Pedagogická fakulta
Faculty of Education
Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Jihočeská univerzita v Českých Budějovicích

Pedagogická fakulta

Katedra slovanských jazyků a literatur

Oddělení českého jazyka a literatury

Bakalářská práce

K problematice Českého národního korpusu

Vypracovala: Žaneta Zemanová

Vedoucí práce: PhDr. Milena Nosková, Ph.D.

České Budějovice 2019

Poděkování

Mé poděkování patří PhDr. Mileně Noskové, Ph.D., za odborné vedení, cenné rady a ochotu při vypracování bakalářské práce. Dále bych chtěla poděkovat také PhDr. Marii Kopřivové, Ph.D., za odbornou spolupráci a pomoc při praktické práci s Českým národním korpusem.

Prohlášení

Prohlašuji, že svoji bakalářskou práci jsem vypracovala samostatně, pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své bakalářské práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledky obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích dne 8. 7. 2019

…….....

Žaneta Zemanová

Anotace

K problematice Českého národního korpusu

Bakalářská práce se zaměřuje na Český národní korpus ve frazeologii. Český národní korpus má několik korpusových manažerů, v praktické části je využit KonText. Korpusová databáze obsahuje materiály psané i mluvené. Práce se bude zabývat psaným materiálem. Frazeologie je často spjata s archaickými nebo dobovými jevy. Frazémy se vyskytují v různých typech textů. Vytváří se nové a některé se obměňují. Po zadání dotazu CQL nám KontText v syn v7 vyhodnotí, jaká je frekvence frazémů v textech a počet výskytů. Budeme sledovat, který typ frazému se využívá nejčastěji v porovnání s ostatními typy, jako jsou lidová rčení. Ve frekvenci nás bude zajímat frekvence skupiny textových typů, textový typ, tematická oblast a periodicita.

Klíčová slova: korpusová lingvistika; KonText; Český národní korpus; frazeologie

Abstract

On the issue of the Czech National Corpus

The bachelor thesis focuses on the Czech National Corpus in phraseology. The Czech National Corpus has several corpus managers, in the practical part, „KonText“ is used. The corpus database contains both written and spoken materials. The thesis will deal with written materials. Phraseology is often associated with archaic or period phenomena. Phrases occur in different types of texts. New ones are being created and some are changing. After entering the CQL, KontText in syn v7 will evaluate the frequency of the phrases in the texts and the number of occurrences. We will observe which type of phrase is most often used in comparison with other types, such as folk saying. We will be interested in the frequency of a group of text types, text type itself, thematic area and periodicity.

Key words: corpus linguistics; KonText; Czech National Corpus; phraseology

Obsah

Seznam použitých značek

Poziční atributy:

col_lemma (collocation lemma): lemma víceslovné jednotky v podobě slovníkového hesla v základním tvaru (nominativ singuláru, infinitiv apod.);

col_type (collocation type): dvoupísmenná značka, jejíž první písmeno určuje druh víceslovné jednotky a druhé slouží k rozlišení hlavního (`H`) a závislého (`Z`) slova v ní;

tag: gramatická informace automaticky přiřazená ke každému tvaru/wordu;

M: přísloví, okřídlená rčení, citace apod.;

Regulární výrazy:

tečka (`.`): představuje jeden libovolný znak;

hvězdička (`*`): představuje libovolný počet (0 a více) opakování předchozího znaku nebo celku;

seznam (`[]`): představuje alternativu. Nabízí možnost vybrat jeden libovolný znak z těch, které jsou uvedeny v seznamu uvnitř hranatých závorek;

ampersand (&): a zároveň (and), platí všechny podmínky zároveň;

Morfologické značky:

N: 1. pozice, slovní druh, substantivum;

V: 1. pozice, slovní druh, verbum;

B: 2. pozice, sloveso, tvar přítomného nebo budoucího času;

N: 2.pozice, substantivum obyčejné;

Značení v korpusu:

doc.txtype_group: skupina textových typů;

doc.txtype: textový typ;

doc.genre: tématická oblast;

doc.periodicity: periodicita;

Filter: filtrování výsledku;

Freq: frekvence;

i.p.m.: relativní frekvence;

p/n: pozitivní/ negativní filtr;

word: slovo**.**

Zdroj:

CVRČEK, Václav a Olga RICHTEROVÁ (eds). cnk: syn2015 [online]. Příručka ČNK; 2016 [cit. 2019-06-17]. Dostupné na: http://wiki.korpus.cz/doku.php?id=cnk:syn2015&rev=1476702845

CVRČEK, Václav a Olga RICHTEROVÁ (eds). kurz:pokrocile_dotazy [online]. Příručka ČNK; 2018, [cit. 2019-06-17]. Dostupné na: http://wiki.korpus.cz/doku.php?id=kurz:pokrocile_dotazy&rev=1535546576

CVRČEK, Václav a Olga RICHTEROVÁ (eds). seznamy:frazemy [online]. Příručka ČNK; 2018 [cit. 2019-06-17]. Dostupné na: http://wiki.korpus.cz/doku.php?id=seznamy:frazemy&rev=1534164400

CVRČEK, Václav a Olga RICHTEROVÁ (eds). kurz:zobrazeni_dotazu [online]. Příručka ČNK; 2018 [cit. 2019-06-17]. Dostupné na: http://wiki.korpus.cz/doku.php?id=kurz:zobrazeni_dotazu&rev=1534344877

CVRČEK, Václav a Olga RICHTEROVÁ (eds). pojmy:dotazovaci_jazyk [online]. Příručka ČNK; 2016 [cit. 2019-06-17]. Dostupné na: http://wiki.korpus.cz/doku.php?id=pojmy:dotazovaci_jazyk&rev=1458408131

Úvod

Tématem předkládané bakalářské práce je problematika Českého národního korpusu ve frazeologii. Domnívám se, že frazémy se v psaném projevu často vyskytují, jelikož jejich užití posouvá význam k jinému smyslu.

Ve své práci využívám data z Českého národního korpusu, přičemž pracuji s korpusem verze syn v7. Korpusový manažer KonText je vhodnou aplikací při práci v korpusu. Celkově obsahuje 6 typů dotazů. Vhodný typ dotazu zvolíme podle toho, co v práci chceme analyzovat. Korpus je spolehlivým zdrojem informací, protože obsahuje rozsáhlé množství dat. Celý korpus je automaticky značkován, což znamená, že po zadání dotazu vyhledá přesné výskyty, a to jak v korpusu psaném, tak i mluveném. Při práci s korpusem je důležitý kritický přístup konkrétního badatele.

Předkládaná bakalářská práce se dělí na teoretickou a praktickou část. V teoretické části představuji korpusovou lingvistiku, popisuji tuto lingvistickou disciplínu a historii korpusů. Dále se věnuji Českému národnímu korpusu a práci s korpusovou databází. Uvádím Františka Čermáka a jeho spolupracovníky Českého národního korpusu a základní problematiku frazeologie.

Praktická část obsahuje analýzu dat psaného korpusu syn v7 výskytu frazémů, mj. i lidových rčení. Ve verzi syn v7 jsem si vytvořila subkorpus2015. Pracovala jsem s korpusovým manažerem KonText, kde jsem do dotazovacího řádku použila dotaz CQL: [tag="VB.*" & col_type="M.*"][tag="NN.*" & col_type="M.*"], tzv. typ podle slovnědruhového vzoru. Analyzovala jsem data výskytu verba v přítomném nebo budoucím čase s obyčejným substantivem víceslovné jednotky druhu M (přísloví, okřídlená rčení, citace apod.) Minimální frekvenci jsem nastavila na 10, jelikož pod touto hranicí byl výskyt malý. Druhou pozici col_typu jsem nechala libovolnou. Výchozím atributem je col_lemma. V kategorii typ textu se zaměřuji na frekvenci skupiny textových typů, textový typ, tematickou oblast a periodicitu. Právě v publicistice je frekvence frazémů vysoká. V závěru práce provádím shrnutí výsledků, kde komentuji a popisuji závěry praktické části.

Teoretická část

1. Korpusová lingvistika

1.1 Co je to korpus

Slovo korpus pochází z latinského slova *corpus*, *-oris* = 1. tělo, těleso hmota; 2. tělo, postava; 3. uspořádané těleso, celek, kmen, soubor, sbor. K nám toto slovo přišlo z angličtiny (*corpus*, pl. *corpora* nebo *corpuses*). Znamená buď sbírku všech psaných textů jednoho určitého druhu nebo jednoho člověka, nebo sbírku informací či materiálů určených ke studiu. V lingvistice je korpus považován (v obecné rovině) za soubor dokladů autentického užití přirozeného jazyka. Je to **materiálová základna** sloužící k lingvistické analýze a popisu, a to jak jazyka psaného, tak mluveného (pro potřeby výzkumů transkribovaného).[1]

V historii lingvistiky můžeme slovem korpus označit celou řadu sbírek lingvistického materiálu, která byla ručně pořízena několika spolupracovníky, většinou se těmto sbírkám říká např. lístkový katalog. V době rozkvětu počítačů ve vědě bylo možné dřívější ručně psané práce přepsat do elektronické podoby. Bylo to přínosem pro nový typ materiálové základny pro lingvistický výzkum. Korpus si tak získal své místo v aplikované lingvistice. V posledních letech se elektronické korpusy natolik využívají, že se korpus využívá téměř výhradně pro elektronicky uložené a uchovávané soubory textů. Korpusový lingvista se od svých kolegů liší tím, že pracuje spíše s introspekcí. Je to dáno tím, že v době korpusů se všechny doklady získávají přímo z textu.[2]

Dnes korpus využívá pro studium a poznání jazyka (vedle jiných oborů) jakákoliv empirická lingvistika. Korpus má mnohem širší využití. Slouží k poznání dobové reality, kterou je jazyk zachycen a zprostředkován obecně i odborně (např. sociologie, historie). Svou povahou korpus předčí jiné zdroje jazykového poznání a studia. Korpusy stále rostou, ať už směrem do budoucnosti nebo minulosti. Příkladem je Český národní korpus. Není zcela zachycena celá minulost dané psané národní kultury v elektronické podobě. Jazyk se stále vyvíjí. **Důvody rostoucího korpusu** a potřeba mapování je následující: *a) korpusy nejsou úplné, zvláště ve složkách mluveného jazyka; b) do jazyka přicházejí stále nová slova; c) objevují se další informace spolu s dalšími kombinacemi slov (už zachycených) při zkoumání kolokací.*[3]

[1] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 9.
[2] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 9-10.
[3] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství

Rozdíl mezi korpusem a běžným elektronickým archivem (prosté úložiště daných textů s omezenými možnostmi prohledávání) je, že korpusu je dodán vnější nástroj, propracovaný korpusový manažer (původně Manatee-Bonito). Tento korpusový manažer je schopný zpracovat několik miliard slov. Kromě základní funkce vyniká i prací s několika speciálními programy. Korpusový manažer je specificky pojatý a celkově je zaměřený na určitý cíl. Zpravidla to bývá na celý jazyk se záměrem ho textově maximálně pokrýt, což znamená, že jde o reprezentativnost korpusu. Předpokladem jazykového korpusu je soubor autentických textů, které nebyly upravovány. Texty pro korpus se shromažďují podle určeného a kvantifikovaného obrazu, který jsou žádoucí pro budoucí korpus. Cílem je získat z nich rozsáhlý základ. Lze tak za pomocí počítače a dalších specializovaných programů provádět výzkum, který vede k rozšíření a větší objektivitě při poznání jazyka v obecné rovině či v určitém směru. Korpusovou databázi můžeme využívat k mnohým účelům jako například k testování hypotéz v teoretické lingvistice nebo k práci nespecializovaného uživatele. Pokud hledáme jakékoliv slovo (nikoliv však přímo význam), dá se v korpusu rychle zjistit a nabídne v častých i méně častých kontextech autentický úzus (toto zachycení bylo a je největší slabinou většiny dosavadních mluvnic a slovníků).[4]

1.2 Typy korpusů

Korpusy vznikají podle potřeby uživatele a rozdělují se podle typologie. Obecně se dají podle jazyka a situace dále modifikovat.[5]

Vytvoření a technické zpracování korpusů probíhá v několika krocích. Nashromážděné texty v určité proporci procházejí, po zdokumentování, katalogizování a zanesení do bibliografické databáze, řadou technických automatických procedur, ve kterých se texty konvertují, technicky se zpracovávají a jejich části se označují.[6]

Dělení korpusových databází: 1. Podle množství obsažených jazyků dělíme korpusy na **korpusy jednoho jazyka** (např. Český národní korpus) a **paralelní korpusy** (parallel corpora) obsahující stejné texty ve dvou i více jazycích (originál a překlady). **2.** Z hlediska časového záběru na **korpusy diachronní** (diachronic

Karolinum, 2017, s. 37-38.
[4] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 38-39.
[5] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 74.
[6] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum,2017, s. 39.

corpora), které zpracovávají jazyk v průběhu delšího časového rozpětí (např. Helsinki Corpus of English Texts obsahuje texty od roku 750 do roku 1700), a **korpusy synchronní** (synchronic corpora), které se zabývají určitým časovým obdobím a není potřeba v něm přihlížet k vývojovým změnám (např. Brown Corpus obsahuje texty americké angličtiny jen z roku 1961). **3.** Některé **synchronní korpusy** byly sestaveny pro lingvistický účel, který nebyl blíže specifikován. Nikterak neupřednostňující oblast lingvistiky, rovinu jazyka či stylovou příslušnost pokrývají všeobecné korpusy. Využívají se ke studiu slovní zásoby, gramatiky a diskurzu (např. SEU Corpus). **Všeobecné korpusy** jsou zpracovány tak, aby byly vyvážené (balanced corpora). Při jejich sestavování musí být vzaty v potaz různé typy žánrů, média přenosu (jazyk mluvený i psaný), různé stupně formálnosti (komunikace soukromá a veřejná) atd. V nich je vyhrazena i část pro denní tisk (s ohledem na čtenost a vliv na jazyk). Tyto korpusy jsou v literatuře citovány jako základní (core corpora), zejména pokud byly využity v komparativních studiích. Většina korpusů je vytvořena ke konkrétnímu výzkumu. Takové korpusy se nazývají **specializované** (specialized corpora). Zkoumají rozvoj jazyka u dětí, nesouhlas v konverzaci, direktivní jazykové techniky učitele, chyby studentů určitého jazyka atd. Vznikají také speciální tréninkové a testovací korpusy pro studenty jazyka. **4.** Z hlediska běžné každodenní komunikace je počet korpusů **mluveného jazyka** (spoken corpora) velmi malý. Pro svoji náročnost a nákladnost mají malý rozsah oproti korpusu psaného jazyka (written corpora). **5.** Jedním z rysů korpusů je způsob **reprezentace jazyka**. V méně častých případech obsahují celou entitu, která je předmětem lingvistického popisu. **6.** Většina korpusů je založena pro obecné zkoumání jazyka. Pracuje se tak se vzorky a musí se dobře zvolit kategorie a typy, které mají celek reprezentovat. Zvažuje se proto kritérium, zda budou do korpusu vkládány celé texty (full-text corpus), nebo zda bude stačit pracovat s úryvky/vzorky (sample-text corpus). **7.** Značnou část korpusových databází tvoří uzavřené projekty, které jsou omezené časem i částkou, a tím omezené svou velikostí. Není tak časté, aby si projekt kladl za cíl pracovat s jazykem obecně (všeobecný korpus), a zároveň měl dostatečnou finanční podporu, aby zůstal korpusem otevřeným. Takové se nazývají korpusy **monitorovací** (monitor corpora). Malou nevýhodou může být nevyváženost při strategii. **8.** Pokud mohou lingvisté přidat k „čistému“ textu dodatečné informace, pak takové korpusy obsahují **tagování** (tagging), tj. značkování slovních druhů a někdy i morfologických kategorií, jednak i korpusy, které obsahují **parsing**, tj. značkování syntaktické. Paralelní korpusy mají ještě **zarovnání**

(alignment), což ukazuje jednotku v textu v jednom jazyce a je k ní přiřazena odpovídající jednotka v jazyce druhém. **9.** Zvláštní typ korpusu je takovou databází, která je nějakým způsobem zpracovaná a neobsahuje všechen vstupní jazykový materiál. Neobsahuje celé texty, ale jen určité citace. Nejčastěji se s tím setkáme při prezentaci lingvistů v jejich odborných pracích, tj. ve slovnících a v seznamech slov.[7]

1.3 Historie korpusové lingvistiky

1.3.1 Počátky do roku 1960

Korpusová lingvistika v dnešním pojetí (v elektronické podobě) se datuje do 60. let 20. století. Předcházející lingvisté také pracovali s autentickým užitím jazyka i s citacemi. Nazývali to kartotékou, archivem atd. První využití citací pro studium používali lexikografové v 17. století. Samuel Johnson se svými pomocníky vytvořil na základě 150 tisíc citací 40 tisíc slovníkových hesel pro slovník *Dictionary of English Language.* Podobná práce byla v roce 1828 vydána jako soubor citací *An American Dictionary of English Language* od Noaha Webstera. Rozsáhlý slovník *Oxford English Dictionary* (OED) má 12 dílů, z toho 12. díl byl publikován v roce 1928. Obsahuje 414 825 hesel a bylo nasbíráno 5 milionů citací. Zájem byl i o biblická studia a studia krásné literatury.[8]

V roce 1736 Alexandr Cruden, knihkupec a korektor, vydal knihu *Complete Concordance to the Old and New Testaments*. Šlo o seznam konkordancí, které byly použity v Bibli. Později vyšla podobná práce, jež byla založena na tvorbě Shakespearea i dalších autorů. V 19. století se ve Velké Británii někteří lexikografové zabývali dialekty, vytvářeli korpusy i s podobou výslovnosti, např. *J. Wright: English Dialect Dictionary* (1898-1905) a *A. J. Ellis: The Existing Phonology of English Dialects* (1889). Práci s korpusy velmi využila v první polovině 20. století jazyková pedagogika. E. L. Thorndike sestavil korpus, jehož úkolem bylo vytvořit frekvenční seznam. Sloužil jako pomůcka pro rodilé mluvčí angličtiny ke studiu ortografie. Své výsledky sepsal do knihy *Teacher´s Wordbook* (1921). Významnější práce pro výuku angličtiny vznikla ve spolupráci s I. Lorgem, jejíž název byl *Teacher´s Wordbook of 30.000 Words* (1944). Lingvisté, kteří se zabývali lexikálně-gramatickým studiem angličtiny, vytvářeli v 50. a 60. letech své vlastní korpusy. Tento způsob práce byl častější i v dalších

[7] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 12-13.
[8] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 28.

zemích. Pro příklad lze uvést Japonsko, Severní Ameriku i Evropu. Větší množství materiálu potřebovala také gramatika, která vycházela ze souborů citací. Patří sem dílo od C. C. Friese *American English Grammar* (1940). Jeho další kniha vyšla v roce 1952, kde pracoval se záznamy telefonní konverzace, *The Structure of English.*[9]

Survey of English Usage (SEU) **Corpus** byl nejvýznamnějším neelektronickým korpusem, který se využíval ke studiu gramatiky. Randolph Quirk začal s tímto projektem v roce 1959 v Londýně, přičemž cílem bylo sesbírat 200 mluvených i psaných vzorků, každý měl mít přibližně 5 tisíc slovních výskytů. Materiály měly sloužit k popisu jazyka dospělého rodilého mluvčího britské angličtiny a byly zaměřené na studium gramatiky. Mluvený korpus usiloval o různé stupně formálnosti, různé náměty, různé situační kontexty jako konverzace, interview, přednáška, seminář, schůze a další. Vědecký jazyk však spíš převládal nad standardním. SEU Corpus se strukturou druhé poloviny původně psaných textů podobal prvním elektronickým korpusům a je významným prvkem v moderní korpusové lingvistice. Gramatické kategorie a jejich značky byly základem, když se v 80. letech tvořil **Brown Corpus**. SEU Corpus byl materiálovou základnou při shromažďování kompletních popisů současné anglické gramatiky: *A Comprehensive Grammar of the English Language* (1985). V elektronické verzi byl vytvořen až ve 2. polovině 70. let 20. století.[10]

1.3.2 První generace

První využití počítačů v lingvistice se vztahuje k začátku 60. let 20. století, začala tím tak nová éra práce s jazykovým materiálem. V roce 1961 Henry Kučera a Nelson Francis vytvářeli *Brown University Standard Corpus of Present-Day Edited American English,* běžně nazýván **Brown Corpus** (BC). Cílem bylo, aby byl „rozumně reprezentativní“ pro psanou americkou angličtinu, a zároveň byl zdrojem pro lingvistický výzkum. Negativní postoj vůči korpusům měl lingvista Noam Chomský, jenž tvrdil, že korpusy nejsou vhodným podkladem pro popis gramatiky. Pro korpus měl být nashromážděn přibližně 1 milion slov psaného textu. Korpus obsahoval několik žánrů imaginativní i informativní prózy (15 žánrů), ne však drama či poezii. V únoru 1963 se konala konference na Brownově univerzitě (H. Kučera, W. N. Francis, R. Quirk, P. B. Gove, P. O´Connor, J. B. Carroll),

[9] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 28-29.
[10] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 29-30.

kde kategorizace žánrů dostala konkrétní podobu. Struktura se tak stala standardem v korpusových výzkumech. Velkým přínosem pro vědce byla možnost využití korpusu zdarma, pokud šlo o vědecký výzkum, a ne komerční využití.[11]

Lancaster-Oslo/Bergen (LOB) **Corpus** se stal 2. významným elektronickým korpusem. Vznikl mezi léty 1970-1978. Na jeho vytvoření se podílelo několik univerzit. Britský korpus byl vytvořen jako protiklad k Brownovu korpusu. Obsahoval také 500 vzorků o průměrné velikosti 2 tisíc slovních výskytů, které vznikly v roce 1961. Struktura byla podobná americké verzi BC a korpus obsahoval psanou britskou angličtinu. Tento mladší projekt oproti americkému korpusu měl výhodu v rozvoji počítačové technologie. Nejprve byl umístěn na magnetickém pásku, dnes je uložen na CTDG, disketách, mikrofiších a na CD-ROMu. Kdokoli, kdo využije korpus ke své práci, může pracovat s původní verzí, s verzí gramaticky otagovanou a s verzí pro konkordantní program typu KWIC.[12]

Nijmegen Corpus byl jedním z prvních korpusů angličtiny, který vznikl v 70. letech pod vedením Jana Aartse mimo anglicky mluvící země. Několik dalších lingvistů chtělo vytvořit materiálovou základnu, jejímž účelem bylo vytvořit např. popis anglické gramatiky, především syntaxi. Bylo sesbíráno 6 úryvků psané standartní angličtiny od 6 různých autorů, doplněno bylo několik mluvených sportovních komentářů. Výsledky práce se potýkaly s několika nedostatky, ale zároveň byl tento projekt přínosem pro projekt **TOSCA Corpus**. **Survey of Spoken English** (SSE). Jde o švédský projekt, který vznikl na Lundské univerzitě v roce 1975. Cílem bylo převést mluvenou část SEU korpusu do elektronické podoby. K přepsaným 87 textů se přidalo dalších 13 a vznikl **London-Lund Corpus** (LLC), jeden z nejčastěji používaných mluvených korpusů angličtiny. Dalším projektem, který vycházel z BC, byl **Kolhapur Corpus of Indian English** (Kolhapur Corpus). Vycházel z textů vydaných v Indii v roce 1978. **Wellington Corpus of Written New Zealand English** svou pozornost zaměřil na vývoj krásné literatury. Nevýhodou bylo, že do korpusu nebyly zahrnuty texty z krásné literatury pro děti a mládež, které byly na Novém Zélandu často publikovány.[13]

Brownovým korpusem byly ovlivněny i další korpusy: **Australian Corpus of English** (ACE), jinak také **Macquarie Corpus of Written Australian English**, a **Corpus of English-Canadian Writing**. Na Freiburské univerzitě v roce 1991 vznikl

[11] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 30.
[12] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 30-31.
[13] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 31-32.

projekt, který se zaměřoval na srovnávací korpus. Zahrnoval texty psané americké a britské angličtiny. Projekt navazoval na modul Brownova korpusu. Kromě srovnávání měl zaznamenávat změny v psané angličtině za posledních 30 let. Do mluveného korpusu se řadí i **Corpus of Spoken American English** (CSAE), který vytvořil první velký korpus angličtiny tak, jak je používána dospělými Američany. Obsahuje 900 mluvčích (80 hodin hovoru) s různými dialekty i sociální různorodostí. Doposud byly korpusy nespecializované jednojazyčné synchronní, od konce 60. let 20. století začaly vznikat korpusy specializované. **American Heritage Intermediate** (AHI) Corpus vznikl v 70. letech pro lexikografické využití American Heritage School Dictionary. Bylo použito 10 043 vzorků z vybraných nejčtenějších knih americké mládeže ve věku od 7 do 15 let v roce 1969.[14]

Na konci 70. let diachronní lingvistika začala využívat nové možnosti zpracování lingvistických dat. **Complete Corpus of English** vznikl na Torontské univerzitě z *Dictionary of Old English*, který obsahoval materiál ke studiu staré angličtiny. Mezi první diachronní korpus angličtiny patří *Helsinki Corpus of English Texts: Diachronic Part* (1984-1991). Je v něm zahrnuto období od roku 750 do roku 1700. Na tento korpus navazuje **Archer Corpus** a zpracovává texty od poloviny 17. století do současnosti. Dalším diachronním korpusem je **Century of Prose Corpus**, který je sestaven texty od 120 autorů, kteří publikovali v letech 1680 až 1780. **Tools for Syntactic Corpus Analysis** (TOSCA) měl pomoci vytvořit pro lingvisty vhodné softwarové nástroje. Sběr dat obsahoval psané texty a publikované z roku 1976 a 1986. **Map Task Corpus** se zaměřil na mluvenou podobu jazyka skotských studentů Univerzity v Glasgow. **Corpus of London Teenager Language** (COLT), který je od roku 1994 součástí Britského národního korpusu, věnoval svou studii slovních tvarů dospívající londýnské mládeže ve věku od 13 do 17 let. Vznikaly korpusy se zaměřením na geografii a gramatiku i v oblasti psycholingvistiky (*Oxford Psycholinguistic Database*).[15]

Korpusy první generace byly počátkem jiné lingvistické práce. Rozdíly byly spíše ve frekvenci než v rozdílech rejstříku gramatických struktur. Z analýzy korpusů bylo zřejmé, že některé zkoumané oblasti lingvistiky daná velikost korpusu limituje v jeho užití. Bylo potřeba mnohem větších korpusů. V 90. letech vznikají mimo národní (neanglické) korpusy také korpusy paralelní (dvoujazyčné) i multilingvální.[16]

[14] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 32-33.
[15] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 33-34.
[16] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 34-35.

1.3.3 Druhá generace

Druhá velká vlna korpusů začala v 80. a 90. letech 20. století projektem **COBUILD Corpus** (*Collins Birmingham University International Language Database*). Vedl ho **J. Sinclair** a projekt se tak stal jedním z prvních lexikografických korpusů od 70. let. Korpus byl primárně projektován jako materiál pro nový slovník angličtiny. Na jeho projektování se podílelo nakladatelství Collins ve spolupráci s katedrou angličtiny univerzity v Birgminghamu. Materiál korpusu byl určen pro reprezentaci jazyka při výuce angličtiny pro studenty a učitele, ale i pro pomoc lingvistům, kteří se zabývají současnou angličtinou. Obsah korpusu zahrnuje materiál od roku 1960, zaměřen je na běžný jazyk a přínosem byly texty s neologismy. Texty imaginativní, informativní a próza obsahovaly 75 %, drama a poezie nebyly zahrnuty. Zbylých 25 % zaujímal mluvený jazyk, přepsány byly přednášky, interview, zvukové nahrávky univerzitního archivu a rádiové přenosy. Protože projekt byl určen pro vytvoření jednojazyčného slovníku (*Student´s Dictionary*), byly v korpusu použity texty a učební materiály pro studenty angličtiny (*TEFL Corpus*). Mělo tak vzniknout nové kurikulum pro učitele angličtiny jako cizího jazyka. V roce 1990 Sinclair oznámil rozšíření databáze a vznikl první monitorovaný korpus na světě **Bank of English** (BoE). BoE pojímal texty knih, časopisů, novin, letáků, návodů, transkribované texty rozhlasových pořadů a neformálních rozhovorů od roku 1989 do současnosti.[17]

Textové typy, které BoE rozlišuje: *Austrálie – australské noviny; UK – knihy (různé), populární časopisy, neformální hovor (schůze, konverzace), efeméra, BBC World Service Radio (interview, telefonáty, pořady s telefonní účastí posluchačů), Times (noviny), Today (noviny); USA - knihy (různé), národní veřejnoprávní rádio, efeméra.*[18]

Z BoE vznikl další projekt **Cobuild Direct**, který byl přístupný veřejnosti a byly do něho vybrány texty z let 1989-1996 o velikosti 50 milionů slovních výskytů. **British National Corpus** (BNC), korpus současné mluvené a psané britské angličtiny, byl projektován v letech 1991-1995. Obsahuje přibližně 60 % knižního textu, 25 % textu z periodik, 5 % z brožur či jiného příležitostného tisku, posledních 5 % bylo ve formě nepublikovaných dopisů, esejí, protokolů aj. Další psané texty byly určeny k mluvenému projevu (projevy, divadelní hry). V psaných textech korpusu byla zajišťována vnitřní klasifikace externích a interních rysů. Transkribované mluvené texty v BNC jsou největším mluveným korpusem na světě. Korpus je svou velikostí

[17] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 35-36.
[18] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 36.

i složením jedním z nejcitovanějších pramenů.[19]

Novější korpus pro angličtinu je **Corpus of Contemporary American English** (COGA) z let 1990-2011. Zároveň je dnes jedním největším americkým nereprezentativním synchronním psaným korpusem s rozsahem 450 milionů slov. Autorem je **M. Davies**. Korpus zaměřený na mluvený jazyk angličtiny pojímá **International Corpus of English**, který z 24 zemí zkoumá užití angličtiny jako mateřského či oficiálního jazyka. Protikladem BNC je ještě nedokončený **American National Corpus**. **DWDS** (Digitales Wörterbuch der Deutschen Sprache) má komplex 5 slovníků s vazbou na 15 podkladových korpusů, z nich nejdůležitějším je Kernkorpus des 20. Jahrhunderts (tj. Jádrový korpus 20. století). Korpus **DeReKo** (Deutsches Referenzkorpus) má kapacitu o velikosti 24 miliard slovních tvarů, tím je největším korpusem na světě. Obsahuje beletristické, odborné, populárně-naučné a publicistické texty, které jsou uloženy v několika korpusech. Veřejný přístup je omezen jen na část textů z důvodu autorských práv. Pro DeReKo byl vytvořen korpusový manažer COSMAS II, který nabízí mnoho možností při práci s korpusem. Korpusem reprezentujícím rakouskou němčinu je **AAC korpus**, který zahrnuje nejrůznější psané texty. V 60. letech 20. století se začal vytvářet projekt **Frantext**, korpus francouzského jazyka, jenž obsahuje více než 4000 literárních neanotovaných textů od středověku do současnosti.[20]

1.3.4 Specializované korpusy současnosti

Helsinský diachronní korpus angličtiny (Helsinki Corpus of English Texts: Diachronic and Dialectal Part) byl vytvořen v roce 1984 na univerzitě v Helsinkách katedrou anglického jazyka. Cílem bylo podpořit a usnadnit diachronní a dialektologická studia angličtiny. V korpusu jsou obsaženy texty, které pocházejí z období od roku 750 do roku 1700. Korpus neobsahuje materiály v cizích jazycích, ani poznámky editorů či korpusových pracovníků. V roce 1991 bylo v databázi 1 572 800 slovních výskytů v diachronních textech. Korpus byl tvořen na základě výběru textů za účelem vytvoření reprezentativního psaného jazyka určitého období. Z důvodu, že každý mluvčí může mít rozdílné varianty v oblasti gramatiky, bylo do korpusu zahrnuto velké množství vzorků každého individuálního mluvčího. Každý vzorek má kód, který identifikuje. Pro příklad lze uvést: *jméno souboru, mluvčího, pohlaví, věk, zaměstnání,*

[19] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 36-38.
[20] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 85-87.

zemi, obec, dialekt, osobu kladoucí otázky, číslo strany. Byla snaha o experiment propojení databáze se zvukem (aplikace Macintosh Hypercard a Mac-Recorder).[21]

Dánský korpus vznikl jako korpusový a lexikografický projekt mimo anglicky mluvící země. Textový korpus *Slovníku dánštiny* byl vytvořen kvůli obavě o dánskou kulturu a jazyk. **Den Danske Ordbog** (DDO) obsahuje slovník moderní dánštiny. V korpusu jsou materiály od roku 1950 a některé materiály jsou z 28 dílného *Slovníku dánského jazyka* (z let 1918-1956). Projekt vedl **O. Norling-Christensen** v letech 1991-1993 a korpus pojímá 40 milionů slovních tvarů. Texty pochází z let 1983-1992 (noviny, časopisy, knihy, rozhlas, televize, dopisy, letáky atd.).[22]

Synchronní korpusy se dostávají do slovanské oblasti velmi pomalu a nerovnoměrně. Příkladem je nejpoužívanější dostupný soubor ruských textů, který se nachází mimo Ruskou federaci, a to v Mannheimu. Nejstarším korpusem je korpus srbštiny, který vznikl v 90. letech, avšak z politicko-vojenských a ekonomických důvodů se dále nerozvíjí. Polský korpus vytváří lingvisté déle než dva roky. Slovenští lingvisté mají svůj korpus obsahově omezený. Práce na korpusu probíhá i v Bulharsku, kdy první přípravy začaly ve Slovinsku projektem **FIDA**. Nejlépe ze slovanské oblasti je na tom **Český národní korpus**. Netypickým korpusem je projekt **Církevněslovanský slovník makedonské redakce**, nástroj STINO nedokáže vyhledat a zobrazit kontexty hledaného slova, ale pouze slovo samotné. Databáze se zaměřuje na církevní slovanštinu makedonské redakce z 12.-16. století. Celý projekt vznikl v Ústavu makedonského jazyka ve Skopji.[23]

MULTEXT-EAST (Multilingual Tools and Resources for Central and East European Languages = MTE) je speciální multilingvální projekt, který byl vytvořen pro jazyky střední a východní Evropy v elektronické podobě jazykového materiálu. Navazuje na projekt z let 1993 až 1995 MULTEXT, který zahrnoval pouze jazyky západní Evropy. MULTEXT-EAST z roku 1995-1997 se týkal češtiny, slovinštiny, maďarštiny, bulharštiny, estonštiny a rumunštiny a zpracovával 3 úkoly: tzv. „srovnatelné korpusy“ (pro všechny výše uvedené jazyky vytvořit 2 malé korpusy); zpracování mluveného jazyka (problematikou této oblasti); tvorba „paralelních korpusů“, kdy vybraná kniha byla přeložena do všech výše uvedených jazyků (Orwell, G. 1984), celá jazyková verze textu byla přenesena do elektronické podoby a texty byly

[21] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 39-41.
[22] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 41-42.
[23] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 42-43.

otagovány.)[24]

Dnešním největším korpusem psané angličtiny se stal **Oxford English Corpus**. Texty pochází z let 2000-2006 z britské i dalších angličtin. Na univerzitě v Bologni vytvořili korpus současné psané italštiny **CORIS/CODIS**. Psaný synchronní korpus španělštiny, **CREA**, vytvořil M. Davis. Existuje několik i mimo evropských korpusů: čínský **Corpus Sinica**, japonský **BCCWJ**, arabský **ICA** atd.[25]

1.4 Korpusová lingvistika

Korpusová lingvistika jako disciplína je novějším odvětvím lingvistiky. Zabývá se výstavbou a analýzou jazykových korpusů na podobných datech jako tradiční lingvistika. Uplatní se všude, kde korpusová data můžou vnést svou informaci. Jako každou vědní disciplínu doprovází i ji rozvoj metodologie, tj. výstavba korpusů i jejich vytěžování, propracování souvisejících oblastí i oblastí nejazykových (ve vazbě komputační lingvistiky a informatiky). Korpusová lingvistika vyvrací postoj N. Chomského, který odmítá studovat skutečný jazyk pro „*jeho nedostatky, pokřivenosti, popř. i mezery*“ pocházející z terénu. Hlavní zásadou korpusové lingvistiky naopak je, oproti Chomského zaměření na výběrové a pro něho vymyšlené příklady, její zásadní, systematická orientace na nevýběrové vyčerpávající studium všech korpusových dat objektivními metodami zvláště statistickými, a to dat autentických a v reálném kontextu. Tato vědní disciplína navazuje na dřívější manuální shromažďování dat. Je to metodologická disciplína a přináší metodologické zkušenosti pro analýzu spolehlivých dat a jejich další studium pro každého. Dostupnost korpusových objektivních informací je snadná, a proto se tento obor stává ve světě stále populárnějším jak v lingvistice teoretické, tak aplikované. Korpusová lingvistika využívá poznatků z komputační lingvistiky. Vedle vlastního výzkumu se obě tyto disciplíny orientují na shromažďování, zpracování a zpřístupnění hromadných dat a značkování dat za využití některých typů softwarů.[26]

Od prvních začátků v 60.- 80. letech se dnes korpusová lingvistika rozvíjí nejvíce, avšak musí se řešit narůstající problém s masivními datovými rozsahy, které se nedají zpracovávat manuálně. Proto vznikají nové programy k analýze dat. Zásadní se

[24] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 43-44.
[25] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 87-89.
[26] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 90.

stala automatická anotace, tagování textů a související problematika, která se zakládá na manuálním označkování malého výchozího korpusu. Velká část jazykových jevů a forem se typově v malém výchozím korpusu už objevila a dá se zobecnit pro celý velký korpus. Do velkého korpusu se pak užité značkování automaticky přenáší a následně se značkují zbylé části. Je to velmi dlouhodobá činnost, která z důvodu přílivu nových slov do korpusu nikdy nekončí. V korpusové lingvistice není zcela shoda ohledně výhodnosti anotace korpusových textů. Je nesplnitelné mít vedle sebe na výběr několik anotací podle různých teorií. U jiných korpusů takové anotace chybí, zdůrazňuje se autentický prostý text, kde si vlastní výklad vytváří jedinec sám. Pro lingvistickou práci je důležité, zda objektivnost informací získaných z korpusových dat vyšla z dat (velkého) korpusu, nebo jestli se o korpus opírají jen zásadně. Mluvíme proto o dvou pojmech, o přístupu **corpus-driven** a **corpus-based**. [27]

1.4.1 Corpus-based, corpus-driven

Corpus-based (založený na výzkumu korpusu) a **corpus-driven** (řízený výzkumem na korpus) se odlišují v míře vlivu, kterou je ochoten lingvista připustit při formulování hypotézy o jazyce. Rozdíly podrobněji popsala A. Čermáková. V corpus-based přístupu lingvista pracuje s jazykovými daty, která mají předem vytvořenou hypotézou. V korpusu hledá důvody pro potvrzení či vyvrácení. Corpus-driven přístup vytváří koncepty a popisuje struktury v závislosti na zkoumání dat. První přístup je již u nás zcela přijat, druhý přístup je teprve na začátku. V corpus – driven přístupu může lingvista kdykoli vytvořené hypotézy zvrátit a na základě dat pracovat na daném úseku jazykové reality znovu a jinak.[28]

Dalším příkladem odlišnosti v přístupech je rozdělení slov do slovních druhů. V Corpus-based budeme rozdělovat slova do deseti tradičních skupin. Můžeme zkoumat formu, funkci a význam jednotlivých slov, u kterých je klasifikace nejasná. Corpus-driven přístupem může lingvista dojít ke zjištění, že slovnědruhová klasifikace může mít i jinou podobu. Například řadové číslovky jsou blízké adjektivům a mohly by se sloučit do jedné významově homogenní skupiny. Mohla by se tak vytvořit alternativní klasifikace lexémů na úrovni slovních druhů, přičemž kritériem by byla kontextová

[27] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 91.
[28] CVRČEK, Václav a Dominika KOVÁŘÍKOVÁ (eds.). Možnosti a meze korpusové lingvistiky. *Naše řeč – Základní informace* [online]. 2011. roč. 94, č. 3 [cit. 2019-03-10]. Dostupné z: http://nase-rec.ujc.cas.cz/archiv.php?art=8191

vlastnost slov. Klasifikace by vycházela z jazykových dat. Třetí odlišností v corpus-driven může být zkoumání terminologie ohledně přístupu v automatickém vyhledávání termínů v korpusu. Při hledání vlastností termínů můžeme o některých slovech říci, že se jedná o termíny, ale při dalších fázích je výzkum veden korpusovými daty. Následuje formulace definice termínu, která probíhá v mnoha fázích. Označí se termíny v odborných textech a automaticky se vyhledají slova, která sdílejí podobné formální, statistické a lingvistické vlastnosti. Zároveň se vyhodnotí, které z těchto rysů mají na označení slova za termín největší vliv. Terminologickou platnost slova určuje kombinace určitých vlastností.[29]

Nejdůležitější vlastnosti pro vyhledávání termínů v textu jsou:

1. *frekvence slova v obecném korpusu ve srovnání s texty odbornými (pokud je frekvence slova výrazně vyšší v odborných textech než v korpusu složeném z beletrie a publicistiky, je pravděpodobnost, že jde o termín, vyšší);*
2. *distribuce slova v jednotlivých odborných disciplínách (v čím menším počtu disciplín se slovo vyskytuje, tím větší je pravděpodobnost, že půjde o termín);*
3. *struktura slova (čím je struktura slova neobvyklejší, měřeno obvyklostí grafémových bigramů, tím je pravděpodobnější, že se jedná o termín).*[30]

1.4.2 Korpusová metodologie

Korpus můžeme zkoumat několika formálními způsoby a metodami. Na korpusovou metodologii z části působí **povaha dat** a **užitý software** (korpusový manažer). Nepřehledná rozsáhlá korpusová data potřebují výsledky svých zkoumání řadit a odstupňovat. Přístup k datům má každý a využívá se **metoda statistická** (formální povaha), na kterou navazují další, hlavně **analýza kvalitativní**, především **sémantická**. Nejlépe by měla být data předem zjištěná, tj. změřená kvantitativně, až poté se dobrat ke kvalitativním závěrům (sémantickým, funkčním nebo pragmatickým). Pokud je výzkum kvalitativní jen zprostředkovaně, záleží na interpretaci výsledků badatele. Například jaká najde kritéria v datech apod. Je nutné upravovat výsledky o nové poznatky z dalších zdrojů. Povaha dat je základní vlastností analýzy, která je

[29] CVRČEK, Václav a Dominika KOVÁŘÍKOVÁ (eds.). Možnosti a meze korpusové lingvistiky. *Naše řeč – Základní informace* [online]. 2011. roč. 94, č. 3 [cit. 2019-03-10]. Dostupné z: http://nase-rec.ujc.cas.cz/archiv.php?art=8191

[30] CVRČEK, Václav a Dominika KOVÁŘÍKOVÁ (eds.). Možnosti a meze korpusové lingvistiky. *Naše řeč – Základní informace* [online]. 2011. roč. 94, č. 3 [cit. 2019-03-10]. Dostupné z: http://nase-rec.ujc.cas.cz/archiv.php?art=8191

škálovité nečernobílé povahy.[31]

Poznatky z práce s korpusovými daty: *lingvistická informace se do korpusu může vkládat* (většinou automaticky), *díky kombinacím (v kontextu) dá vysoudit mnohem víc než jen informace vložená, byť kvantifikovaná.*[32]

Vysoká frekvence má spojitost s typičností korpusových výstupů. Může tím být materiál pro frekvenční seznamy (pedagogické hledisko – tvorba učebnic, lexikografie) nebo frekvenční slovník, tak i pro studium dichotomie centrum – periferie. Musíme však brát zřetel na nepravidelnosti a anomálie. **Obecně platí**: *Co je typické, je nejvíc ustálené a funkčně základní.* Není zcela vyřešena otázka, například je-li pevná a přímá souvislost mezi vysokou frekvencí a ustáleností formy, i její příslušnost do systému. Je několik nefrekventovaných frazémů, které v systému nejsou. Je potřeba více studovat poznání jazykové variability a jeho variant, protože dříve byla u kodifikovaných jazyků nežádoucí. Korpus je proto vhodným a důležitým zdrojem.[33]

Oporou korpusové informace je kombinace hledané formy (prosté nebo vícečlenné v řetězci). Pro analýzu je východiskem hromadný výskyt hledaných forem v souvislosti v podobě konkordance. Uspořádaná konkordance (většinou v abecedním uspořádání okolních slov) nabízí typické kolokace nebo náhodné shluky slov a tvarů. Pokud v daném kontextu sousedící data tvoří s hledaným slovem běžnou kombinaci, jedná se o kolokace. Posoudit se dá analýza lexikálního okolí z hlediska vyšší abstrakce, nebo také valencí. Možností je i automaticky zjistit n-gramy, což jsou všechny kombinace dvou, tří, čtyř, pěti a více forem (obvykle slov), bigramy, trigramy, tetragramy, pengramy apod. Analyzují se z celého textu. Základem zobecnění je analogie (doložená), a to zobecnění gramatické, sémantické, funkční. Jde o induktivní přístup. Opačným jevem je anomálie. Analogie i anomálie působí vedle i proti sobě v rámci obecného principu vývoje jazyka. Výsledky analýzy korpusu jsou takové, že klasické jazykové příručky, gramatiky apod. mají starý přístup, který je založen na pravidle a výjimce. Korpusová data mají škálovitou povahu na rozdíl od starého přístupu „škatulkování“, který je často nepoužitelný. Škálovitá povaha se nejvíce uplatňuje v sémantických aspektech. Protože korpusová data i výstupní konkordance mají veliký rozsah, dochází k velkému množství výsledků, u kterých není možno je

[31] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 92-93.
[32] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 93.
[33] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 94-95.

mentálně a manuálně zvládnout. Proto se využívá metoda několika opakovaných náhodných vzorců, které zkoumají jev, dokud nejsou nalezeny rozhodující výsledky. Tato metoda se zakládá na matematických metodách a vzorečcích.[34]

Data výsledků z korpusu v podobě konkordance se třídí formálně nebo kvalitativně. Formální klasifikací jsou výsledky uspořádány abecedně nebo retrográdně. Kvalitativní klasifikací vycházejí výsledky z kvantifikovaných dat, z toho pak můžeme vyhodnotit kvalitativní výsledky a výsledky interpretovat. Sémantika lexému byla ve slovnících uváděna v podobě lemmatu, není to totéž jako úhrn chování a významů každého tvaru. Existuje větší množství lexikálních tvarů, vzhledem k sémantice, které je v příručkách zachyceno. V tradičním přístupu se sémantika propojuje s lemmatem. Pokud je málo dat, opakuje se analýza na větších datech. Dojde-li k tomu, že korpus nabízí pouze jeden výsledek k dané formě (popř. několik málo) v omezeném kontextu, nemůžeme dojít k obecnému závěru. Tento jev se nazývá hapax. Podobným jevem jsou monokolokabilní slova, tvary vázané jako komponenty na jeden nebo více slov a kontextů. Platí pouze jen pro systém. Hapaxů v korpusu stále přibývá. Hapaxy jsou v diachronním smyslu zdrojem k vývoji jazyka, v synchronním korpusu jsou zdrojem pro budoucnost. Ukazují potencialitu tvoření forem i významu. Umožňují lingvistům studovat vývoj jazyka z hlediska slovotvorného smyslu. Kompatibilita je základem pro tvoření kolokace (jazykové kombinace), zároveň je důležitá i kompozičnost sémantiky smysluplných kombinací. Existují i výjimky, tj. kombinatorické anomálie, které nemají kompoziční význam. Inkompatibilními jsou sémanticky neslučitelná slova. Pro speciální výzkumy v korpusu si lingvista může nastavit v korpusovém manažeru určité druhy textu, tj. subkorpus. Kvantitativní lingvistika, zvláště kvantitativní přístup, je základem pro kvalitativní analýzu.[35]

1.4.3 Lingvistický výzkum, lexikon a frazeologie

Výzkumný zájem se velmi často zabývá lexikonem včetně frazeologie. Slovníky bývají většinou týmovou prací, zaměřují se na frekvenci, oblast autorských slovníků a další. První vliv na korpus měl frekvenční slovník z redakce Čermáka a Křena. Zakládal se na stomilionovém korpusu. Slovník obsahuje 50 tisíc hesel podle frekvence i abecedy. Mimo se uvádí slovník proprií, zkratek, interpunkce a grafémů. Je jak

[34] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 95-97.
[35] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 97-100.

pomůckou pro další obory studií, tak základem pro vytváření učebnic a slovníků. Tento slovník vychází z psaných dat, proto vznikl Frekvenční slovník mluvené češtiny. Data pro tento slovník vychází z Pražského mluveného korpusu. Lingvistům nabízí frekvenci lemmat i tvarů. Pro studenty vznikl i anglicky psaný malý frekvenční slovník, který obsahuje i mluvený jazyk. Vedle knižních publikací jsou dostupné i ke stažení frekvenční slovníky tvarů i lemmat z korpusů SYN2000, SYN2005 a SYN2010 na stránkách Ústavu Českého národního korpusu, dále retrográdní slovníky nebo Abecední a retrográdní slovníky lemmat a tvarů od M. Křena. Knižně vyšel slovník Základní slovník českých přísloví. Z Českého národního korpusu vychází velké dílo Slovník české frazeologie a idiomatiky. Frazeologií v mluveném jazyce a korpusu se zabývá Kopřivová, problematikou identifikace frazémů na pozadí bigramů se věnuje Čermák.[36]

[36] ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017, s. 114-116.

2. Český národní korpus

2.1 Vznik ČNK

Před vznikem **Českého národního korpusu** (ČNK) a **Ústavu Českého národního korpusu** (ÚČNK) se u nás objevovaly první pokusy o počítačové slovníky ve formě Retrográdního slovníku současné češtiny a Anglicko-českého výkladového slovníku výpočetní techniky. Velkým přínosem pro ČNK byly zkušenosti z práce na několika slovníkových databázích, které vznikly v 80. - 90. letech. Několik autorů vytvořilo slovníkovou část automatických korektorů pro počítačové programy. Výzkumný ústav matematických strojů a Středisko pro výpočetní techniku ČSAV sestavily korektor pro databázi se 100 tisíci českých lexikálních jednotek. Katedra ČJ Filozofické fakulty Masarykovy univerzity v Brně měla databázi o 170 tisících českých kmenech. Ústav teorie informace a automatizace obsahoval 400 tisíc položek českých odborných textů se sémantickým kódem. Z iniciativy Kybernetické společnosti vznikla skupina pro přípravu počítačových korpusů v roce 1988. Cílem byla spolupráce nejen v našem státě, ale i se zahraničím a také sjednocení metodiky lexikografických prací. V roce 1991 se skupina lingvistů a odborníků, zejména z oboru matematiky, rozhodla vytvořit **„Počítačový fond češtiny“**. Jednalo se o databázi textů v elektronické podobě. Na projektu pracovali pracovníci z FF UK Praha, FF MU Brno, MFF UK Praha, ÚJČ a VÚMS Praha. Byla snaha vytvořit fungující projekt, standardizaci kódů záznamu, standardizaci uložení a zpracování, standardizaci editorů pro komentáře, vyřešení právních vztahů a finanční zajištění. To bylo první podobou ČNK. Jedno z prvních jednání vedl **F. Čermák** s vedením Ústavu pro jazyk český AV ČR. Hlavní jednání probíhalo s představiteli FF UK, při kterém také vznikl dne **9. 9. 1994 ÚČNK**. V roce 1996 získal ÚČNK grant *„Počítačový korpus českého jazyka“*. Tím bylo vybudováno pracoviště pro fungování ČNK. Od **1. 10. 1996** měl ČNK své vlastní prostory v budově FF UK.[37]

Projekt ČNK byl vytvořen a zpřístupněn pro veřejnost, aby svou databází jazykových dat pomohl výuce a výzkumu ve formě elektronických korpusů. V současnosti existuje přístup k více než 3 miliardám slov, které jsou uspořádány v synchronních, diachronních, mluvených, psaných, paralelních a jednojazyčných korpusech. ČNK spravují pracovníci dvou ústavů FF UK: Ústav Českého národního korpusu a Ústav teoretické a komputační lingvistiky. Na práci a dalších dílčích

[37] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 45-46.

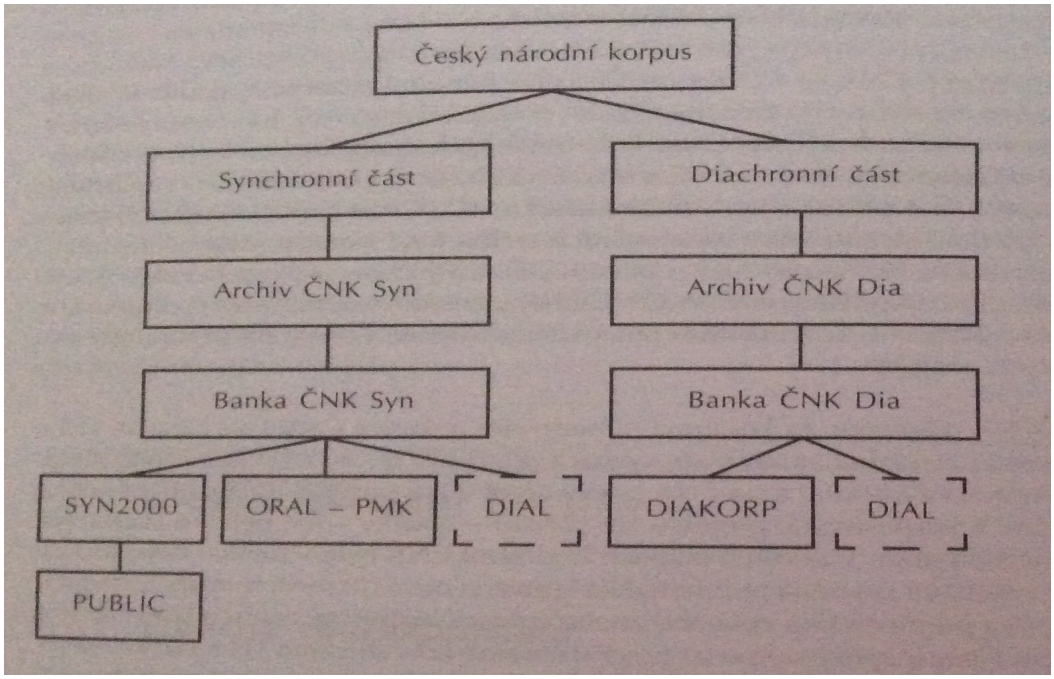
činnostech ČNK se podílí více než 200 externích pracovníků z celé ČR.[38]

Spolupracovníci na práci ÚČNK:

- Ústav teoretické a komputační lingvistiky – FF UK Praha;
- Ústav bohemistických studií – FF UK Praha;
- katedra českého jazyka – FF UK Praha;
- Ústav formální a aplikované lingvistiky – MFF UK Praha;
- Ústav českého jazyka a slovanské jazykovědy – FF MU Brno;
- katedra informačních technologií – FI MU Brno;
- katedra počítačů – Elektrotechnická fakulta ČVUT, Praha;
- Ústav pro jazyk český – AV ČR, Praha.[39]

2.2 Členění ČNK

Obrázek 1 - členění ČNK



KUČERA, Karel (ed.). Český národní korpus: úvod a příručka uživatele. Praha: Filozofická fakulta UK, 2000, s. 12.

Základní členění ČNK, jak je výše znázorněno ve schématu, má pod sebou několik složek v různém vývoji, které celkově tvoří korpus. Složky mají různou povahu i rozsah. Největší zastoupení má *synchronní psaný korpus*, který má rozsah přes

[38] CVRČEK, Václav a Olga RICHTEROVÁ (eds). *Start* [online]. Příručka ČNK: 25.05.2018 [cit. 2019-03-24]. Dostupné na: http://wiki.korpus.cz/doku.php?id=start&rev=1527255653.
[39] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 47.

100 milionů slovních tvarů. Z něho pak vychází korpus (PUBLIC), který je veřejně přístupný. Menší je *diachronní psaný korpus* (DIAKORP), jenž obsahuje okolo 1 750 000 tvarů. *Synchronní mluvený korpus* (ORAL-PMK) má 700 tisíc tvarů. Nářeční korpusy synchronního i diachronního typu (DIAL) mají zatím menší kapacitu. Jednotlivé korpusy vycházejí z celých textů, nikoli ze vzorků. V korpusu jsou zahrnuta díla (tj. dochované části textu), která pochází ze staršího vývoje češtiny. Složkami nekorpusové ČNK jsou **archivy**: Archiv ČNK Syn a ČNK Dia (zde jsou uloženy všechny získané texty ve výchozí podobě) a **banky**: Banka ČNK Syn a Banka ČNK Dia (zde jsou uloženy všechny texty ve formátu SGML).[40]

Synchronní a diachronní korpus

V jazyce nejsou striktní kritéria pro stanovení časového rozmezí jednotlivých druhů korpusů. Časové rozdělení ČNK je závislé na vnějších činitelích (tj. historických). V novinové a časopisecké oblasti bylo mapování textů zařazeno do synchronního psaného korpusu rokem 1990. Starší noviny měly ideologický podtext a nemohou dnes reprezentovat současný jazyk, protože jazyk se v publicistice neustále mění. Téhož roku byl přijat i jako počátek krásné literatury. Z mnoha důvodů (př. literární texty se znovu a znovu přetiskují) se stanovila kritéria, a tím se do synchronního korpusu řadí: čtení autoři narozeni roku 1880 a později; publikované knihy od roku 1945. Rovněž byly zařazeny do synchronního korpusu odborné texty po roce 1989. Všechny texty, které jsou časově za těmito hranicemi, patří do diachronního korpusu, dosud ještě nejsou k dispozici všichni vhodní kandidáti pro zařazení do ČNK. Do synchronního psaného korpusu byly zařazeny texty z časového rozmezí 1998/9, které byly získány s časovým odstupem. Časová etapa 1990-1999 byla prezentována v roce 2000, kdy byla dokončena první etapa ČNK. Práce ČNK pokračovala dále v práci na další etapě a verzi korpusu (rozsáhlejší a úplnější). Diachronní korpus ČNK (DIAKORP) byl vytvořen pro základnu elektronického materiálu. Má pomoci při výzkumu vývoje českého jazyka od prvních dochovaných souvislých záznamů (2. pol. 13. stol.) do poloviny 20. století. V diachronním korpusu jsou dobově a útvarově autentické texty. DIAKORP je projektem, který se obtížně vytváří, jelikož většina textů se musí manuálně přepsat nebo skenovat či korigovat. Texty pocházející z doby před rokem 1849 jsou v diachronním korpuse v transkribované podobě. Důvodem je různorodost paleografických podob textů,

[40] KUČERA, Karel (ed.). *Český národní korpus: úvod a příručka uživatele*. Praha: Filozofická fakulta UK, 2000, s. 12-13.

které pochází z odlišných období vývoje českého pravopisu. V současnosti ČNK neumožňuje připojovat k transkribovaným starším textům jejich transliterovanou podobu. V budoucnu se plánuje užší spojení transkripce s originálem. V roce 2000 DIAKORP obsahoval 1 750 000 textových slov.[41]

Mluvený synchronní korpus

Mluvený synchronní korpus ORAL-PMK (Pražský mluvený korpus) je samostatnou složkou ČNK. Původně sloužil k výzkumu frekvence autentické mluvené češtiny. Rozsah a povaha projektu byla pragmatická. Dnešní rozsah ORAL-PMK činí přes 700 tisíc slov. Vlivem limitujících faktorů je zatím omezen na oblast Prahy a okolí (nejvyšší reprezentativnost). Časové rozpětí mluveného korpusu je z období let 1988-1996. I přes jistá omezení je rozsah záznamu autentické mluvené češtiny největší. Mluvený korpus se skládá z přibližně 300 magnetofonových nahrávek v upraveném přepisu a sleduje reprezentativnost čtyř hlavních sociolingvistických proměn: pohlaví mluvčích, věk, vzdělání a typ textu. Těmito čtyřmi znaky vzniklo několik desítek kombinací. Nahrávky se pak manuálně přepisovaly do elektronické verze a anotovaly.[42]

InterCorp

Projekt InterCorp je paralelním korpusem a je součástí Českého národního korpusu. InterCorp je podporován Ministerstvem školství ČR, jelikož je součástí programu – Projekty velkých infrastruktur pro vědu, výzkum a inovace. Paralelní korpus poskytuje zdroj dat pro teoretické studie, lexikografii, odborné práce, výuku (zejména výuku cizích jazyků), počítačové aplikace, překladatele i veřejnost. Korpus je dostupný pro registrované uživatele. Celý projekt má za cíl budovat paralelní synchronní korpus pro většinu jazyků studovaných na FF, pro daný jazyk a češtinu. Jedná se o akademický a nekomerční projekt.[43]

2.3 Jak korpus vzniká

Získání textu: Základem je získání textů v elektronické podobě. Nejlépe je možné získat materiál u nakladatele nebo vydavatele. ÚČNK tomuto věnuje největší úsilí. Pracovník ústavu osloví poskytovatele textů, vysvětlí projekt a jeho důležitost. Musí se dodržovat stanovená „Dohoda o spolupráci“, získané texty se nesmí poskytovat

[41] KUČERA, Karel (ed.). *Český národní korpus: úvod a příručka uživatele*. Praha: Filozofická fakulta UK, 2000, s. 13-15.
[42] KUČERA, Karel (ed.). *Český národní korpus: úvod a příručka uživatele*. Praha: Filozofická fakulta UK, 2000, s. 15-16.
[43] *Ústav Českého národního korpusu* [online]. *Intercorp.* Copyright © FF UK 2015 [cit.2019-04-03]. Dostupné z: http://ucnk.korpus.cz/intercorp/

třetímu subjektu. Další možností, jak text získat, je skenování. Je zapotřebí určitého hardwaru a softwaru. Časově je to náročnější. Nejméně se používá přepis písařkou. Tento způsob je nejnáročnější a finančně nákladný. Využívá se u příliš malých textů (např. lístky MHD, lékařské předpisy, volební letáky atd.).[44]

Počáteční evidence: Veškeré materiály, které se dostanou do archivu ČNK, musí projít počáteční evidencí. Evidenční databázi připravili pracovníci ČNK, do které se při práci s textem zaznamenávají relevantní informace (tj. typ textu a zpracování). První záznam je zpracován odpovědným pracovníkem za celou databázi. Provádí se úkony jako typování editoru, kódování a záznam o původní organizaci, odkud je získaný text. Následně je materiál rozepsán po jednotlivých souborech. Zkopírovaná data se uloží jako jedna verze do archivu ČNK. Další verze projde konverzí do mezi-formátu a k vnějšímu lingvistickému značení.[45]

Přechod surových textů do archivu ČNK: Původní texty od poskytovatelů se uloží do archivu ČNK, kde jsou uchovány na záložních zdrojích. Kopie textů jsou konvertovány a zpracovány do budoucího korpusu tak, aby přicházely texty větné podoby. Pomůže to lingvistům, kteří by se zabývali např. obsahem knih, rejstříky, tirážemi.[46]

Mezi-konverze – mezi-formát: Původní získané texty je třeba přeformátovat do jednotného formátu, tj. odejmout z textu obrázky, tabulky, speciální formátové značky. Tato první etapa se nazývá mezi-formát. Některé formátovací prostředky jsou uchovány (např. rozlišení typu písma a označení odstavců). Všechny značky vložené do mezi-formátu jsou v konečné fázi ve formátu SGML. Mezi-konverze není jen pro konečnou podobu textu, ale má pomoci lingvistům při práci s materiálem.[47]

Vnější lingvistická anotace: Text před vstupem do korpusu musí projít poslední konverzí, přiřadí se mu tzv. *„hlavička“* (header). Jedná se o pár nejdůležitějších informací o textu. Hlavní funkcí hlavičky je identifikace textu a poskytnutí lingvistovi při práci s konkordancí základních informací o zdroji (textový typ, žánr, rok vydání). Hlavičky textů vznikají při vnější lingvistické anotaci, lingvista projde text v mezi-formátu a informace z tiráže i z textu přenese do své varianty evidenční databáze. **Jedná se o položky:** *autor, název díla, překladatel, místo vydání.* Pro odlišení jazyka např. beletristického díla od jazyka jeho předmluvy, jsou texty značkovány zvlášť.

[44] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 52.
[45] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 52-53.
[46] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 53.
[47] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 53.

V tomto případě je potřeba, aby byly uchovány informace, že texty jsou z jedné knihy, a proto se vyplňují pro předmluvu položky: *autor celku, název celku*, ve kterém je dílo publikováno. **Další položky:** *korpus, text veršovaný/ neveršovaný, typ textu, odborná oblast, žánr, médium přenosu, pohlaví autora, jazyk díla, pohlaví překladatele, původní jazyk díla, rok vydání, rok prvního vydání, opus (jednoznačná identifikační značka).* Informace jsou označeny zkratkami, nebo čísly. Pokud některé informace nemůžeme zjistit, můžeme je u některých položek vynechat. Některé položky se musí povinně vyplnit, jelikož jejich zkratka je pak součástí hlavičky (např. korpus, typ textu, žánr, rok vydání, opus).[48]

Databáze evidence: Před vložením informace do hlavní evidenční databáze je nutno ji zkontrolovat. Pracovník ÚČNK projde položky a překopíruje je do hlavní databáze evidence.[49]

Konverze do formátu SGML a čištění: Lingvistický software, který pracuje s korpusy, vyžaduje jednotný formát a to SGML (Standard Generalized Markup Language = Standardní generalizovaný značkovací jazyk). Jedná se o mezinárodní standard pro popis a značkování elektronických textů. Každý dokument SGML nese řadu informací, tj. „SGML" prolog, „deklarace DTD" (definice typu dokumentu), přidané informace v textu a jejich funkce. Značky jsou v textu ve špičatých závorkách a jsou ve dvojici. Není třeba se více zabývat významem značek, protože vyhledávací program se SGML pracuje, ale korpusový manažer většinou generické identifikátory nezobrazuje. Čištění nastává po konverzi textů. Během čištění jsou odstraněny duplikáty a cizojazyčný text.[50]

Zařazení do textové banky ČNK: Přenesený text v SGML je uložen do banky ČNK. Takto velký soubor veškerých technicky a lingvisticky zpracovaných dat se převádí do vnitřního formátu vyhledávacích programů, tím vzniká ČNK. **ČNK (reprezentativní pohled na jazyk):** Byly provedeny výzkumy a statistiky z různých pohledů, zabývaly se vztahem čtenář – čtený text. Výzkum ÚČNK se zaměřil na recepci, výzkum zkoumal problematiku nejčtenějších typů beletrie i odborné literatury. Výsledky se staly pravidlem pro sestavování prvního reprezentativního korpusu českého jazyka. Reprezentativnost se bude odvíjet jak od nových výzkumů čtenosti, tak i od závislosti na pracovních zkušeností s prvním reprezentativně sestaveným korpusem češtiny. Postupně budou vznikat i další modifikované korpusy, jak se bude

[48] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 54-55.
[49] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 58.
[50] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 58-60.

měnit pohled na reprezentativnost.[51]

2.4 Reprezentativnost ČNK

Vymezení textových zdrojů korpusu s kvantitativním pohledem na jejich strukturu se nachází ve vzájemné spolupráci. Ideální textový korpus podle matematické statistiky obsahuje neuskutečnitelný počet textů, které bychom měli k dispozici. Reálný korpus nazývají statistikové *vzorkem, výběrem, sample*. Reprezentativnost nenarůstá lineárně se zvětšováním rozsahu, ale záleží na tom, co zkoumáme, na pravděpodobnostních charakteristikách výskytů prvků. V případě korpusu jde však o prvky, kdy je všestrannost reprezentativnosti vyloučena. Na reprezentativnosti korpusu můžeme zkoumat jeho možnosti využití. Oficiálně není univerzálně uznávána koncepce o využití korpusů, proto byl ČNK vytvořen pro zájemce o práci s korpusem. Z průzkumů bylo vyhodnoceno několik stupňů poznání. Výchozím bodem první úrovně se stal výsledek o poměru *čtení knih, časopisů a novin* českou populací. Zastoupení novin v SYN2000 získalo 60 %. Druhou úrovní členění zdrojů ČNK (nenovinové texty na imaginativní a informativní) bylo porovnávání zájmu o literaturu naučnou a literaturu krásnou u různých skupin čtenářů. Výsledek ukázal zastoupení naučené literatury (informativních, nenovinových textů) v hodnotě 25 %, krásné literatury (imaginativních textů) pak bylo 15 %. Třetí úroveň zkoumala zájem o naučnou literaturu, strukturu zájmu o literaturu ze strany čtenářů nových knih a strukturu katalogu domácích periodik.[52]

IMAGINATIVNÍ TEXTY	**15,00%**
Krásná literatura	**15,00%**
poezie	0,81%
drama	0,21%
próza	11,02%
jiné imaginativní texty	0,36%
přechodové pásmo	2,60%

KUČERA, Karel (ed.). Český národní korpus: úvod a příručka uživatele. Praha: Filozofická fakulta UK, 2000, s. 18.

[51] ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999, s. 60-61.
[52] KUČERA, Karel (ed.). *Český národní korpus: úvod a příručka uživatele*. Praha: Filozofická fakulta UK, 2000, s. 16-17

INFORMATIVNÍ TEXTY	**85,00%**
Publicistika	**60,00%**
Odborné texty	**25,00%**
vědy o umění	3,48%
sociální vědy	3,67%
právo a bezpečnost	0,82%
přírodní vědy	3,37%
technika	4,61%
ekonomie a řízení	2,27%
víra, náboženství	0,74%
životní styl	5,55%
administrativa	0,49%

KUČERA, Karel (ed.). Český národní korpus: úvod a příručka uživatele. Praha: Filozofická fakulta UK, 2000, s. 18.

2.5 Korpusový manažer

Korpusový manažer je speciálním souborem programů, jenž umožňuje využití korpusu. V rámci Ústavu Českého národního korpusu můžeme pracovat s verzí **NoSketch Engine** nebo s **KonText**. Uživatel zvládne práci s programy bez speciálního školení. Pokud ne, pak jsou dostupné příručky jak tištěné, tak i elektronické.[53] Korpusový manažer efektivně vyhledává v korpusových datech. Kromě konkrétních konkordancí má i další funkce (např. vyhledávání kolokací, vyhodnocení základní frekvence, statistiky a složitější analýzy). Práce manažeru je lokální, nebo je založena na principu *klient-server*. Při lokální práci musí být na počítači celý textový korpus. Taková práce se většinou využívá v začátcích s korpusovou lingvistikou. V druhém principu se na počítači pracuje pouze s klientskou částí, jež nabízí určité uživatelské rozhraní a hledání probíhá na vzdáleném serveru, kde se korpus nachází. Tento princip je vhodný pro velké korpusy. Aktuálními serverovými aplikacemi jsou **Corpus Workbench** (CWB) s korpusovým manažerem CQP (Corpus Query Processor) a **Manatee**, což je modernější korpusový manažer, který byl vytvořen Pavlem Rychlým. Ke korpusovým manažerům patří uživatelské rozhraní. Pro příklad lze uvést: Bonito, Sketch Engine, Word Sketches, NoSketch Engine, Park, KonText atd.[54]

[53] OSOLSOBĚ, Klára. *Česká morfologie a korpusy*. Praha: Karolinum, 2014, s. 8.
[54] CVRČEK, Václav a Olga RICHTEROVÁ (eds). *"pojmy: korpusovy_manazer"*. [online]. Příručka ČNK: 03.08.2016 [cit. 2019-04-03].
Dostupné z: http://www.wiki.korpus.cz/doku.php/pojmy:korpusovy_manazer

3. Práce s korpusovou databází

3.1 První seznámení s ČNK

Pro plnohodnotnou práci s ČNK je nutná registrace na www.korpus.cz. Tato registrace je bezplatná. Bez ní není možné využívat všechny nástroje korpusu. Mezi veřejně přístupné aplikace korpusu, které jsou k dispozici patří **KonText, SyD, Morfio, Kwords, Treq.** Pro registrované uživatele, kteří odsouhlasili podmínky užívání korpusu, je k dispozici širší nabídka korpusů, poradna ČNK a několik dalších speciálních funkcí, jež nabízí aplikace KonText. Pro uživatele a výzkumné instituce, kterým není přístup ke korpusům v určitém rozhraní dostačující, může ČNK připravit jazyková data „na míru“. [55]

3.2 Sedm základních lekcí

Veškeré základní lekce práce s korpusem v rozhraní KonText jsou k dispozici na internetu, kde je vše v rámci sedmi lekcí podrobně popsáno a vysvětleno i s obrázky. Součástí manuálu jsou i bonusové lekce – hledání v paralelním korpusu, mluvených korpusech, diachronních korpusech, syntakticky anotovaném korpusu. Knižně vydané jsou i příručky pro uživatele korpusu.[56]

1. úvodní lekce – zadávání dotazů do korpusu SYN2015, v čem se liší dané typy a citace korpusu. **Typy dotazů:** V KonTextu je celková nabídka 6 typů dotazů (základní, lemma, fráze, slovní tvar, část slova, CQL). Každý typ se hodí na jiný druh dotazu, záleží na zkoumaném objektu. Ne ve všech korpusech je k dispozici každý typ dotazu. Typ dotazu – **slovní tvar (word)** je nejjednodušší. S tímto typem dotazu najdeme v korpusu přesnou shodu, tj. tvar, tak jak byl zadán. Jedinou odlišností může být velikost písma. V nastavení můžeme zadat velikost písma za pomocí *„Shoda velikosti písmen“* pod dotazovacím řádkem. Typ **lemma** je označení pro základní tvar výrazu. Nejčastěji lemma hledáme ve slovníku *(např. slovní tvar chytrého hledáme ve slovníku tvar chytrý)*. Lemmatem u substantiv bývá 1. pád jednotného čísla. Další typ dotazu – **základní** napomáhá při rychlém informativním hledání v situacích, kdy není třeba příliš velká přesnost. K zadanému slovnímu tvaru (lemma) se vygenerují všechny jeho tvary. Pokud nejde o lemma, jsou nalezeny jen tvary, které se s dotazem shodují.

[55] CVRČEK, Václav a Olga RICHTEROVÁ (eds). kurz: zaciname [online]. Příručka ČNK; 2018 [cit. 2019-05-19]. Dostupné na: http://wiki.korpus.cz/doku.php?id=kurz:zaciname&rev=1533556421
[56] CVRČEK, Václav a Olga RICHTEROVÁ (eds). kurz:uvod [online]. Příručka ČNK; 2018 [2019-06-15]. Dostupné na: https://wiki.korpus.cz/doku.php?id=kurz:uvod&rev=1534164412

K vyhledávání víceslovných výrazů je vhodný dotaz – **fráze**, který je podobný základnímu dotazu, ale je v nich rozdíl. V typů dotazech slovní tvar a lemma není možno zadat více slov najednou. Při hledání všech slov, které mají řetězec znaků, se využívá typ dotazu – **část slova**. Tento dotaz nám vyhledá danou sekvenci znaků (předcházející nebo následující libovolné znaky, ale i třeba žádný). Poslední typ – **CQL** patří mezi nejobecnější způsoby hledání a nabízí mnoho možností. Všechny předchozí dotazy jsou v KonTextu převáděny na CQL.[57]

2. lekce – vše, co se dá u dotazu zobrazit, jaká data jsou k dispozici, konkordantní řádek, vnitřní struktura dat, zobrazení dostupných metainformací, ukládání, export dat. Základní informace můžeme získat z prvního zobrazení konkordance. Jednak zvýrazněný **KWICK** (klíčové slovo v kontextu) a název textu, tak i informace o četnosti slova v liště nad konkordancí. Nejprve je k dispozici **frekvence absolutní**, jež udává počet výskytů slova v rámci jednoho korpusu. Díky **relativní frekvenci** (i.p.m.) máme možnost srovnávat frekvenci slov v různých korpusech. Z průměrné redukované frekvence (ARF) získáváme data o rovnoměrnosti rozložení slova v korpusu.[58]

3. lekce – vyhodnocení dotazu, frekvence a konkordance, třídění řádků. Třídit řádky je užitečné v případě, že celá konkordance není natolik rozsáhlá. Vhodnější je práce s náhodnými vzorky. Náhodný vzorek lze nastavit v menu Konkordance. Vzorek pak můžeme nechat abecedně seřadit.[59]

4. lekce – využití regulárních výrazů při dotazování, kvantifikátory, sekvence libovolných znaků, speciální symboly. Za pomoci **regulárních výrazů** (tj. sekvence znaků) můžeme vyhledat množinu slov. Regulární výrazy pracují se znaky se speciálním významem a běžnými znaky, tj. znaky abecedy, číslovky apod.[60]

5. lekce – dotazovací jazyk CQL, kombinace dotazů, posloupnost pozic. Dotazovací jazyk **CQL** je formálním jazykem korpusu. Dotaz musíme formulovat v CQL podle určitého typu dotazu. S CQL pracujeme, když zadáváme složitý dotaz

[57] CVRČEK, Václav a Olga RICHTEROVÁ (eds). kurz:prvni_dotaz [online]. Příručka ČNK; 2018 [cit. 2019-06-15]. Dostupné na: https://wiki.korpus.cz/doku.php?id=kurz:prvni_dotaz&rev=1552483926
[58] CVRČEK, Václav a Olga RICHTEROVÁ (eds). kurz:zobrazeni_dotazu [online]. Příručka ČNK; 2018 [cit. 2019-06-15]. Dostupné na: https://wiki.korpus.cz/doku.php?id=kurz:zobrazeni_dotazu&rev=1534344877
[59] CVRČEK, Václav a Olga RICHTEROVÁ (eds). kurz:vyhodnoceni_dotazu [online]. Příručka ČNK; 2018 [cit. 2019-06-15]. Dostupné na: https://wiki.korpus.cz/doku.php?id=kurz:vyhodnoceni_dotazu&rev=1535534150
[60] CVRČEK, Václav a Olga RICHTEROVÁ (eds). kurz:regularni_vyrazy [online]. Příručka ČNK; 2018 [cit. 2019-06-15]. Dostupné na: https://wiki.korpus.cz/doku.php?id=kurz:regularni_vyrazy&rev=1533716282

(např. najdi výskyt lemmatu *oko*). Nabízí proto mnoho možností (např. kombinace více atributů za pomoci logických operátorů).[61]

6. lekce – pokročilé dotazy, kolokace, filtrování konkordance, asociační míry. Na pojem *kolokace* není obecná shoda, názory se neshodují v tom, co do kolokace všechno patří. Rozlišuje se kolokace v užším slova smyslu, také běžné kolokace, a kolokace v širším slova smyslu: běžné kolokace *(letní šaty, vejce naměkko)*, frazémy a idiomy *(ležet ladem, růžové brýle)*, víceslovné termíny *(infarkt myokardu, červí díra)*, víceslovná vlastní jména *(Andělská Hora, Kostelec nad Černými Lesy).*[62]

7. lekce – vytváření vlastních subkorpusů a podmínek, veřejné subkorpusy. Pokud si vytvoříme trvalý vlastní subkorpus, je nám k dispozici opakovaně. Zůstane v našem uživatelském účtu. Subkorpus se tvoří v menu Korpusy v položce Vytvořit nový subkorpus.[63]

3.3 Rozhraní KonText

Uživateli webová aplikace KonText napomáhá k přístupu a práci v korpusech ČNK. První dvě položky hlavního menu umožňují zadávat dotazy, organizovat korpusy a vytvářet subkorpusy. Další položky menu (s výjimkou nápovědy) jsou založeny na kontextu – pracují s vybraným korpusem, dodatečně upravují nebo vyhodnocují dotaz. V menu KonTextu se nachází tyto položky: dotaz, korpusy, uložit, konkordance, filtr, frekvence, kolokace, zobrazení a nápověda.[64]

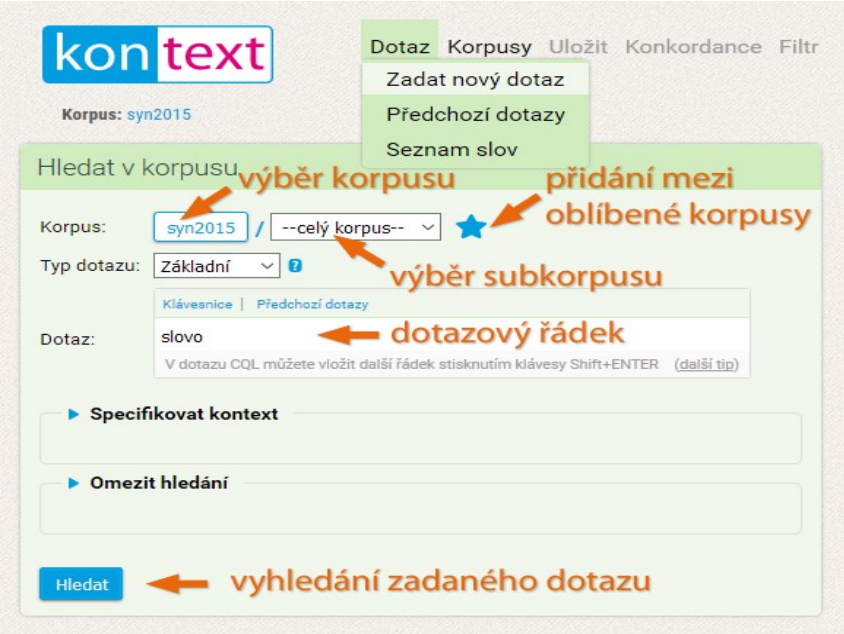
[61] CVRČEK, Václav a Olga RICHTEROVÁ (eds). kurz:pokrocile_dotazy [online]. Příručka ČNK; 2018 [cit. 2019-06-15]. Dostupné na: https://wiki.korpus.cz/doku.php?id=kurz:pokrocile_dotazy&rev=1535546576

[62] CVRČEK, Václav a Olga RICHTEROVÁ (eds). kurz:hledani_kolokaci [online]. Příručka ČNK; 2018 [cit. 2019-06-15]. Dostupné na: https://wiki.korpus.cz/doku.php?id=kurz:hledani_kolokaci&rev=1544785461

[63] CVRČEK, Václav a Olga RICHTEROVÁ (eds). kurz:subkorpusy [online]. Příručka ČNK; 2019 [cit. 2019-06-15]. Dostupné na: https://wiki.korpus.cz/doku.php?id=kurz:subkorpusy&rev=1558522243

[64] CVRČEK, Václav a Olga RICHTEROVÁ (eds). manualy:kontext:index [online]. Příručka ČNK; 2018 [cit. 2019-06-19]. Dostupné na: http://wiki.korpus.cz/doku.php?id=manualy:kontext:index&rev=1541457083

Obrázek 2 – zadání nového dotazu do rozhraní KonText



CVRČEK, Václav a Olga RICHTEROVÁ (eds). Zadání nového dotazu do rozhraní KonText [foto]. In: kurz:prvni_dotaz [online]. Příručka ČNK; 2018 [cit. 2019-06-19]. Dostupné na: http://wiki.korpus.cz/doku.php?id=kurz:prvni_dotaz&rev=1552483926

4. František Čermák a spolupracovníci ČNK

František Čermák, prof. PhDr., DrSc. (1940)

V roce 1962 vystudoval na Filozofické fakultě Univerzity Karlovy v Praze češtinu, angličtinu a holandštinu. V roce 1991 se stal docentem obecné jazykovědy a roku 1994 profesorem českého jazyka. V letech 1991-93 pracoval v akademickém Ústavu pro jazyk český v Praze jako externí vedoucí lexikografického oddělení. Několik let byl ředitelem Ústavu českého národního korpusu na filozofické fakultě v Praze. Ve své vědecké činnosti se zabývá obecnou lingvistikou, jazykovědnou teorií, bohemistikou, a to zejména lexikologií a lexikografií. Je autorem dvojjazyčných slovníků, učebnic, odborných prací a překladů lingvistických příruček. Za svou vědeckou práci byl vyznamenán Akademií věd a Karlovou univerzitou. V roce 2003 získal Cenu za vědu od ministryně školství.[65]

Je také členem několika lingvistických sdružení (Jazykovědné sdružení, Pražský lingvistický kroužek, Societas linguistica Europea aj.) a členem redakční rady International Journal of Corpus Linguistics. Mezi některá jeho díla patří *Základní učebnice češtiny I-V*, *Česká lexikologie*, *Slovník české frazeologie a idiomatiky*, *Jazyk ajazykověda*, *Český národní korpus: 100miliónový SYN 2000*, *Základy lingvistické metodologie*, *Stručná mluvnice češtiny pro zahraniční studenty* a mnoho dalších. [66]

Na akademickém projektu ČNK pracuje Ústav Českého národního korpusu na Filozofické fakultě Univerzity Karlovy v Praze. Od roku 1994 ÚČNK buduje ČNK, který se neustále rozvíjí zejména v oblasti výuky a oboru korpusové lingvistiky.[67]

Ústav Českého národního korpusu:

ředitel: Mgr. Michal Křen, Ph.D.

zástupce ředitele: doc. Mgr. Václav Cvrček, Ph.D.

čestná afiliace: prof. PhDr. František Čermák, DrSc.

profesor: prof. PhDr. Karel Kučera, CSc.[68]

---

[65] VEČERKA, Radoslav. *Biografickobibliografické medaloinky českých lingvistů: bohemistů a slavistů (2. vydání)* [online]. Linguistica online. Brno: electronic journal of the Department of Linguistics and Baltic Languages at Masaryk University, Czech Republic, 2008, s. 22 [cit. 2019-03-22]. ISSN 1801-5336. Dostupné z: http://www.phil.muni.cz/linguistica/art/vecerka/vec-medailonky2.pdf

[66] TŘEŠTÍK, Michael (ed.). *Kdo je kdo: osobnosti české současnosti: 5000 životopisů*. Praha: Agentura, Kdo je kdo, 2002, s. 84.

[67] *Ústav Českého národního korpusu.* [online]. *Co je korpus?* Copyright © FF UK 2015 [cit. 2019-04-03]. Dostupné z: http://ucnk.korpus.cz/co_je_korpus.php

[68] *Lidé | Ústav Českého národního korpusu.* [online]. Copyright © FF UK 2015 [cit. 2019-03-23]. Dostupné z: https://ucnk.ff.cuni.cz/cs/ustav/lide/

Lingvistická sekce (vědecká, publikační a pedagogická činnost):

doc. Mgr. Václav Cvrček, Ph.D. (vedoucí)

Mgr. Anna Čermáková, Ph.D. (Marie Curie Fellowship)

Mgr. Dominika Kováříková, M.A., Ph.D.

Mgr. Lucie Lukešová, Ph.D.

PhDr. Jiří Milička, Ph.D.

Mgr. Michal Škrabal, Ph.D.[69]

Komputační sekce (zpracování dat, vývoj software, správa IT, DTP):

Mgr. Pavel Vondřička, Ph.D. (vedoucí)

Tomáš Čapka

Tomáš Jeziorský

Mgr. Jan Kocek

Ing. Tomáš Machálek

Mgr. Jakub Pejcha

Pavel Procházka[70]

Sekce mluvených korpusů (koordinace sběru dat a anotace pro mluvené a nářeční korpus):

PhDr. Marie Kopřivová, Ph.D. (vedoucí)

Mgr. Lucie Benešová

Mgr. Hana Goláňová, Ph.D.

Mgr. Petra Poukarová

Mgr. et Mgr. Zuzana Komrsková

Mgr. David Lukeš

Mgr. Martina Waclawičová[71]

[69] *Lidé | Ústav Českého národního korpusu.* [online]. Copyright © FF UK 2015 [cit. 2019-03-23]. Dostupné z: https://ucnk.ff.cuni.cz/cs/ustav/lide/

[70] *Lidé | Ústav Českého národního korpusu.* [online]. Copyright © FF UK 2015 [cit. 2019-03-23]. Dostupné z: https://ucnk.ff.cuni.cz/cs/ustav/lide/

[71] *Lidé | Ústav Českého národního korpusu.* [online]. Copyright © FF UK 2015 [cit. 2019-03-23]. Dostupné z: https://ucnk.ff.cuni.cz/cs/ustav/lide/

Sekce diachronních korpusů (koordinace sběru dat a anotace pro diachronní korpusy):

Mgr. Martin Stluka, Ph.D. (vedoucí)

Mgr. Kateřina Najbrtová, Ph.D.

Mgr. Klára Pivoňková

Mgr. Anna Řehořková[72]

Sekce lingvistické analýzy a anotace (morfologická a syntaktická anotace):

Mgr. Tomáš Jelínek, Ph.D. (vedoucí, ÚTKL)

RNDr. Milena Hnátková, CSc. (ÚTKL)

doc. RNDr. Vladimír Petkevič, CSc. (ÚTKL)

RNDr. Hana Skoumalová, Ph.D. (ÚTKL)[73]

Sekce paralelních korpusů (koordinace sběru dat a anotace pro paralelní korpus InterCorp):

Ing. Alexandr Rosen, Ph.D. (vedoucí, ÚTKL)

PhDr. Michala Adamová

Bc. Martin Vavřín

Mgr. Adrian Jan Zasina [74]

[72] *Lidé | Ústav Českého národního korpusu.* [online]. Copyright © FF UK 2015 [cit. 2019-03-23]. Dostupné z: https://ucnk.ff.cuni.cz/cs/ustav/lide/

[73] *Lidé | Ústav Českého národního korpusu.* [online]. Copyright © FF UK 2015 [cit. 2019-03-23]. Dostupné z: https://ucnk.ff.cuni.cz/cs/ustav/lide/

[74] *Lidé | Ústav Českého národního korpusu.* [online]. Copyright © FF UK 2015 [cit. 2019-03-23]. Dostupné z: https://ucnk.ff.cuni.cz/cs/ustav/lide/

5. Základní problematika frazeologie

5.1 Obecná teorie

Frazém je ustálený pojem, jenž je víceslovný, obrazný a často expresivní s omezenou spojovatelností. Slovník české frazeologie a idiomatiky dělí frazémy na přirovnání, frazémy nevětné a větné. Obsahově sémantická stránka frazému se nazývá **idiom**.[75] Typickým znakem frazému, ale ne zcela nutným, bývá **metaforičnost** a **expresivnost** *(udělat díru do světa, být v balíku, lézt někomu krkem)*. Ve frazémech se často vyskytují lexikální a gramatické **archaismy** *(kout pikle, nemá kouska rozumu)*. Monokolokabilní prvky jsou takové, které se nevyskytují v jazyce, ale pouze v konkrétním frazému.[76]

Frazémy a idiomy se z hlediska své formy liší i svou funkcí (jsou součástí věty, větou nebo kombinací věd apod.).[77]

Lexikografický zájem o idiomatiku a frazeologii není příliš nový, vždy se ve slovnících vyskytovaly některé frazeologické i idiomatické výrazy. Byly uváděny nedůsledně k jednotlivým významům určitých slov. Mělo to řadu nevýhod jako například rozlišení jeho daného významu i celého výrazu. První nedokonalé pokusy sbírek jsou od J.M. Sychry, J. Šacha i F. Šebka. Obsahově i zpracováním měly jejich práce s dnešním pojetím frazeologie a idiomatiky jen velmi málo společného. Ceněná práce je až od J. Zaorálka *Lidová rčení*. Poučné mravní a jiné zásady, postřehy a myšlenky o životě bychom našli v antických tradicích. Přísloví se stalo nástrojem vyjádření pro národní svébytnost jako u Komenského *Moudrost starých Čechů*, Dobrovského *Českých příslovích sbírka* a Čelakovského *Mudrosloví národu slovanského v přísloví*ch. [78]

5.2 Frazémy nevětné

Frazémy nevětné se připojují do věty až v konkrétním kontextu a jsou v nich gramaticky formovány *(lev salónů – Stal se lvem salónů)*. Nevětných frazémů je velké množství a volnost slovních spojení není ohraničena *(získat motiv, mít motiv, ztratit*

[75] ČECHOVÁ, Marie. *Čeština – řeč a jazyk*. 2., přeprac. vyd. Praha: ISV, 2000, s. 66.
[76] KARLÍK, Petr et al. *Příruční mluvnice češtiny*. Vyd. 2., opr. [i.e. 4. vyd.]. Praha: NLN, Nakladatelství Lidové noviny, 2012, s. 71.
[77] ČERMÁK, František a Jiří HRONEK (eds.). *Slovník české frazeologie a idiomatiky*. Praha: Academia, 1983, s. 10.
[78] ČERMÁK, František. *Idiomatika a frazeologie češtiny*. Praha: Univerzita Karlova, 1982, s. 11-12.

motiv).[79]

Nejčastější strukturní typy nevětných frazémů:

1. frazémy s funkcí slovesnou obsahující sloveso: *měnit barvu, dělat kličky, ukápnout jedu, praštit se přes kapsu, dopadnout bledě;*
2. frazémy s funkcí jména v nominativu, které vznikly kombinací adjektiva a substantiva: *jidášský groš, šedá eminence, růžové brýle;*
3. frazémy s funkcí jména v nominativu vzniklé kombinací substantiv: *hodina duchů, lev salónů, zkouška ohněm;*
4. frazémy s funkcí adverbiální, jež vznikly kombinací substantiva a adjektiva, případně několika substantiv v jiných pádech než v nominativu: *levou rukou, živou mocí, s klidem Angličana, na dosah ruky;*
5. frazémy z ne autosémantických komponentů: *podle mého, pro nic za nic, jako takový, bodejť by neljo, s takovou.*[80]

Ustálená přirovnání

Ustálená přirovnání bývají řazena k nevětným frazémům. Naznačují větnou strukturu, ale některé pozice v ní jsou teprve až v konkrétních kontextech *(spát jako na vodě – někdo spí jako na vodě).* Za ustálená přirovnání v nevětném frazému se považuje, když jejich součást tvoří větu a v písmu se pak neoddělují čárkou *(rozprchli se (,) jako když do vrabců střelí)*. Podoba přirovnání je stabilní. K charakteristice skutečnosti se užívá ustáleného přirovnání, hlavní funkcí bývá **intenzifikace**, **zdůraznění** a kladné či záporné **zhodnocení** znaku.[81]

5.3 Frazémy větné

Frazémy mají také podobu věty buď slovesné *(Ranní ptáče dál doskáče)*, nebo neslovesné *(Všude dobře, doma nejlíp),* popřípadě souvětí *(Čiň čertu dobře, peklem se ti odmění)*. V kontextu působí jako celek.[82]

Tradiční vymezení podle Příruční mluvnice češtiny:

- **pořekadla** – *anonymní výroky vystihující určitou situaci (I mistr tesař se utne);*

[79] KARLÍK, Petr et al. *Příruční mluvnice češtiny.* Vyd. 2., opr. [i.e. 4. vyd.]. Praha: NLN, Nakladatelství Lidové noviny, 2012, s. 71.
[80] KARLÍK, Petr et al. *Příruční mluvnice češtiny.* Vyd. 2., opr. [i.e. 4. vyd.]. Praha: NLN, Nakladatelství Lidové noviny, 2012, s. 71.
[81] KARLÍK, Petr et al. *Příruční mluvnice češtiny.* Vyd. 2., opr. [i.e. 4. vyd.]. Praha: NLN, Nakladatelství Lidové noviny, 2012, s. 72.
[82] KARLÍK, Petr et al. *Příruční mluvnice češtiny.* Vyd. 2., opr. [i.e. 4. vyd.]. Praha: NLN, Nakladatelství Lidové noviny, 2012, s. 72.

- **pranostiky –** *specifický druh pořekadel vyjadřující vztah mezi časovým obdobím roku a atmosférickými jevy, případně zemědělskými pracemi (Vánoce na blátě, Velikonoce na ledu);*
- **přísloví –** *anonymní výroky podávající mravní poučení poukazem na kolektivní zkušenost (S poctivostí nejdál dojdeš);*
- **okřídlená slova –** *(Nikdo není prorokem ve své vlasti – evangelium sv. Lukáše, Veni, vidi, vici – výrok přiřčený Caesarovi);*
- **dialogické** – *vztahují se minimálně ke dvěma komunikantům (Pozdrav pánbu! Dejž to pánbu! Na zdraví! Ať slouží!).*[83]

Vymezení frazémů podle Čeština, řeč a jazyk:

- **rčení** (úsloví) – *domácí drak (o ženě), zaječí úmysly, má za ušima, má pod čepicí, poradí se i s medvědem, pálí mu to, je to hlava otevřená, ...*
- **přirovnání –** *se srovnávacím výrazem jako: je chytrý jako liška, je jako dubnové počasí (nestálý), je jako slon v porcelánu (nešikovný), točí se jako korouhvička; ustálená přirovnání se srovnávacím výrazem než: udělá to, než bys řekl švec (rychle, hned);*
- **pořekadla –** *charakter větný, shrnují často zkušenost lidí, hodnotící a imperativní: Kam vítr, tam plášť, Panská láska po zajících skáče (je nestálá), shnují i zkušenost protikladnou: Spěchej pomalu, Ráno moudřejší večera x Co můžeš udělat dnes, neodkládej na zítřek;*
- **pranostiky –** *výroky zabývající se roční dobou, počasí a závislosti prací v zemědělství: Svatá Lucie noci upije a dne nepřidá; Úhor bílý pole sílí; Na svatého Řehoře šelma sedlák, který neoře;*
- **přísloví –** *ustálené obrazné výpovědi dvousložkové (první část přísloví něco konstatuje, druhá část z toho vyvozuje poučení): Kdo jinému jámu kopá, sám do ní padá; Tak dlouho se chodí se džbánem pro vodu, až se ucho utrhne; některá jsou významově protikladná: Mluviti stříbro-mlčeti zlato x Líná huba, holé neštěstí.*[84]

Příklady frazémů podle dělení Čeština, řeč a jazyk jsou **tradiční** (tzv. lidové). Existuje i frazeologie **kulturní,** která má antický původ *(Kostky jsou vrženy – Alea icta*

[83] KARLÍK, Petr et al. *Příruční mluvnice češtiny*. Vyd. 2., opr. [i.e. 4. vyd.]. Praha: NLN, Nakladatelství Lidové noviny, 2012, s. 73.
[84] ČECHOVÁ, Marie. *Čeština – řeč a jazyk*. 2., přeprac. vyd. Praha: ISV, 2000, s. 66-67.

est – výrok Gaia Julia Caesara), ale i středověký *(Cuius regio, eius religio – Čí je země, toho je náboženství).* K tomu patří i dvojslovná **rčení cizojazyčná:** *in memoriam (na paměť, k uctění památky), in natura (v přirozené/přírodní době), sine anno, sine loco (bez roku, bez místa – tj. u vydání spisu, u kterého není tento údaj uveden).* Další kulturní frazémy se do češtiny dostaly z francouzštiny, italštiny a angličtiny *(faux pas, salto mortale, My home, my castle)*. Mezi kulturní a tradiční frazeologií je pohyblivá hranice. V současnosti vznikají nové frazémy nebo se starší frazémy znovu objevují.[85]

5.4 Sémantika komponentů

V kombinatorice komponentů v celkovém významu celku se dělí na dvě skupiny. Jedním typem je výskyt stejného typu sémantiky komponentů, tj. **monosémie** (u každého členu), nebo druhý typ **polysemie**, srov. mýdlová bublina jako zástupce monosémní kombinace a otevřená hlava jako zástupce polysemní kombinace (oba komponenty mají více slovníkových významů). Vyskytují se případy smíšené kombinace monosémního komponentu s polysemním., srov. hořká pilulka, tekutý chléb, jít přes mrtvoly, šlápnout někomu na krk apod. Odlišností těchto typů je, že první typ vytváří sémanticky anomální výrazy (nemotivované apod.), zatímco u druhého typu se sémantika komponentů částečně podílí na významu celku. U prvního typu se připouští vznik homonym (mýdlová bublina), kde se jedná o dva paralelní a formálně identické výrazy. V druhém typu se homonyma vyskytují řidčeji. Na existenci homonym závisí výskyt dalších faktorů (tj. kompatibilita a určitý typ transformace).[86]

5.5 Lidová rčení

Vymezení názvu:

1. *Rčení je lexikálně obrazné spojení slov.*
2. *Pořekadlo je obrazné spojení slov, vyjadřuje jistotu situace.*
3. *Úsloví je tvarem, který se řadí mezi rčení a pořekadlo*.[87]

V knize J. Zaorálka *Lidová rčení* najdeme rčení dialektická, argotická a slangová, která jsou užívána na venkově i ve velkoměstech. Kromě lidových rčení jsou sem zařazena i rčení s biblickým původem a rčení zlidovělá.[88]

[85] ČECHOVÁ, Marie. *Čeština – řeč a jazyk*. 2., přeprac. vyd. Praha: ISV, 2000, s. 67.
[86] ČERMÁK, František. *Idiomatika a frazeologie češtiny*. Praha: Univerzita Karlova, 1982, s. 27-28.
[87] ZAORÁLEK, Jaroslav. *Lidová rčení*. Vyd. 4., V nakl. Academia 2. Praha: Academia, 2000, s.7.
[88] ZAORÁLEK, Jaroslav. *Lidová rčení*. Vyd. 4., V nakl. Academia 2. Praha: Academia, 2000, s. 8.

Praktická část

V praktické části se zabývám výskytem frazémů, všímám si výskytu lidových rčení, psaného korpusu v manažeru KonText, verze syn v7. Vytvořila jsem si subkorpus2015. Do dotazovacího řádku jsem použila typ dotazu CQL: [tag="VB.*" & col_type="M.*"][tag="NN.*" & col_type="M.*"], tzv. typ podle slovnědruhového vzoru. Jedná se o výskyt verba v přítomném nebo budoucím čase s obyčejným substantivem víceslovné jednotky druhu M: přísloví, okřídlená rčení, citace apod. Druhá pozice col_typu je libovolná.

Výchozí atribut je col_lemma. Minimální frekvenci jsem stanovila na 10, pod touto hranicí se mimo jiné frazémy vyskytovala 4 lidová rčení. Ve frekvenci v kategorií „Typ textu“ se zaměřuji na frekvenci skupiny textových typů, textový typ, tematickou oblast a periodicitu. S tímto typem dotazu mi KonText vyhodnotil konkordanční seznam s těmito výsledky:

počet celkových výskytů: 1 432 (absolutní frekvence); **i.p.m:** 7,1; publikováno v roce 2017

Minimální frekvence: 10

	Filter	**Word**	**Freq**
1	p/n	nejsou žerty	119
2	p/n	je krása	87
3	p/n	roste chuť	76
4	p/n	prochází žaludkem	64
5	p/n	je život	58
6	p/n	je síla	54
7	p/n	nejsou koláče	49
8	p/n	světí prostředky	46
9	p/n	dělají člověka	46
10	p/n	dělá zloděje	43
11	p/n	jsou peníze	42
12	p/n	je vůle	28
13	p/n	je útok	28
14	p/n	potvrzuje pravidlo	26
15	p/n	kácí les	26
16	p/n	létají třísky	22

17	p/n	je jistota	22
18	p/n	je základ	18
19	p/n	není žalobce	18
20	p/n	není voda	18
21	p/n	jsi člověkem	16
22	p/n	je chleba	15
23	p/n	není posvícení	15
24	p/n	okrádá rodinu	14
25	p/n	je zisk	14
26	p/n	tahá pilku	14
27	p/n	pije víno	13
28	p/n	káže vodu	13
29	p/n	poznáš přítele	13
30	p/n	je matka	13
31	p/n	není soudce	12
32	p/n	je změna	12
33	p/n	je pes	11
34	p/n	je cesta	11
35	p/n	začíná svoboda	11
36	p/n	je žena	11
37	p/n	je boj	10
38	p/n	není kouře	10
39	p/n	dělaj člověka	10
40	p/n	předchází pád	10

KŘEN, M. et al. Korpus SYN, verze 7 z 29. 11. 2018 [online]. Ústav Českého národního korpusu FF UK, Praha 2017. [cit. 2019-06-18] Dostupný z WWW: http://www.korpus.cz

Frekvence nižší 10

	Filter	**Word**	**Freq**
64	p/n	není šprochu	5
81	p/n	kuje železo	3
82	p/n	hodí kamenem	3
110	p/n	nezarmoutíš slibem	1

KŘEN, M. et al. Korpus SYN, verze 7 z 29. 11. 2018 [online]. Ústav Českého národního korpusu FF UK, Praha 2017. [cit. 2019-06-18] Dostupný z WWW: http://www.korpus.cz

Frekvenční seznam

	Filter	doc.txtype_group	Freq	i.p.m.
1	p/n	NMG: publicistika	1292	7,1
2	p/n	NFC: oborová literatura	140	7,12

KŘEN, M. et al. Korpus SYN, verze 7 z 29. 11. 2018 [online]. Ústav Českého národního korpusu FF UK, Praha 2017. [cit. 2019-06-18] Dostupný z WWW: http://www.korpus.cz

	Filter	doc.txtype	Freq	i.p.m.
1	p/n	NEW: tradiční publicistika	795	5,67
2	p/n	LEI: volnočasová publicistika	497	11,86
3	p/n	PRO: profesní literatura	79	6,19
4	p/n	POP: populárně naučná	61	9,32

KŘEN, M. et al. Korpus SYN, verze 7 z 29. 11. 2018 [online]. Ústav Českého národního korpusu FF UK, Praha 2017. [cit. 2019-06-18] Dostupný z WWW: http://www.korpus.cz

	Filter	doc.genre	Freq	i.p.m.
1	p/n	NTW: celostátní tisk	413	5,95
2	p/n	REG: regionální tisk	382	5,4
3	p/n	SCT: společenský život	167	15,47
4	p/n	LIF: životní styl	163	14,04
5	p/n	MIX: společnost	80	10,92
6	p/n	HOU: bydlení, zahrada, hobby	55	16,37
7	p/n	ECO: ekonomie, obchod, logistika	47	8,63
8	p/n	TEC: technika	35	6,42
9	p/n	SPO: sport	26	3,46
10	p/n	PSY: psychologie	13	18,51
11	p/n	EDU: pedagogika	12	19,34
12	p/n	ICT: výpočetní technika	10	6,57
13	p/n	LAN: filologie	8	14,07
14	p/n	INT: zajímavosti ze světa	6	4,59

15	p/n	AGR: zemědělství	6	2,16
16	p/n	REC: sport, rekreace, hobby	5	20,71
17	p/n	LAW: právo	2	4,25
18	p/n	MED: lékařství	2	1,53

KŘEN, M. et al. Korpus SYN, verze 7 z 29. 11. 2018 [online]. Ústav Českého národního korpusu FF UK, Praha 2017. [cit. 2019-06-18] Dostupný z WWW: http://www.korpus.cz

	Filter	**doc.periodicity**	**Freq**	**i.p.m.**
1	p/n	DA: deník	796	5,54
2	p/n	WE: týdeník, čtrnáctideník	462	12,05
3	p/n	MO: měsíčník	172	9,76
4	p/n	BI: nižší než měsíčník	2	0,97

KŘEN, M. et al. Korpus SYN, verze 7 z 29. 11. 2018 [online]. Ústav Českého národního korpusu FF UK, Praha 2017. [cit. 2019-06-18] Dostupný z WWW: http://www.korpus.cz

Závěr

V teoretické části představuji korpusovou lingvistiku, osvětluji, co korpus znamená a představuji typy korpusů, historii korpusové lingvistiky a korpusovou metodologii. Dále zmiňuji Český národní korpus, jeho členění a reprezentativnost a práci s korpusovou databází, ze které jsem vycházela při analýze praktické části v psaném korpusu. V teorii zmiňuji Františka Čermáka a jeho spolupracovníky Českého národního korpusu. V kapitole Základní problematika frazeologie uvádím obecnou teorii, dělení frazémů a vymezení frazémů podle Příruční mluvnice češtiny a Čeština, řeč a jazyk. Formuluji pojem lidových rčení podle Jaroslava Zaorálka.

V praktické části jsem pracovala v korpusové databázi, verze syn v7, s korpusovým manažerem KonText. Další důležité informace pro práci v korpusu jsem získala v elektronické Příručce ČNK a ve studijní příručce Jak využívat Český národní korpus od F. Čermáka. Zabývala jsem se výskytem frazémů v psaném korpusu, zejména lidových rčení. Vytvořila jsem si subkorpus2015. Do konkordančního řádku jsem použila dotaz CQL, který se využívá při složitějších dotazech a má specifický formát. Je proto vhodný při analýze frazémů, protože s přesností vyhledá výskyty v textu podle zadaného typu vzoru. Zadala jsem si dotaz [tag="VB.*" & col_type="M.*"] [tag="NN.*" & col_type="M.*"], tzv. typ podle slovnědruhového vzoru. Tento dotaz byl zaměřený na verbum v přítomném nebo budoucím čase s obyčejným substantivem víceslovné jednotky druhu M: přísloví, okřídlená rčení, citace apod. Druhá pozice byla libovolná, může tedy být hlavní či závislá. Výchozí atribut jsem nastavila na col_lemma. Počet celkových výskytů byl 1 432, (absolutní frekvence); i.p.m. (relativní frekvence) 7,1, publikovaných textů v roce 2017. Z těchto výskytů jsem minimální frekvenci nastavila na 10. Na základě výsledků korpusu jsem si sestavila tabulky a potvrdila si domněnku, že frekvence frazémů se v textu objevuje, a to nejčastěji v podobě přísloví *(příležitost dělá zloděje, když se kácí les létají třísky, každý den není posvícení, pije víno káže vodu, pýcha předchází pád, šaty dělaj člověka atd.)*, citátů *(nejlepší obrana je útok, v jednoduchosti je síla, rodina je základ státu, čas jsou peníze atd.)*, rčení *(výjimka potvrzuje pravidlo, v jednoduchosti je krása, účel světí prostředky atd.)*. U pranostiky a pořekadla byla frekvence minimální. Pod hranicí frekvence 10 jsem vyhledala 4 lidová rčení *(není šprochu, aby na něm nebylo pravdy trochu; kuje železo, dokud je žhavé; ať hodí kamenem; nezarmoutíš slibem)*.

Ve frekvenční skupině typu textu mě zajímala frekvence skupině textových typů,

textový typ, tematická oblast a periodicita. Jednoznačně nejvyšší frekvence ve skupině textových typů byla v publicistice, kde bylo zaznamenáno 1 292 výskytů. Oborová literatura byla méně frekventovaná, bylo vyhledáno pouze 140 výskytů. U obou skupin textových typů byla i.p.m. hodnota velmi podobná (7,1 a 7,12). V textovém typu frekvence bylo zastoupení tradiční publicistiky (freq. 795), volnočasové publicistiky (freq. 497), profesní literatury (freq. 79) a populárně naučné literatury (freq. 61). Tematická oblast byla početná, jedná se o oblasti celostátního tisku (freq. 413); regionálního tisku (freq. 382); společenského života (freq. 167); životního stylu (freq. 163); společnosti (freq. 80); bydlení, hobby, zahrady (freq. 55); ekonomie, obchodu, logistiky (freq. 47); techniky (freq. 35); sportu (freq. 26); psychologie (freq. 13); pedagogiky (freq. 12); výpočetní techniky (freq. 10); filologie (freq. 8); zajímavostí ze světa (freq. 6); zemědělství (freq. 6); sportu, rekreace, hobby (freq. 5); práva (freq. 2) a lékařství (freq. 2). Poslední položkou frekvenčního seznamu, jež mě zajímala, byla periodicita. Z hlediska nejvyšší frekvence byl na první pozici deník (freq. 796); dále týdeník, čtrnáctideník (freq. 462); měsíčník (freq. 172) a nižší než měsíčník (freq. 2). Míra hodnoty i.p.m. byla vysoká u týdeníku a čtrnáctideníku, a to 12,05. S těmito výsledky jsem si potvrdila domněnku, že vysoká míra frekvence se vyskytuje v publicistice.

Psaní bakalářské práce pro mě nebylo úplně snadné, a to zejména z důvodu náročnosti zorientování se v korpusovém manažeru KonText při dotazu CQL v Českém národním korpusu. Díky pomoci a cenným radám PhDr. Marie Kopřivové, Ph.D., jsem se s Českým národním korpusem brzy seznámila a práce tak byla snadnější.

Seznam literatury a zdrojů

Knižní publikace

ČECHOVÁ, Marie. *Čeština – řeč a jazyk*. 2., přeprac. vyd. Praha: ISV, 2000.

ČERMÁK, František. *Idiomatika a frazeologie češtiny*. Praha: Univerzita Karlova, 1982.

ČERMÁK, František. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum, 2017.

ČERMÁK, František a Renata BLATNÁ. *Jak využívat český národní korpus*. Praha: Lidové noviny, 2005.

ČERMÁK, František a Jiří HRONEK (eds.) *Slovník české frazeologie a idiomatiky*. Praha: Academia, 1983.

KARLÍK, Petr, Marek NEKULA, Zdenka RUSÍNOVÁ a Miroslav GREPL. *Příruční mluvnice češtiny*. Vyd. 2., opr. [i.e. 4. vyd.]. Praha: NLN, Nakladatelství Lidové noviny, 2012.

KUČERA, Karel (ed.). *Český národní korpus: úvod a příručka uživatele*. Praha: Filozofická fakulta UK, 2000.

OSOLSOBĚ, Klára. *Česká morfologie a korpusy*. Praha: Karolinum, 2014.

ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999.

TŘEŠTÍK, Michael (ed.). *Kdo je kdo: osobnosti české současnosti: 5000 životopisů*. Praha: Agentura, Kdo je kdo, 2002.

ZAORÁLEK, Jaroslav. *Lidová rčení*. Vyd. 4., V nakl. Academia 2. Praha: Academia, 2000.

Časopis

CVRČEK, Václav a Dominika KOVÁŘÍKOVÁ (eds.). Možnosti a meze korpusové lingvistiky. *Naše řeč – Základní informace* [online]. 2011. roč. 94, č. 3 [cit. 2019-03-10]. Dostupné z: http://nase-rec.ujc.cas.cz/archiv.php?art=8191

Internetové zdroje

CVRČEK, Václav a Olga RICHTEROVÁ (eds). [online]. Příručka ČNK: 25.05.2018. Dostupné na: http://wiki.korpus.cz/doku.php?id=start&rev=1527255653

KŘEN, M. et al. *Korpus SYN, verze 7 z 29. 11. 2018* [online]. Ústav Českého národního korpusu FF UK, Praha 2017. Dostupný z WWW: http://www.korpus.cz

Ústav Českého národního korpusu [online]. Copyright © FF UK 2015. Dostupné z: http://ucnk.korpus.cz

Ústav Českého národního korpusu. [online]. Copyright © FF UK 2015. Dostupné z: https://ucnk.ff.cuni.cz

VEČERKA, Radoslav. *Biografickobibliografické medaloinky českých lingvistů: bohemistů a slavistů (2. vydání)* [online]. Linguistica online. Brno: electronic journal of the Department of Linguistics and Baltic Languages at Masaryk University, Czech Republic, 2008. ISSN 1801-5336.

Dostupné z: http://www.phil.muni.cz/linguistica/art/vecerka/vec-medailonky2.pdf