *co nejlepší*, *co nejrychleji*). V celém korpusu ORAL bylo nalezeno pouze 186 dokladů (tedy přibližně 0,5 % celého lemmatu *co*) užití výrazu *co* v tomto významu.

(43) *Myslím, že bychom měli zavolat na policii a **co** nejdříve to ohlásit.* (SYN2020)

Pro výraz *co* ve významu citoslovce jsem v analyzovaném vzorku v korpusu ORAL našla 54 dokladů, v korpusu SYN2020 pouze 7 (všechny v beletrii). Ve většině případů se jedná o spojení výrazu *co* s dalším slovem (příklad 44), např. *víš co*, *nebo co*, *co jako*, *tak co*, *no a co*, případně je výraz *co* užit jako tázací dovětek (příklad 45). Mluvčí často užívá tyto výrazy z pragmatických důvodů, mají funkci diskurzního markeru.

(44) *hele máš to v pohodě ségra taky . **tak co*** (ORAL)

(45) *jo a hulilo se tu . vesele **co** ?* (ORAL)

Cílem této analýzy bylo zjistit, v jakém významu se lemma *co* užívá v mluvené a psané komunikaci a zda se v psané a mluvené komunikaci užívají všechny zmíněné významy. Při vlastní analýze jsem vycházela především z významů představených v ASSČ. Z analýzy náhodného vzorku o velikosti 200 konkordančních řádků z korpusů SYN2020 a ORAL

vyplývá, že v obou korpusech je tento výraz nejčastěji užíván ve funkci zájmena. Funkce příslovce je častější v korpusu SYN2020 (7 % zkoumaného vzorku). V korpusu SYN2020 se výraz *co* užívá také ve funkci částice (4 %), v korpusu ORAL jsem ve zkoumaném vzorku pro toto užití nenašla žádný doklad. Naopak v korpusu ORAL se tento výraz poměrně často (27 % zkoumaného vzorku) užívá ve funkci citoslovce (příp. diskurzního markeru), v korpusu SYN2020 jsem pro toto užití našla ve zkoumaném vzorku pouze 7 dokladů. Výraz *co* ve funkci spojky se ani v jednom korpusu téměř nepoužívá.

#### 6.2.3.3.3 Lemma *vědět*

Slovníky SSČ a SSJČ uvádějí, že základním významem výrazu *vědět* je mít vědomost o někom či o něčem, být s něčím seznámen, znát něco či někoho. Přestože se toto lemma nachází pouze ve frekvenční špičce korpusu ORAL, v korpusu SYN2020 je také poměrně frekventované (jedná se o 57. nejčastější lemma).

V korpusu ORAL se tento výraz objevuje nejčastěji ve tvaru *nevím* a ve tvaru *víš*. Tyto tvary tvoří 84 % celého lemmatu, v korpusu SYN2020 pouze 19 % celého lemmatu. Z důvodu jejich vysoké frekvence v korpusu ORAL jsem se rozhodla udělat analýzu těchto tvarů. Cílem analýzy je stanovit, zda se tyto tvary užívají pouze v původním (výše představeném) významu, nebo jestli v mluveném jazyce nabývají nových významů. Na náhodném vzorku o velikosti 100 konkordančních řádků z korpusu ORAL jsem provedla analýzu tvarů *nevím* (resp. *nevim*) a *víš* (resp. *viš*). Nejprve představím analýzu tvaru *nevím*, poté analýzu tvaru *víš*.

První skupinu tvoří výrazy *nevím*, které vyjadřují původní význam. Za původní význam jsou považována užití výrazu *nevím*, ke kterému se pojí valenční doplnění v akuzativu (či vedlejší věta), příp. jde o spojení *vědět o* + *lokál* (např. *nevím nic*; *nevím, že*; *nevím, kdo*; *nevím o tom*). Za výrazy s původním významem jsou dále považovány ty výrazy, pomocí kterých mluvčí reaguje na otázku druhého mluvčího (příklad 46). Do této skupiny spadá většina analyzovaných výrazů (65 %).

(46) *A: a proč to maj tady takhle ?*

*B: no* ***nevim*** *ale já jí říkam tak si se měla podívat vidíš i tady to je uděláš linku* (ORAL)

Do druhé skupiny spadají zbylé výrazy (35 %), které mají funkci diskurzního markeru. Tyto výrazy nevyjadřují původní význam slovesa *vědět*, ale mluvčí je užívají z různých pragmatických důvodů. Častou funkcí výrazu *nevím* je konektivní funkce, mluvčí pomocí

tohoto výrazu propojuje jednotlivé části repliky. V příkladu 47 jej mluvčí za tímto účelem užije dvakrát.

(47) *tak no ale tak jako pak sem to vyměnila tak sem fotila ještě na Martin* Markétin digitál ale . jako zdá .* ***nevim*** *zdá se mi že nemá tak dobrej . digitál jako naše mamka třeba nebo* ***nevim*** *nefotilo se mi na něj moc dobře* (ORAL)

V analyzovaném vzorku bylo šest případů, kdy mluvčí užije výraz *nevím* k zakončení výpovědi. Mluvčí tento výraz užívají, pokud neví, jak výpověď zakončit (často pomocí spojení *nebo nevím*), případně pokud chtějí dát najevo nejistotu ohledně svého tvrzení (často pomocí spojení *ale nevím*). Toto užití je uvedeno v příkladu 48.

(48) *no jo no s tím dítětem jako na to se tak těšil že jo na to dítě a teď tady to no* ***nevim*** (ORAL)

Mluvčí výraz *nevím* užívají, pokud uvádějí nějaký příklad. Funguje podobně jako výrazy *například*, *dejme tomu*, *třeba*. Těmito výrazy může být výraz *nevím* často doprovázen, jako je tomu v příkladu 49.

(49) *i posedět to jako ani náhodou . spíš třeba já* ***nevim*** *s Markétou NP jak sem k ní chodila* (ORAL)

V analyzovaném vzorku jsem našla čtyři případy, kdy výraz *nevím* funguje jako responzní diskurzní marker (podobně jako výrazy *hmm*, *no* a *ne*). Těmito výrazy posluchač vyjadřuje zpětnou vazbu, dává najevo, že rozhovor sleduje, vnímá, případně že s mluvčím ne/souhlasí. Toto užití je uvedeno v příkladu 50, kde mluvčí pomocí *nevím* dává najevo, že hovor sleduje, a v příkladu 51, kde mluvčí pomocí *nevím* vyjadřuje nesouhlas s jiným mluvčím (resp. vyjadřuje pochybnosti ohledně tvrzení jiného mluvčího).

(50) *A: když si to dá jako jediný jako že se snaží tim doplnit jako kalorie místo jako jídla*

*B: já* ***nevim*** *já* ***nevim***

*A: když si dá ty voříšky tak to máš ale hrozně kilojaulů* (ORAL)

(51) *no . (se smíchem) já* ***nevim*** *no nechci vám brát iluze ale myslim si že to . neuděláte* (ORAL)

Podobně byla provedena analýza tvaru *víš*. Stejně jako u výrazu *nevím* byla za původní význam považována užití výrazu *víš*, ke kterému se pojí valenční doplnění v akuzativu (či vedlejší věta),

příp. jde o spojení *vědět o + lokál* (např. *už to víš*; *víš, že*; *víš, kdo*; *víš o tom*). Do této skupiny spadá 23 % analyzovaných výrazů.

Zbylé výrazy (77 %) jsou použity ve funkci diskurzního markeru. V těchto případech dochází k oslabení (či zániku) původního významu, mluvčí výraz *víš* užívají spíše z pragmatických důvodů. Třetinu analyzovaných výrazů tvoří víceslovné jednotky obsahující tvar *víš*, např. *to víš*, *víš co*. Nejedná se však o spojení plnovýznamového slovesa *vědět* s valenčním doplněním, ale celé spojení funguje jako diskurzní marker, jako je tomu v příkladech 52 a 53.

(52) *A: já nechápu jak tam můžeš žít nahoře já bych tam chcípnul*

*B: no* ***to víš*** *tak sem zvyklej no (...)* (ORAL)

(53) *jednú jednú sem byl enom ale tak* ***víš co*** *dávat každý den kilo za to* (ORAL)

Třetina výrazů *víš* v analyzovaném vzorku je užívána za účelem propojení jednotlivých replik. Výraz *víš* má tak konektivní funkci, jako je tomu v příkladu 54.

(54) *ale to už bylo jak sem mu to dělal ten počitač* ***viš*** *ja řikam . to maš ale jenom deset giga (...)* (ORAL)

Pomocí zbylých výrazů (11 % z analyzovaného vzorku) mluvčí vyjadřuje výzvu k reakci, obrací se na komunikačního partnera a očekává od něj určitou odezvu. Svou funkcí se tyto výrazy velmi podobají tázacím dovětkům *chápeš* a *rozumíš*, kterým se ve své práci věnuje Z. Komrsková (2018), či dalším výrazům, např. *viď*, *že jo*, *ne*, *že*. Tato užití jsou obvykle doprovozena stoupavou intonací a objevují se na konci repliky (příklad 55).

(55) *A: a . voni ho vostříhaly* ***víš***

*B: jo jasně to má dobře zařízený* (ORAL)

Cílem této analýzy bylo zjistit, zda se výrazy *nevím* (resp. *nevim*) a *víš* (resp. *viš*) užívají pouze v původním významu, nebo jestli v mluveném jazyce nabývají nových významů. Z analýzy náhodného vzorku o velikosti 100 konkordančních řádků z korpusu ORAL vyplývá, že mluvčí oba výrazy užívají nejen v původním významu, ale také z různých pragmatických důvodů. V analyzovaném vzorku má funkci diskurzního markeru 35 % zkoumaných výrazů *nevím* a 77 % zkoumaných výrazů *víš*. V mluvené komunikaci tak u těchto výrazů často dochází k oslabení (či zániku) původního významu.

#### 6.2.3.3.4 Lemma *tak*

V SSČ je výraz *tak* charakterizován jako zájmenné příslovce vyjadřující způsob a míru (*tak se tehdy žilo*; *ještě není tak starý*). Dále může mít funkci souřadicí spojky důsledkové (*bál se, (a) tak si pískal*), spojky slučovací (ve spojení s *jak*) a částice navazovací (*tak pojď*), částice zdůrazňovací (*kdyby tu tak byl*) a částice vyjadřující přibližnost (*může mu být tak pět let*). SSJČ navíc uvádí funkci citoslovce vyjadřujícího podiv, překvapení či souhlas.

Výrazem *tak* se ve své práci zabývá L. Jílková (2016a,b), která zkoumá užití výrazu *tak* v korpusu DIALOG. Autorka píše, že výraz *tak* má v mluvené komunikaci často přípravnou funkci, pomocí tohoto výrazu mluvčí signalizují, že jsou připraveni převzít slovo. V závěru práce dochází ke zjištění, že čím je mluvený text formálnější, tím méně se v něm slovo *tak* vyskytuje (a naopak) (Jílková, 2016b, s. 40). Kromě zmíněné práce se L. Jílková podrobněji věnuje spojení *až tak* (Jílková, 2018) a *takže tak* (Jílková, 2021).

Tomuto výrazu se ve své práci o diskurzních markerech věnovaly autorky A. Čermáková, L. Jílková, Z. Komrsková, M. Kopřivová a P. Poukarová (2019). Na náhodném vzorku 100 konkordančních řádků z korpusu ORTOFON provedly analýzu výrazu *tak*. Z této analýzy vyplynulo, že přibližně dvě třetiny zkoumaných dokladů se užívají ve funkci spojky či adverbia, zbylá třetina plní funkci diskurzního markeru, kde mluvčí pomocí výrazu *tak* přebírají slovo (Čermáková a kol., 2019, s. 330).

V rámci vlastní analýzy výrazu *tak* jsem zkoumala, jak často tento výraz v korpusu SYN2020 a ORAL funguje jako adverbium, spojka a kdy má funkci diskurzního markeru. Z analýzy vzorku o velikosti 300 konkordančních řádků z korpusů SYN2020 a ORAL vyplývá, že třetina zkoumaných výrazů v obou korpusech má funkci příslovce. V korpusu SYN2020 je výraz *tak* nejčastěji užit ve funkci souřadicí spojky (56 % zkoumaných dokladů). V korpusu ORAL má funkci spojky pouze 18 % zkoumaných dokladů. V těchto případech je často v dané replice užit výraz *kdyby* nebo *když* (příklad 56).

(56) *ale **když** to dáš někde v cizině **tak** tě to stejně fur* tahá na ty dálnice* (ORAL)

Ve zbylých případech má výraz *tak* funkci diskurzního markeru. V korpusu SYN2020 má tuto funkci 11 % analyzovaných dokladů. Většina z nich se nachází v přímé řeči postav v beletrii (příklad 57). V korpusu ORAL má funkci diskurzního markeru 48 % zkoumaných dokladů. Často se nacházejí na začátku repliky a plní navazovací či preparativní funkci, kterou zmiňuje L. Jílková (2016a,b). Mluvčí pomocí výrazu *tak* dává najevo, že souhlasí s druhým mluvčím. V takových případech je výraz *tak* často doprovázen výrazy *přesně*, *jo*, *no* (příklad 58).

(57) *„Klidně se mě zeptej, na co chceš.“ „Vážně můžu, na co chci?“ „Vážně.“ „**Tak** jo,“* (SYN2020)

(58) *A: dyť se říká jako zážitek nemusí byt dobrý ale hlavně když je silný*

*B: **přesně tak** . jako hnedka bych se tam chtěl vrátit (...)* (ORAL)

Na základě této analýzy vyplývá, že většina zkoumaných výrazů (89 %) v korpusu SYN2020 funguje jako příslovce či spojka. V korpusu ORAL má tuto funkci přibližně polovina zkoumaných výrazů, zbylé výrazy fungují jako diskurzní markery, které mají často navazovací či preparativní funkci.

#### 6.2.3.3.5 Lemma *tam*

SSČ uvádí, že výraz *tam* ukazuje nebo odkazuje na místo vzdálené od mluvčího dané promluvy, ukazuje nebo odkazuje směrem od mluvčího dané promluvy na vzdálený (prostorový) cíl. Na základě této charakteristiky není překvapivé, že se výraz *tam* užívá především v mluvené komunikaci. V korpusu SYN2020 je tento výraz také poměrně frekventovaný (jedná se o 80. nejfrekventovanější lemma), ale zdaleka není tak frekventovaný jako v korpusu ORAL (kde se jedná o 13. nejfrekventovanější lemma).

Vysoká frekvence lemmatu *tam* v korpusu ORAL je způsobena tím, že se mluvená komunikace odehrává v rámci určitého sdíleného kontextu (srov. oddíl 4.5). Díky sdílenému kontextu nemusí být mluvčí při jazykovém projevu vždy explicitní. V mluvené komunikaci často nastane situace, kdy mluvčí pojmenuje nějaký jev, předmět, objekt apod. explicitně, dále ale pro jeho pojmenování používá výrazy s vágním významem (srov. Čermák, 1993). K tomuto jevu často dochází i v případě užití výrazu *tam*. V příkladu 59 mluvčí k pojmenování obchodu použije poprvé konkrétní název (*Delvita*), ve svém projevu jej ale dále nahrazuje výrazem *tam*.

(59) *no tak to je Delvita . no a **tam** sem jí koupila původně když sem jí vezla do Jizerek ale pak jí **tam** přestali mít protože mam* sem to říkala mamce ale ona jí **tam** nenašla tak třeba jí **tam** už zas maj (...)* (ORAL)

Mluvčí může výraz *tam* použít i z pragmatických důvodů, pokud např. potřebuje získat čas, aby mohl lépe formulovat své myšlenky, případně pokud se snaží udržet si slovo. V takových případech dochází k hromadění výrazu *tam* v krátkém úseku, jako je tomu v příkladu 60. Pro tento jev jsem v korpusu SYN2020 nenašla žádný doklad.

(60) ***tam** je **tam** je **tam tam** sem třeba jednou **tam** měli divočáka a skutečně dobrý* (ORAL)

V korpusu SYN2020 výraz *tam* tvoří ustálená spojení typu *sem tam*; *sem a tam*; *tam a sem*; *tu a tam*; *hned sem, hned tam; krok sem, krok tam* apod. V korpusu ORAL se kromě *sem tam* ostatní spojení téměř nevyskytují.[43]

Podobně jako výraz *tam* funguje v mluvené komunikaci i výraz *tady*, který je v korpusu ORAL také poměrně frekventovaný. Zajímavé je, že synonymní výraz *zde* má v korpusu ORAL pouze 21 výskytů, v korpusu SYN2020 je ale poměrně frekventovaný (přibližně 33000 výskytů).

## 6.3 Shrnutí výsledků výzkumu

V úvodu jsem představila cíl této práce – odpovědět na následující otázky, které je možné shrnout do tří bodů:

1) Která lemmata se vyskytují ve frekvenční špičce korpusu SYN2020 a ORAL? Proč jsou tato lemmata tak frekventovaná, proč jsou frekventovanější než jiná lemmata?
2) Jaké jsou významy a funkce nejfrekventovanějších lemmat v korpusu SYN2020 a ORAL? V čem se tato lemmata shodují a v čem se liší? Užívají se v korpusu SYN2020 a ORAL shodně, nebo odlišně?
3) Co lze na základě nejfrekventovanějších lemmat v korpusu SYN2020 a ORAL vyvodit ohledně psané a mluvené komunikace?

Hledání odpovědí na tyto otázky bylo předmětem výše představeného výzkumu. V následujících odstavcích shrnu zásadní poznatky a zjištění, ke kterým jsem dospěla na základě analýzy zkoumaného materiálu.

Lemmata tvořící frekvenční špičku korpusů SYN2020 a ORAL zobrazuje tabulka č. 11. Prostřední sloupec uvádí rank (pořadí) lemmat v korpusu SYN2020, pravý sloupec uvádí rank (pořadí) lemmat v korpusu ORAL. Tabulka je rozdělena do tří částí, které odpovídají tomu, v jakém pořadí byla jednotlivá lemmata popisována. Nejprve jsem představila lemmata nacházející se ve frekvenční špičce obou korpusů. Následně jsem se zaměřila na lemmata nacházející se pouze ve frekvenční špičce korpusu SYN2020. Jako poslední jsem popsala lemmata, které se nacházejí pouze ve frekvenční špičce korpusu ORAL.

Frekvenční špičky obou korpusů tvoří převážně synsémantika, tedy slova, jejichž význam se rozvíjí až ve spojení s ostatními slovy. Tyto výrazy jsou velmi frekventované z toho důvodu, že mají v jazyce především gramatickou funkci a podílejí se na výstavbě a organizaci textu.

[43] Spojení *sem tam* (hledáno pomocí dotazu [lemma="sem"][lemma="tam"]) má sice v korpusu ORAL 216 výskytů, ve většině případů se ale jedná o chybnou lemmatizaci výrazu *sem*, který by měl být správně lemmatizován jako *být*.

Frekvenční špičky obou korpusů tvoří především zájmena, předložky, spojky, částice, příslovce a slovesa. Ani v jedné frekvenční špičce se nenachází substantivum, adjektivum či číslovka.

| **Frekvenční špičky korpusů SYN2020 a ORAL** | | |
|---|---|---|
| **Lemmata nacházející se ve frekvenční špičce obou korpusů** | | |
| lemma | rank (SYN2020) | rank (ORAL) |
| *a* | 2 | 3 |
| *ale* | 16 | 15 |
| *jako* | 19 | 12 |
| *že* | 9 | 9 |
| *ten* | 6 | 1 |
| *já* | 18 | 6 |
| *on* | 7 | 7 |
| *být* | 1 | 2 |
| *mít* | 12 | 11 |
| *na* | 5 | 14 |
| *v* | 4 | 16 |
| *se* | 3 | 5 |
| **Lemmata nacházející se pouze ve frekvenční špičce korpusu SYN2020** | | |
| lemma | rank (SYN2020) | rank (ORAL) |
| *který* | 11 | 80 |
| *moci* | 20 | 38 |
| *do* | 13 | 28 |
| *k* | 17 | 76 |
| *o* | 14 | 60 |
| *s* | 8 | 27 |
| *z* | 10 | 31 |
| *i* | 15 | 70 |
| **Lemmata nacházející se pouze ve frekvenční špičce korpusu ORAL** | | |
| lemma | rank (SYN2020) | rank (ORAL) |
| *hmm* | 4 104 | 17 |
| *jo* | 579 | 10 |
| *no* | 212 | 4 |
| *ne* | 90 | 19 |
| *co* | 24 | 20 |
| *vědět* | 57 | 18 |
| *tak* | 26 | 8 |
| *tam* | 80 | 13 |

*Tabulka 11: Frekvenční špičky korpusů SYN2020 a ORAL*

Frekvenční špičky obou korpusů jsou dále tvořeny lemmaty, která mají více významů. Tato lemmata bývají řazena k více slovním druhům. Řada autorů upozorňuje na to, že slovnědruhová platnost těchto lemmat není jasná (srov. Čermáková – Kopřivová, 2018; Hoffmannová – Zeman, 2017). Často se jedná o výrazy s vágním významem, případně o výrazy, u kterých došlo k oslabení (či zániku) původního významu a následnému přechodu k jinému slovnímu druhu.

Tradiční slovnědruhové zařazení může být problematické především u výrazů v mluvené komunikaci, z toho důvodu bývají tyto výrazy často charakterizovány jako diskurzní markery.

V obou korpusech jsou velmi frekventovaná lemmata *být*, *mít* a *moci*. Tato slovesa spojuje to, že mají plný lexikální význam, ale na rozdíl od většiny ostatních sloves fungují také jako slovesa pomocná. Tato pomocná funkce je jeden z důvodů, proč jsou lemmata *být*, *mít* a *moci* velmi frekventovaná jak v mluvené, tak i v psané komunikaci.

Oproti frekvenční špičce korpusu ORAL spadá mezi nejfrekventovanější lemmata v korpusu SYN2020 poměrně velké množství předložek. Ve frekvenční špičce korpusu SYN2020 se nacházejí výrazy *na*, *v*, *do*, *k*, *o*, *s* a *z*. Ukázalo se, že poměrně velký počet výskytů těchto lemmat v korpusu SYN2020 je součástí víceslovných předložek (např. *v rámci*, *na rozdíl od*, *na základě*, *vzhledem k*)[44]. Tyto předložky jsou v psané komunikaci velmi frekventované, naopak v mluvené komunikaci se nevyskytují příliš často.

Ve frekvenční špičce korpusu SYN2020 se nachází více spojek. Ukázalo se, že řada výrazů tvořících frekvenční špičku korpusu SYN2020 je součástí víceslovných spojek (např. *a proto*, *a tedy*, *ale i*, *dokonce i*, *i kdyby*). Naopak v korpusu ORAL nejsou tyto víceslovné spojky příliš frekventované. S tímto zjištěním souvisí i další jev – spojky v korpusu SYN2020 často vyjadřují více poměrů. Například spojka *a* (příp. víceslovná spojka, která obsahuje výraz *a*) vyjadřuje slučovací, stupňovací, vylučovací, důsledkový či odporovací poměr. V korpusu ORAL se spojka *a* užívá především pro vyjádření slučovacího poměru.

Frekvenční špičku korpusu ORAL tvoří obecně více zájmen a výrazů s deiktickou funkcí (např. *ten*, *já*, *on*, *tam*, *co*). Vysoká frekvence těchto výrazů je způsobena tím, že se mluvená komunikace odehrává v rámci určitého sdíleného kontextu (srov. oddíl 4.5). Díky sdílenému kontextu nemusí být mluvčí při jazykovém projevu vždy explicitní. V mluvené komunikaci často nastane situace, kdy mluvčí pojmenuje nějaký jev, předmět, objekt apod. explicitně, dále ale pro jeho pojmenování používá výrazy s vágním významem (srov. Čermák, 1993). Vysoká frekvence těchto výrazů (především lemmat *já* a *on*) je způsobena participací více mluvčích. Mluvčí často potřebuje dát najevo, které argumenty či myšlenky jsou jeho vlastní (užití lemmatu *já*) a které náleží někomu jinému (užití lemmatu *on*). Velmi frekventované lemma *ten* může v některých případech nabývat povahy určitého členu (srov. Schneiderová, 1993).

[44] O frekvenci víceslovných předložek viz tabulky č. 8 a 9.

Výrazy tvořící frekvenční špičku korpusu ORAL mají často pragmatickou funkci. Na základě definice A. Čermákové a kol. (2019) (představené v oddílu 4.1) je možné tyto výrazy považovat za diskurzní markery. Vysoká frekvence některých výrazů je způsobena dialogickou povahu korpusu ORAL, především vysoká frekvence výrazů interaktivního charakteru, např. *hmm*, *jo*, *no* (tyto výrazy nejsou v korpusu SYN2020 příliš frekventované, viz tabulka č. 11).

Na základě výše představené analýzy se ukázalo, že některá lemmata, která jsou velmi frekventovaná jak v korpusu SYN2020, tak v korpusu ORAL, mohou v těchto korpusech často plnit odlišné funkce. Například výrazy *jako*, *ale* a *tak* fungují v korpusu SYN2020 primárně jako spojka. Naopak v korpusu ORAL jsou tyto výrazy užívány především z pragmatických důvodů a fungují jako diskurzní markery. Také u ostatních zkoumaných výrazů v korpusu ORAL se ukázalo, že mohou často fungovat jako diskurzní markery, nejedná se však o jejich primární funkci (např. *a*, *že*, *ten*, *vědět*).

Zajímavé jsou případy, ve kterých jsou některé zkoumané výrazy z korpusu SYN2020 užity ve funkci diskurzních markerů. Nejedná se o časté případy, přesto jsem ve zkoumaných vzorcích z korpusu SYN2020 našla několik dokladů pro tato užití. Většinou se vyskytují v beletrii, kde se snaží napodobit mluvenou řeč. Často jsou užity v přímé řeči postav, případně se mohou vyskytovat ve vnitřním monologu některé z postav. V publicistice a v oborové literatuře se objevují nejčastěji v doslovných přepisech rozhovorů (je tedy otázkou, do jaké míry se jedná o psanou komunikaci, případně jestli se nejedná spíše o transkripci komunikace mluvené). Užití diskurzních markerů v psané komunikaci bylo dosud věnováno poměrně málo pozornosti, jedná se však o zajímavý jev, který by měl být v budoucnu prozkoumán podrobněji.

# Závěr

V této práci jsem zkoumala 20 nejfrekventovanějších lemmat v psaném korpusu SYN2020 a mluveném korpusu ORAL. Pro tato lemmata jsem v práci užívala pojem *frekvenční špička*. Cílem práce bylo stanovit, která lemmata tvoří frekvenční špičky zmíněných korpusů, zkoumat, jaké významy a funkce mají jednotlivá lemmata, a především zjistit, zda se tyto významy a funkce v použitých korpusech liší.

V teoretické části jsem představila příslušnou terminologii a klíčové pojmy pro tuto práci. Postupně jsem uvedla, co je to frekvence a jaké je její využití v jazyce. Následně jsem se zabývala vztahem mezi mluvenou a psanou komunikací. V souvislosti s tím jsem zmínila paradox reprezentace mluveného jazyka prostřednictvím grafického systému a upozornila jsem, jaké problémy či omezení s sebou může tento paradox nést. Zároveň jsem představila dva přístupy k výzkumu mluveného a psaného jazyka. V další části jsem se věnovala slovní zásobě psané a mluvené komunikace. Důležitou částí této kapitoly byl popis diskurzních markerů, při kterém jsem vycházela z jejich charakteristiky podle A. Čermákové a kol. (2019). V poslední kapitole teoretické části jsem představila některé práce či výzkumy, které se věnují podobnému tématu jako tato diplomová práce.

V praktické části jsem představila analýzu jednotlivých lemmat. Při analýze jsem vycházela především z významů popsaných ve výkladových slovnících (ASSČ, SSČ a SSJČ) a z odborných prací, které se zkoumaným lemmatům věnují. Z důvodu lepší přehlednosti jsem zkoumaná lemmata rozdělila do tří následujících skupin: (1) Lemmata nacházející se ve frekvenční špičce obou korpusů; (2) Lemmata nacházející se pouze ve frekvenční špičce korpusu SYN2020; (3) Lemmata nacházející se pouze ve frekvenční špičce korpusu ORAL.

Přínos této diplomové práce spočívá především v prohloubení a rozšíření poznatků ohledně psané a mluvené komunikace. Z analýzy lemmat vyplývá, že obě frekvenční špičky tvoří především synsémantika. Frekvenční špička korpusu SYN2020 obsahuje více předložek a spojek. Tyto předložky a spojky často tvoří víceslovné jednotky (víceslovné předložky či víceslovné spojky), které ale v korpusu ORAL nejsou příliš frekventované. Vysoká frekvence spojek poukazuje na složitější syntax psané komunikace. Syntax mluvené komunikace je jednodušší a méně strukturovaná. V mluvené komunikaci jsou k sobě syntaktické celky řazeny volně, často mezi nimi není explicitně vyjádřen významový vztah. Naopak v psané komunikaci jsou velmi frekventované spojky, které vyjadřují různé významové vztahy.

Frekvenční špičku korpusu ORAL tvoří více zájmen a výrazů s deiktickou funkcí. V mluvené komunikaci se vyskytuje větší množství zájmen, která poukazují k účastníkům komunikace (především *já*, *on* a *ty*). Díky sdílenému kontextu, v rámci kterého mluvená komunikace probíhá, se mluvčí nemusí vyjadřovat explicitně a může použít výrazy s vágním významem. V mluvené komunikaci často dochází k nadbytečnému užívání ukazovacích zájmen. Někdy u těchto slov dochází ke ztrátě původního významu, v takových případech se jedná spíše o pseudoukazovací zájmena (srov. Hoffmannová – Zeman, 2017).

Analýza ukázala, že některá lemmata, která jsou velmi frekventovaná jak v korpusu SYN2020, tak v korpusu ORAL, se mohou v těchto korpusech užívat různě. Zatímco výrazy užívané v korpusu SYN2020 obvykle odrážejí významy popsané v použitých výkladových slovnících, v korpusu ORAL u těchto výrazů často narážíme na užití a významy, které nejsou ve zmíněných slovnících popsány. Často se jedná o případy, kdy mluvčí tyto výrazy užije z pragmatických důvodů, mohou plnit například následující funkce: pomáhají mluvčímu získat, předat či udržet si slovo; slouží ke zpomalení promluvy tehdy, pokud si mluvčí potřebuje rozmyslet, jak bude promluva dále pokračovat; potvrzují posluchačovu pozornost; označují novou či starou informaci; vyjadřují postoj mluvčích k řečenému apod. Tyto výrazy odpovídají charakteristice diskurzních markerů (srov. A. Čermáková a kol., 2019). Diskurzní markery jsou v mluvené komunikaci velmi frekventované, protože mluvená komunikace probíhá v reálném čase a vzniká lineárně. Mluvčí obvykle nemá čas si promluvu předem rozvrhnout, proto má řada výrazů v mluvené komunikaci pragmatickou funkci. Naopak v psané komunikaci má autor možnost text opravit, přeformulovat apod., z toho důvodu nejsou výrazy s pragmatickou funkcí v psané komunikaci příliš časté.

Na základě relativní frekvence lemmat tvořících frekvenční špičku obou korpusů (viz tabulka č. 6, oddíl 6.2.3) je možné konstatovat, že mluvený jazyk je více repetitivní a omezený, naopak psaný jazyk je lexikálně bohatší a rozmanitější. Omezenost a repetitivnost mluveného jazyka se projevuje různými rutinními konverzačními obraty (srov. Čermáková – Kopřivová, 2018). V mluvené komunikaci často dochází ke kumulaci stejných výrazů (především u výrazů *jo*, *hmm* a *no*). Lemmata tvořící frekvenční špičku korpusu ORAL se často vzájemně kombinují, tvoří například tato spojení, která jsou v korpusu ORAL velmi frekventovaná: *že jo*, *to jo no*, *ale nevim*, *to víš*, *víš co*, *tak jo*, *to ne* apod. Pomocí těchto spojení na sebe mluvčí vzájemně

reagují. Dávají najevo, že druhého mluvčího poslouchají, rozumí řečenému, orientují se v hovoru a s řečeným ne/souhlasí.

Práce se soustředila zejména na sémantické vlastnosti lemmat tvořících frekvenční špičky korpusů SYN2020 a ORAL. Poznatky o frekvenčních špičkách by bylo možné dále rozšiřovat jejich zkoumáním skrze další jazyková hlediska, např. z hlediska syntaktického. Užitečné by dále bylo prozkoumat frekvenční špičky jiných korpusů či textových makrotypů zastoupených ve vybraných korpusech. Zajímavé by bylo např. srovnání frekvenčních špiček mluveného korpusu ORATOR (který obsahuje monologické projevy) a textových makrotypů zastoupených v korpusu SYN2020 (či jiném psaném korpusu). Toto srovnání je stručně představeno v oddílu 6.2.3 (tabulka č. 7). Otevírá tak prostor pro další, podrobnější výzkum.

# Seznam použité literatury


ADAMOVIČOVÁ, A. (2017): Čeština v korpusech. In Škodová, S. (ed.): *Přednášky z 60. běhu Letní školy slovanských studií*. Praha: FF UK, s. 44–60.

*Akademický slovník současné češtiny* [online] (2017–2022): Praha: Ústav pro jazyk český AV ČR, v. v. i. Dostupné z https://slovnikcestiny.cz [Poslední přístup 15. 11. 2022].

ALLWOOD, J. (1998): Some Frequency based Differences between Spoken and Written Swedish. In *Proceedings of the 16th Scandinavian Conference of Linguistics*. Turku University, Department of Linguistics, s. 18–29.

BAKER, P. – HARDIE, A. – MCENERY, T. (2006): *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.

BALHAR, J. a kol. (2002): *Český jazykový atlas 4*. Praha: Academia.

BANÁŠOVÁ, M. – HANSEN, B. – NEKULA, M. (2011): Nová konstrukce „*Karla Gotta nemusím*“ v češtině a slovenštině: případ lexikalizace, pragmatikalizace nebo začínající degramatikalizace?. *Slovo a slovesnost* 72, č. 4, s. 243–267.

BIBER, D. – JOHANSSON, S. – LEECH, G. – CONRAD, S. – FINEGAN, E. (1999): *Longman Grammar of Spoken and Written English*. London: Longman.

BRAZIL, D. (1995): *A Grammar of Speech*. Oxford: Oxford University Press.

BRINTON, L. J. (1996): *Pragmatic Markers in English: Grammaticalization and Discourse Functions.* Berlin – New York: Walter de Gruyter.

BROWNE, C. – CULLIGAN, B. – PHILLIPS, J. (2013): *The New General Service List*. Dostupné z http://www.newgeneralservicelist.org [Poslední přístup 18. 9. 2022].

BŘEZINA, V. – GABLASOVÁ, D. (2015): Is There a Core General Vocabulary? Introducing the New General Service List. *Applied linguistics* 36, č. 1, s. 1–22.

CARTER, R. – MCCARTHY, M. (1995): Grammar and the Spoken Language. *Applied Linguistics* 16, č. 2, s. 141–158.

CARTER, R. – MCCARTHY, M. (2017): Spoken Grammar: Where Are We and Where We Are Going? *Applied Linguistics* 38, č.1, s. 1–20.

CVRČEK, V. (2017a): Frekvence. In Karlík, P. – Nekula, M. – Pleskalová, J. (eds.): *CzechEncy – Nový encyklopedický slovník češtiny*. Dostupné z https://www.czechency.org/slovnik/FREKVENCE [Poslední přístup 8. 9. 2022].

CVRČEK, V. (2017b): Type-token. In Karlík, P. – Nekula, M. – Pleskalová, J. (eds.): *CzechEncy – Nový encyklopedický slovník češtiny*. Dostupné z https://www.czechency.org/slovnik/TYPE-TOKEN [Poslední přístup 28. 9. 2022].

CVRČEK, V. (2019): *Calc: Korpusová kalkulačka.* Praha: FF UK. Dostupné z http://www.korpus.cz/calc [Poslední přístup 20. 11. 2022].

CVRČEK, V. a kol. (2015): *Mluvnice současné češtiny 1. Jak se píše a jak se mluví*. Praha: Karolinum.

CVRČEK, V. – KOVÁŘÍKOVÁ, D. (2011): Možnosti a meze korpusové lingvistiky. *Naše řeč* 94, č. 3, s. 113–133.

CVRČEK, V. – RICHTEROVÁ, O. (eds.): *Příručka ČNK*. Dostupné z https://wiki.korpus.cz/doku.php [Poslední přístup 24. 10. 2022].

ČERMÁK, F. (1993): Lexis in Spoken and Written Language. *Slovo a slovesnost* 54, č. 3, s. 173–177.

ČERMÁK, F. (2008): Partikule, jejich syntagmatika a kumulace v mluvené češtině. In Kopřivová, M. – Waclawičowá, M. (eds.): *Čeština v mluveném korpusu*. Praha: NLN. s. 63–74.

ČERMÁK, F. (2009): Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics* 14, č. 1, s. 113–123.

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

ČERMÁK, F. a kol. (2007): *Frekvenční slovník mluvené češtiny*. Praha: Karolinum.

ČERMÁK, F. – KŘEN M. (2004): *Frekvenční slovník češtiny*. Praha: NLN.

ČERMÁK, F. – KŘEN, M. (2011): *A Frequency Dictionary of Czech: Core Vocabulary for Learners*. London: Routledge.

ČERMÁK, F. – SGALL, P. (1997): Výzkum mluvené češtiny: jeho situace a problémy. *Slovo a slovesnost* 58, č. 1, s. 15–25.

ČERMÁKOVÁ, A. – JÍLKOVÁ, L. – KOMRSKOVÁ, Z. – KOPŘIVOVÁ, M. – POUKAROVÁ, P. (2019): Diskurzní markery. In Hoffmannová, J. – Homoláč, J. – Mrázková, K. (eds.): *Syntax mluvené češtiny*. Praha: Academia, s. 244–351.

ČERMÁKOVÁ, A. – KOPŘIVOVÁ, K. (2018): Korpusový výzkum mluveného jazyka na příkladu češtiny a angličtiny: současný stav. *Slovo a slovesnost* 79, č. 3, s. 217–240.

DEJMEK, B. (2002): Frekvence slov a tvarů substantiv a sloves v běžně mluveném jazyce. *Slovo a slovesnost* 63, č. 1, s. 15–25.

FRIED, M. (2011): Vztažné věty s nesklonným *co*. In Štícha, F. (ed.): *Kapitoly z české gramatiky*. Praha: Academia, s. 1126–1143.

HOFFMANNOVÁ, J. (1992): Mluvené a psané texty ve vzájemných citacích. *Stylistyka* 1, s. 67–81.

HOFFMANNOVÁ, J. (2010): České *jako* a slovenské *ako/akože* v mluvených projevech (malý konfrontační pokus). In Šimková, M. (ed.): *Slovo – Tvorba – Dynamickosť: Na počest Kláry Buzássyovej*. Bratislava: Veda – Jazykovedný ústav Ľ. Štúra SAV, s. 359–371.

HOFFMANNOVÁ, J. (2017): „Malá slova“ a jejich podíl na stylu mluvených projevů. *Stylistyka* 26, s. 173–181.

HOFFMANNOVÁ, J. – KOMRSKOVÁ, Z. – POUKAROVÁ, P. (2017): Reprodukce řeči/myšlení v mluvených projevech jako předmět korpusového výzkumu. *Korpus – gramatika – axiologie* 15, s. 3–25.

HOFFMANNOVÁ, J. – ZEMAN, J. (2017): Výzkum syntaxe mluvené češtiny: inventarizace problémů. *Slovo a slovesnost* 78, č. 1, s. 45–66.

JELÍNEK, J. – BEČKA, J. V. – TĚŠITELOVÁ, M. (1961): *Frekvence slov, slovních druhů a tvarů v českém jazyce*. Praha: SPN.

JÍLKOVÁ, L. (2016a): *Tak* v současných mediálních mluvených rozhovorech: *ty zákony co se vztahujou teďkon na nás soukromníky tak to asi budu muset zavřít* (část 1). *Korpus – gramatika – axiologie*, č. 14, s. 12–33.

JÍLKOVÁ, L. (2016b): *Tak* v současných mediálních mluvených rozhovorech: *ty zákony co se vztahujou teďkon na nás soukromníky tak to asi budu muset zavřít* (část 2). *Korpus – gramatika – axiologie*, č. 15, s. 31–46.

JÍLKOVÁ, L. (2017): Ne/přízvukování slov *no* a *jo*. *Naše řeč* 100, č. 3, s. 177–185.

JÍLKOVÁ, L. (2018): Slovní spojení *až tak* v současné mluvené češtině. *Korpus – gramatika – axiologie*, č. 17, s. 23–41.

JÍLKOVÁ, L. (2021): Ke slovnímu spojení *takže tak* (*no nehlídala jsem nic .. sebe .. no .. takže tak*). *Korpus – gramatika – axiologie*, č. 23, s. 3–13.

KARLÍK, P. (2017): Auxiliár. In Karlík, P. – Nekula, M. – Pleskalová, J. (eds.): *CzechEncy – Nový encyklopedický slovník češtiny*. Dostupné z https://www.czechency.org/slovnik/AUXILIÁR [Poslední přístup 20. 11. 2022].

KIRK, J. – ANDERSEN, G. (2016): Compilation, Transcription, Markup and Annotation of Spoken Corpora. *International Journal of Corpus Linguistics* 21, č. 3, s. 291–298.

KLIMEŠOVÁ, P. – KOMRSKOVÁ, Z. – KOPŘIVOVÁ, M. – LUKEŠ, D. (2015): Slovo *to* v mluvených korpusech ČNK, jeho prefixace a reduplikace. *Časopis pro moderní filologii* 97, č. 1, s. 21–30.

KOCH, P. – OESTERREICHER, W. (1985): Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld und Sprachgeschichte. *Romanistisches Jahrbuch* 36, s. 15–43.

KOCHOVÁ, P. – OPAVSKÁ, Z. (eds.) (2016): *Kapitoly z koncepce Akademického slovníku současné češtiny.* Praha: Ústav pro jazyk český AV ČR, v. v. i.

KOMRSKOVÁ, Z. (2018): *Chápeš* a *rozumíš* v mluveném spontánním dialogu. *Nová čeština doma a ve světě*, č. 1, Praha: FF UK, s. 53–61.

KOPŘIVOVÁ, M. – LAUBEOVÁ, Z. – LUKEŠ, D. – POUKAROVÁ, P. (2020): *ORATOR v2: Korpus monologů*. Praha: Ústav Českého národního korpusu FF UK. Dostupný z https://www.korpus.cz.

KOPŘIVOVÁ, M. – LUKEŠ, D. – KOMRSKOVÁ, Z. – POUKAROVÁ, P. – WACLAWIČOVÁ, M. – BENEŠOVÁ, L. – KŘEN, M. (2017): *ORAL: korpus neformální mluvené češtiny, verze 1 z 2. 6. 2017.* Praha: Ústav Českého národního korpusu FF UK. Dostupný z http://www.korpus.cz.

KOVÁŘÍKOVÁ, D. – KOVÁŘÍK, O. (2021): *Akalex*. Praha: FF UK. Dostupné z http://www.korpus.cz/akalex.

KŘEN, M. (2017): Frekvenční slovník. In Karlík, P. – Nekula, M. – Pleskalová, J. (eds.): *CzechEncy – Nový encyklopedický slovník češtiny*. Dostupné z https://www.czechency.org/slovnik/FREKVENČNÍ%20SLOVNÍK [Poslední přístup 15. 9.].

KŘEN, M. – CVRČEK, V. – HENYŠ, J. – HNÁTKOVÁ, M. – JELÍNEK, T. – KOCEK, J. – KOVÁŘÍKOVÁ, D. – KŘIVAN, J. – MILIČKA, J. - PETKEVIČ, V. – PROCHÁZKA, P. – SKOUMALOVÁ, H. – ŠINDLEROVÁ, J. –ŠKRABAL, M. (2020): *SYN2020: reprezentativní korpus psané češtiny*. Praha: Ústav Českého národního korpusu FF UK. Dostupný z http://www.korpus.cz.

LEECH, G. (2000): Grammars of Spoken English: New Outcomes of Corpus-Oriented Research. *Language Learning* 50, č. 4, s. 675–724.

LEECH, G. – RAYSON, P. – WILSON, A. (2001): *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman.

MCCARTHY, M. (2001): Is There a Basic Spoken Vocabulary: Technology and Common Sense. *The Journal of TESOL France* 8, s. 29–36.

MÜLLEROVÁ, O. (1996): Výskyt a funkce slova *no* v českých textech prostě sdělovacího stylu. *Stylistyka* 5, s. 222–229.

NATION, P. – WARING, R. (1997): Vocabulary Size, Text Coverage and Word Lists. In Schmitt, N. – McCarthy, M. (eds.): *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University, s. 6–19.

*Slovník spisovné češtiny pro školu a veřejnost* (1994): Praha: Academia.

*Slovník spisovného jazyka českého* (1960–1971). Praha: Nakladatelství ČSAV.

SCHNEIDEROVÁ, E. (1993): Užívání zájmena *ten* (v přívlastkové pozici) v mluvených projevech. *Naše řeč* 76, č. 1, s. 31–37.

SLÁMA, J. – ŠIMÍK, R. (2021): Vztažné věty s relativizátorem *jak*. *Naše řeč* 104, č. 4, s. 207–224.

STOECKEL, T. – BENNETT, P. (2015): A Test of the New General Service List. *Vocabulary Learning and Instruction* 4, č. 1, s. 1–8.

VÁHALA, F. (1960): Tam mne jistě najdeš — to se mně líbí. *Naše řeč* 43, č. 7–8., s. 251–252.

WEST, M. (1953): *A General Service List of English Words*. London: Longman, Green & Co.