KORPUSY V JAZYKOVÉM VYUČOVÁNÍ

Kateřina Šormová, Karel Šebesta a kol.

PRÁCE FILOZOFICKÉ FAKULTY UNIVERZITY KARLOVY

Publikace Korpusy v jazykovém vyučování je určena učitelům, studentům učitelství i dalším zájemcům o vyučování jazyka. V první části se čtenář seznámí s teoriemi osvojování jazyka a jazykovým výzkumem, jsou mu představeny nejznámější a největší korpusy angličtiny, češtiny a výběrově i dalších jazyků s důrazem na korpusy akviziční, zejména na projekt AKCES. Pozornost je zaměřena také na typy korpusů a jejich parametry, tedy velikost, vyváženost, autentičnost, strukturní a poziční atributy. V druhé části publikace je představena práce s korpusy Českého národního korpusu, vyhledávání pomocí rozhraní KonText a pomocí aplikací SyD, Morfio, KWords, Treq a ukázka zapojení frekvenční analýzy do jazykového vyučování. Poslední část publikace tvoří pracovní listy věnované využití korpusových dat v jazykovém vyučování. Součástí pracovních listů je i podrobný popis metodického postupu a řešení zadaných cvičení.

FILOZOFICKÁ FAKULTA
Univerzita Karlova

KORPUSY V JAZYKOVÉM VYUČOVÁNÍ

Kateřina Šormová, Karel Šebesta a kol.

FILOZOFICKÁ FAKULTA UNIVERZITY KARLOVY, 2019

Kniha vznikla díky podpoře projektu Zvýšení kvality vzdělávání a začleňování žáků s OMJ spojené s jeho radikální inovací (CZ.07.4.68/0.0/0.0/16_037/0000299) a Kreativita a adaptabilita jako předpoklad úspěchu Evropy v propojeném světě (CZ.02.1.01/0.0/0.0/16_019/0000734).

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
OP Praha – pól růstu ČR

Recenzovali
Mgr. Svatava Škodová, Ph.D.
doc. PhDr. Milan Hrdlička, CSc.
Mgr. Jarmila Valková, Ph.D.

ISBN 978-80-7308-897-2 (online : pdf)

Obsah

Úvod 7

Seznámení s korpusy a korpusovou lingvistikou

1. Korpusy, jejich typy; korpusy češtiny (*Karel Šebesta*) 11
2. Teorie osvojování jazyka a korpusový výzkum (*Tomáš Gráf*) 16
   - 2.1. Korpus a jazykové testování 17
   - 2.2. Žákovský korpus a jazyk pro specifické účely 19
   - 2.3. Analýza žákovského jazyka 21
     - 2.3.1. Gramatika 22
     - 2.3.2. Slovní zásoba 24
     - 2.3.3. Frazeologie 27
     - 2.3.4. Diskurz 29
     - 2.3.5. Pragmatika 31
     - 2.3.6. Plynulost 32
3. Významné parametry korpusů (*Karel Šebesta*) 34
   - 3.1. Velikost 34
   - 3.2. Vyváženost 37
   - 3.3. Autentičnost jazyka 40
   - 3.4. Strukturní atributy 43
   - 3.5. Poziční atributy – lingvistická (a chybová) anotace 44
4. Práce s korpusem (*Věra Hejhalová*) 47
   - 4.1. Český národní korpus (ČNK) 47
     - 4.1.1. Materiál/Obsah 47
     - 4.1.2. Vyhledávač neboli korpusový manažer 47
     - 4.1.3. Vyhledávání 48
     - 4.1.4. Volba korpusu 49
     - 4.1.5. Typ dotazu 49
     - 4.1.6. Zadání dotazu 51
     - 4.1.7. Specifikace kontextu 52
     - 4.1.8. Omezení hledání 53
     - 4.1.9. Výsledky hledání 54

4.2. Korpusové aplikace 60
4.2.1. Aplikace SyD 61
4.2.2. Aplikace Morfio 65
4.2.3. Aplikace KWords 67
4.2.4. Aplikace Treq 70

Využití korpusů ve výuce

5. Náměty pro strategie vyučování a pro tvorbu pracovních listů 75
5.1. Frekvenční analýza (*Kateřina Šormová*) 75
5.2. Pracovní listy (*Barbora Kukrechtová*) 78
5.2.1. Prostředí KonText 79
5.2.2. Morfio 89
5.2.3. SyD 93
5.2.4. KWords 96

Použitá literatura 103
Profily autorů 105
Rejstřík 107
Resumé/Summary 109

Úvod

Milí čtenáři,

právě jste otevřeli publikaci Korpusy v jazykovém vyučování. Je určena učitelům, studentům filologických oborů, jazykovým lektorům i všem dalším zájemcům o jazyk a jeho vyučování. Při přípravě této příručky jsme měli na mysli dva cíle. Prvním bylo představení jazykového korpusu jako nástroje, jehož prostřednictvím můžeme zajímavým způsobem zkoumat jazyk, pohlížet na něj z různých úhlů pohledu a v určitém smyslu si s ním i hrát. Zejména se zaměřujeme na korpusy akviziční, obsahující jazyková data z projevů mluvčích, kteří se jazyk teprve učí. Druhým cílem bylo poskytnout Vám inspiraci v tom, jak můžeme jazykový korpus využít při výuce češtiny jako cizího/druhého jazyka. Z toho důvodu tvoří poslední kapitolu příručky několik pracovních listů, které by Vám měly posloužit jako návodná ukázka aktivit, které lze na základě práce s korpusem do vyučování přinést. Budeme rádi, pokud Vás přivedou k tvorbě Vašich vlastních aktivit a pracovních listů založených na korpusových datech. Přejeme Vám příjemné čtení a radost z práce s jazykem.

Autoři

Seznámení s korpusy a korpusovou lingvistikou

1. Korpusy, jejich typy; korpusy češtiny

Corpus znamená latinsky tělo. Ve společenskovědním výzkumu se tohoto výrazu (resp. českého korpus) užívá přeneseně pro označení základního souboru dat, s nímž se dále pracuje. Pokud např. zkoumáme komunikaci při vyučovacích hodinách, můžeme pro tuto potřebu shromáždit určitý počet nahrávek vyučovacích hodin a tento materiál používat při výzkumu jako korpus.

Jazykový korpus je rovněž takovým souborem (jazykových) dat, s nimiž je možné dále pracovat. Nejde ovšem o soubor jakýchkoli dat, ale o soubor vybudovaný (jazyková data do něj jsou vybrána, zpracována a uspořádána) podle přesných a explicitních lingvistických kritérií tak, aby reprezentoval určitou oblast jazykové praxe, jazykového úzu.

Elektronický či počítačový jazykový korpus (dále jen korpus) bývá vymezován několika různými znaky. Klasickou definici jazykového korpusu uvádějí F. Čermák a V. Schmiedtová: *Korpus se obvykle vymezuje jako strukturovaný, unifikovaný (a často též označkovaný) rozsáhlý soubor jazykových dat, který je elektronicky uložený i zpracovávaný; skládá se zpravidla z jednotlivých textů a jako celek si činí nárok na reprezentativnost vzhledem k vytčenému cíli* (Čermák – Schmiedtová 2004: 154). Obecně dnes platí, že jde o soubor strukturovaný, unifikovaný (aby bylo možné použít pro jeho zpracovávání a vyhledávání v něm automatických nástrojů) a elektronicky uložený a zpracovávaný.

Vedle široce přístupných korpusů nekomerčních známe i korpusy komerční. Bývají to často korpusy speciální, které budují např. velká nakladatelství specializující se na vydávání slovníků, gramatik a učebnic pro jazykovou výuku, zvl. angličtiny (*Longman, Cambridge University Press* apod.).

Podle funkce a obsahu (podle zahrnutého jazykového materiálu a jeho vztahu k obvyklému, normálnímu užívání jazyka) lze rozlišit dva hlavní typy korpusů:

a. korpusy obecné, tedy takové, které byly vytvořeny s cílem nabídnout pro výzkumné a jiné účely dostatečně velký vzorek normálního, obvyklého užívání příslušného jazyka jako celku a o nichž lze předpokládat, že normální, obvyklý jazykový úzus reprezentují;[1]

1 Reprezentativnost ve vztahu k jazyku jako celku je pouze relativní, můžeme ale říci, že tvůrci korpusů k reprezentativnosti jako cíli směřují. Srov. vyjádření F. Čermáka (in Čermák a kol., eds. 2011: 16) o tom, že lingvistický korpus je takový korpus, „*který umožňuje vy-*

b. korpusy speciální, které takovou ambici, reprezentovat normální, obvyklý jazykový úzus, nemají, jsou zaměřeny často na užívání jazyka nějakou zvláštní skupinou či jednotlivcem, příp. zachycují jazyková data, o nichž z nějakého jiného důvodu nelze předpokládat, že normální užívání jazyka v jeho celku reprezentují.
Jedním z rozšířených typů speciálních korpusů jsou korpusy akviziční, tedy takové, které *slouží primárně studiu procesů osvojování jazyka, včetně tzv. pozdějšího jazykového vývoje a užívání jazyka mluvčími, kteří (daný) jazyk neovládají na úrovni odpovídající úrovni dospělého rodilého mluvčího*, sekundárně pak *mohou plnit a plní řadu důležitých funkcí v didaktice jazyka: jsou významným zdrojem dat při tvorbě učebnic a učebních pomůcek, jako jsou slovníky nebo gramatiky, při přípravě testů a jazykových cvičení různého typu a uplatňují se i přímo jako didaktický nástroj v jazykové výuce* (Šebesta – Škodová 2012: 6).

Pokud jde o povahu jazykových dat akvizičních korpusů, jde v naprosté většině případů o doklady užívání jazyka např. dětmi předškolního věku, mládeží ve školním věku, nerodilými mluvčími apod. K tomu se přiřazují (především z praktických důvodů) data spojená s péčí o osvojování jazyka, tedy nahrávky vyučovacích hodin, jazyková data z učebnic a dalších učebních materiálů apod.

Akviziční korpusy takto vymezené mohou sloužit i výzkumům jinak zaměřeným: např. studiu jazyka mládeže jako sociální skupiny, popř. studiu jazyka některých subkultur mládeže či jejích sociálně vymezených podskupin, studiu vývojových trendů jazyka obecně (např. při srovnání jazyka mládeže s jazykem starších generací) apod.

Podtypem akvizičních korpusů jsou korpusy žákovské, sloužící studiu osvojování jazyka nerodilými mluvčími. Obsahují data z užívání jazyka nerodilými mluvčími a také jazyková data z výuky cizího jazyka.

Trochu volněji se k akvizičním korpusům řadí korpusy zaměřené na studium tzv. zděděného jazyka (*heritage language*) a jeho osvojování, tedy osvojování a užívání jazyka příslušníky krajanských komunit v zahraničí. S akvizičními korpusy je spojuje skutečnost, že i ony slouží studiu osvojování jazyka a sekundárně didaktickým účelům (např. praxi různých krajanských škol a kurzů pro krajany, Českých škol bez hranic, učebních pomůcek pro ně apod.).

Od akvizičních korpusů se odlišují tím, že korpusy zaměřené na jazyk krajanů mohou být budovány a využívány také pro potřeby studia jazykové atrice, rozpadu zděděného jazyka, tedy jako korpusy někdy označované

vážené a reprezentativní zkoumání relativně celého jazyka, nikoliv jeho části, často v nezdůvodněných proporcích".

termínem korpusy atriční. Krajanské korpusy jsou také důležitým zdrojem dat pro výzkumy jinak zaměřené, poskytují např. unikátní možnosti studia nářečí, jejichž prvky bývají někdy v krajanském jazyce zachovány, pro studium jazykového kontaktu obecně apod.

Z praktických důvodů je ovšem účelné korpusy akviziční a atriční neoddělovat příliš striktně: při budování i využívání obou těchto typů korpusů je nutné se vyrovnávat s podobnými problémy, jako jsou např. nestandardní jazyková data, obtížnost jejich získávání, vysoký podíl manuální práce při jejich budování, a užívá se k tomu podobných specificky pro tyto účely vyvinutých korpusových nástrojů (viz dále).

Podstatně volnější vztah k akvizičním korpusům mají korpusy sloužící studiu jazykových poruch a jejich terapie. Je možné označit je jako korpusy terapeutické (nejde o termín zavedený). S akvizičními korpusy je sbližuje skutečnost, že zachycují data z užívání jazyka mluvčími, kteří daný jazyk neovládají na úrovni dospělého rodilého mluvčího; při jejich budování se tedy mohou využít podobné nástroje a pravidla (např. přepisů) jako u ostatních akvizičních korpusů. Je proto z praktických důvodů účelné budovat je ve spolupráci s korpusy akvizičními.

Pro výuku češtiny jako cizího jazyka můžeme využívat především obecných korpusů češtiny a také specificky pro tento účel budovaných žákovských korpusů češtiny jako cizího/druhého jazyka. V jisté míře se mohou uplatnit i ostatní typy korpusů akvizičních, především akviziční korpusy češtiny jako prvního jazyka (L1). V další části kapitoly se tedy budeme věnovat prioritně oběma prvně uvedeným typům korpusů, korpusům obecným a žákovským; v některých případech, kde to pro nás může být zajímavé, se dotkneme i akvizičních korpusů prvního jazyka.

Počátky počítačových korpusů se zpravidla kladou do 60. a 70. let, a to jak korpusů obecných (jako první se většinou uvádí tzv. *Brown Corpus*, celým názvem *Brown University Standard Corpus of Present-Day Edited American English*, budovaný Henrym Kučerou a Nelsonem Francisem od r. 1961 a zveřejněný 1964, a *Lancaster-Oslo/Bergen Corpus* (*LOB Corpus*), vytvořený v letech 1970–1978 a zpracovávající jako britský protějšek *Brown Corpusu* angličtinu britskou), tak i speciálních – jako jeden z prvních korpusů speciálních se uvádí např. *American Heritage Intermediate Corpus*, zahrnující vzorky z nejčtenějších knih americké mládeže ve školním věku v roce 1969.

První obecné počítačové korpusy ovšem navazují na podstatně delší tradici systematických sběrů jazykových dat a tvorby jazykových korpusů neelektronických. Odkazuje se v této souvislosti většinou na sběry citací, zpravidla pro lexikografické účely.[2]

2 Přehled korpusových počátků pro oblast angličtiny i češtiny podává např. Michal Šulc (Šulc, 1999).

To platí i pro korpusy češtiny, také česká korpusová lingvistika navazuje na dlouhou a bohatou tradici jinak než elektronicky zpracovávaných sběrů. První snahy o vytvoření centrální národní instituce, která by koncentrovaně rozvíjela korpusové aktivity pro češtinu, se objevují přibližně od druhé poloviny 80. let. V r. 1994 vznikl na FF UK pod vedením F. Čermáka Ústav Českého národního korpusu, který budování korpusů českého jazyka, obecných i speciálních, zaštiťuje.

V současné době nabízí:

- synchronní korpusy psaného jazyka řady *SYN*,
- specializované korpusy psaného jazyka různého zaměření, včetně akvizičních (viz dále),
- synchronní korpusy mluveného jazyka řady *ORAL* a nový korpus *ORTOFON*,
- specializované korpusy mluveného jazyka různého zaměření, včetně akvizičních (viz dále),
- diachronní korpus *DIAKORP*,
- paralelní korpus *InterCorp*,
- srovnatelné korpusy *Aranea*, psané jednojazyčné webové a speciální korpusy jiných jazyků, než je čeština.

Pro učitele češtiny jako cizího jazyka mají význam vedle korpusů akvizičních primárně obecné korpusy češtiny, tedy korpusy řady *SYN* a *ORAL*, resp. *ORTOFON*, pro vyšší a nejvyšší úrovně a pro některé speciální účely je využitelný korpus paralelní – *InterCorp*.

Vedle korpusů Českého národního korpusu je pro učitele významný rovněž *Pražský závislostní korpus*, budovaný Ústavem formální a aplikované lingvistiky MFF UK; z něj vycházejí i některé didaktické pomůcky zaměřené na výuku syntaxe.

Korpusy akviziční navazují na neméně dlouhou tradici systematického sběru jazykových dat pro účely studia osvojování a pozdějšího vývoje jazyka dětí předškolního i školního věku, studia jazyka mládeže jako specifické jazykové variety, jazyka nerodilých mluvčích, jazyka mluvčích postižených jazykovou poruchou apod.[3]

V roce 2005 byl pod vedením K. Šebesty zahájen projekt *AKCES* (Akviziční korpusy češtiny) při Ústavu českého jazyka a teorie komunikace FF UK, na jeho vytváření se však významně podílejí i další pracoviště.[4]

3 K této akviziční tradici sběrů v mezinárodním i domácím výzkumu a ke vzniku prvních akvizičních korpusů počítačových podrobněji viz Šebesta, 2010.

4 Ústav teoretické a komputační lingvistiky, Ústav formální a aplikované lingvistiky MFF UK, Ústav Českého národního korpusu, Ústav jazykové a odborné přípravy UK a další.

Projekt *AKCES* je pojat poměrně široce:[5] zahrnuje postupné vytváření databází a návazně budování korpusů psaných i mluvených zaměřených na:

a. jazyk dětí předškolního věku,
b. jazyk dětí a mládeže od 5 do 24 let,
c. jazyk nerodilých mluvčích češtiny,
d. jazyk sociokulturně znevýhodněných skupin,
e. jazyk krajanských komunit,
f. jazyk ve vzdělávacím kontextu.
g. Poskytuje podporu rovněž korpusům terapeutickým.

Jednotlivé korpusy jsou zveřejňovány zčásti v rámci Českého národního korpusu, zčásti v rámci Centra jazykové výzkumné infrastruktury *LINDAT/CLARIN*[6] Ústavu formální a aplikované lingvistiky MFF UK.

Pro výuku nerodilých mluvčích jsou relevantní zejména:

- korpusy žákovské (jazyka nerodilých mluvčích); z nich byly dosud v rámci Českého národního korpusu[7] zveřejněny korpusy řady *CZESL* (polovina korpusu *CZESL-PLAIN*, korpusy *CZESL-SGT* a *CZESL-MAN* – druhý z nich není dosud veřejně přístupný), v rámci Centra *LINDAT/CLARIN* pak korpus AKCES 5; volně se k nim řadí i česká část korpusu MERLIN, zpracovaná v rámci samostatně financovaného německo-italsko-českého projektu;[8]
- korpusy jazyka žáků ze sociokulturně znevýhodňujícího prostředí – z korpusů zveřejněných v rámci Českého národního korpusu sem patří cca čtvrtina korpusu *CZESL-PLAIN*, připravuje se k zveřejnění srovnávací korpus *Compar*, jehož polovinu tvoří data mluvčích ze sociokulturně znevýhodněných komunit; z korpusů zveřejněných v rámci *LINDATu* jsou to korpusy *AKCES 4* a mluvený korpus *ROMi_1.0*;
- korpusy písemných i mluvených projevů českých žáků – v rámci Českého národního korpusu byl zveřejněn korpus písemných prací žáků *SKRIPT2012*; jejich mluvené projevy jsou zachyceny zčásti ve školních dialozích přepsaných a zveřejněných v korpusech *SCHOLA 2010* (na stránkách ČNK) a *AKCES 2 ver 2*.

5 Podrobnější informace o celém projektu jsou na stránkách http://akces.ff.cuni.cz/.
6 https://lindat.mff.cuni.cz/cs/
7 https://kontext.korpus.cz/first_form
8 https://merlin-platform.eu/

2. Teorie osvojování jazyka a korpusový výzkum

Nauka o jazykové akvizici se zabývá zkoumáním procesů osvojování mateřského či cizího jazyka. Jakkoliv doklady o spekulacích o podstatě tohoto procesu nacházíme již ve starověku, jako vědní obor a součást aplikované lingvistiky se ustavuje až v polovině 20. století, kdy nachází své místo v univerzitních studijních programech.

Jako každý vědecký obor pracuje i nauka o jazykové akvizici s konkrétními daty a od 80. let 20. století tato data stále častěji pocházejí z tzv. akvizičních korpusů. Jde o speciální typ jazykových korpusů, které si nejlépe představíme jako počítačově zpracované a prohledávatelné sbírky textů, jejichž původci jsou ti, kdo si jazyk osvojují: děti osvojující si mateřštinu nebo studenti (ať už v dětském, či dospělém věku) osvojující si další, zpravidla cizí, jazyk. Existují přitom různé typy těchto korpusů v závislosti na médiu (psaný či mluvený korpus), typu jazyka (mateřský či cizí) či charakteristice mluvčího (dítě, dospělý, mluvčí s řečovou poruchou, rodilý či nerodilý mluvčí, mluvčí ze sociokulturně či jinak znevýhodněné komunity aj.). Společné pak mají tyto korpusy to, že zachycují tzv. mezijazyk, tedy jazyk ve vývojové fázi, která ještě nedosahuje úrovně obvyklé pro dospělého rodilého mluvčího. Akviziční korpusy jsou tak sbírkou dokladů o tom, jak osvojování probíhá a jaké jsou typické rysy osvojovaného jazyka. Umožňují tak hledat odpověď na řadu klíčových otázek, kterými se nauka o akvizici zabývá, jako jsou například role jazykového transferu (tj. vlivu mateřštiny na osvojování cizího jazyka) a odlišnosti mezi mezijazykem a jazykem rodilého mluvčího, jakož i odlišnosti ve vyjadřování různých mluvčích. Je tak možné srovnávat jazykový projev studentů s různou mateřštinou, kteří si ale osvojují stejný cizí jazyk (např. angličtinu).

Tradiční výzkum v oblasti jazykové akvizice často pracoval s omezenými vzorky dat a nezřídka i s intuicí badatelů a učitelů cizích jazyků. Často tak byly výsledky výzkumu přijímány s výhradou, že se zjištěné skutečnosti mohou týkat právě jen onoho omezeného vzorku mluvčích a nedají se zobecnit. Akviziční korpusy naproti tomu poskytují velké množství dat sesbírané od desítek, stovek či tisíců mluvčích a poskytuje tak rozsáhlou bázi pro empirické ověřování hypotéz s velkou možností zobecnění závěrů.

Specifickým a zřejmě i nejběžnějším typem akvizičních korpusů jsou korpusy žákovské (angl. learner corpora). Jsou vystavěny z textů nerodilých

mluvčích, kteří se učí cizí jazyk. Jazyk, který vyprodukují, pak označujeme jako jazyk žákovský (jakkoliv nezáleží na věku studenta; žákem je označován každý, kdo se učí). Zdrojem textů mohou být psané školní práce, materiály z jazykových zkoušek, přepsané nahrávky mluveného projevu či záznamy řečových úloh sestavených za účelem vytvoření samotného korpusu. V květnu 2018 čítal seznam *Learner corpora around the world*[9] 167 žákovských korpusů. Další korpusy budou jistě i nadále vznikat.

Dalším typem akvizičních korpusů jsou korpusy osvojování mateřštiny. Nejznámější z nich, korpus *CHILDES*[10] (*Child Language Data Exchange System*), začal vznikat již v r. 1984 a shromažďuje data ze 130 specializovaných korpusů dětské řeči v 26 jazycích. Korpus se posléze připojil k projektu *TalkBank*, který obsahuje korpusy akvizice mluvčích s poruchami řeči (afázie), korpusy osvojování cizích jazyků, korpusy jazyka ve školních třídách a korpusy pro konverzační analýzu.

Akviziční korpusy se vybavují různými systémy anotace. Ty přejímají buď z obecné korpusové lingvistiky (např. anotace slovnědruhová), nebo vytvářejí vlastní anotační systémy zachycující specifické rysy akvizice (např. jazykové odchylky, chyby či jiná specifika). Takto vybavené korpusy pak umožňují sofistikované metody vyhledávání a na nich založené analytické postupy, které dokážou odkrýt charakteristické rysy vyvíjejícího se jazyka, na jejichž základě je nezřídka možné formulovat i pedagogické závěry.

2.1. Korpus a jazykové testování

Využití korpusů pro účely jazykového testování je jedním z nejnovějších příkladů aplikace jazykových korpusů a je to bezesporu oblast, kde se do budoucna očekává velká aktivita, a to především v oblasti definování a popisování úrovní pokročilosti. Současný stav výzkumu v této oblasti vychází především z možností sběru dat při jazykových zkouškách. Takto získané vzorky žákovského jazyka, u nichž je školenými examinátory stanovena jazyková úroveň např. dle Společného evropského referenčního rámce (SERR), se analyzují například z hlediska četnosti určitých jevů a na základě těchto analýz je pak možné popsat, které jazykové prostředky jsou charakteristické pro danou úroveň pokročilosti. Lingvistům se tak daří reagovat na časté kritické hlasy, které si všímají, že Společný evropský referenční rámec není dostatečně specifický a v rámci popisu dovedností na jednotlivých

9 Centre for English Corpus Linguistics (date of access): Learner Corpora around the World. Louvain-la-Neuve: Université catholique de Louvain, https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

10 https://childes.talkbank.org

úrovních užívá příliš obecný jazyk. Výsledky tohoto výzkumu mají potenciál přispět k lepší srovnatelnosti.

Zatím nejvýznamnějším a nejrozsáhlejším počinem v této oblasti je projekt *English Profile*[11] (organizovaný University of Cambridge, Cambridge University Press a Cambridge English Language Assessment a podpořený Evropskou komisí). Z dat z jazykových zkoušek na všech úrovních angličtiny dle SERR byl sestaven korpus a na jeho základě pak seznamy slovní zásoby (*English Vocabulary Profile*) a gramatiky (*English Grammar Profile*) typické pro jednotlivé úrovně SERR. Tuto bezplatnou databázi mohou využívat učitelé, studenti či autoři jazykových materiálů a ověřit si, jaký konkrétní jazyk mluvčí různých úrovní angličtiny dokážou používat. Očekává se, že tento výzkum se zpětně promítne i do vývoje dokonalejších, empiricky založených technik zjišťování jazykové pokročilosti.

Největším žákovským korpusem, který vznikl na základě sběru materiálů z testů, je *Cambridge Learner Corpus*, který obsahuje v současnosti přes 57 milionů slov psaného žákovského jazyka od více než čtvrt milionu mluvčích. Korpus je přitom i nadále rozšiřován. Dalším významným projektem je *The Longman Learners' Corpus*, který poskytl podklady pro sestavení jazykových doporučení v žákovském slovníku *Longman Active Study Dictionary*. Mnoho nových výzkumných projektů nyní vzniká na základě mluveného korpusu žákovské angličtiny *Trinity-Lancaster Corpus*, který je sestaven z nahrávek a přepisů ústních zkoušek z angličtiny při zkouškách *Graded Examinations in Spoken English* organizovaných Trinity College v Londýně. Zatím zde však jde pouze o úrovně B1–C1.

Pro účely vývoje jazykových testů se však využívají i korpusy jazyka rodilých mluvčích. Například *T2K-SWAL Corpus* je souborem akademických textů, které se analyzují z pohledu analýzy diskurzu a žánrů. Výsledek tohoto výzkumu vedl k sestavení *Test of English as a Foreign Language, TOEFL iBT*.

Největší výhodou využití korpusové lingvistiky v oblasti jazykového testování je bezesporu to, že jazykové korpusy (ať už žákovského jazyka, nebo jazyka rodilých mluvčích) umožňují empiricky ověřit intuici tvůrců testových materiálů a připravit tak testy, které ověřují odpovídající jazykovou úroveň. To se může dít nejen na základě výzkumu toho, co mluvčí skutečně v jazyce dokážou, ale i toho, kde se odchylují od běžných norem (např. je možné zkoumat, jaké chyby jsou typické pro různé úrovně pokročilosti). Tvůrci jazykových testů se přitom mohou opřít o seznamy slov a jazykových struktur vygenerovaných z žákovských korpusů pro jednotlivé úrovně pokročilosti a mohou na datech z korpusu rovněž zkoumat, jak různé typy úloh ovlivňují výkon v testu a jak se určité gramatické či lexikální jevy

11 http://www.englishprofile.org

v souvislosti s rostoucí pokročilostí proměňují. Cílem tohoto výzkumu je přispět ke zlepšování podoby a vlastností jazykových testů.

Pro účely jazykového testování se však dají využít i korpusy běžné (pro angličtinu např. korpusy *British National Corpus* či *Corpus of Contemporary American English*), a to jako zdroj textů, které si mohou učitelé dle zadaných kritérií vyhledat a vytvořit si tak testy na slovní zásobu či gramatiku. Takto získaná korpusová data jsou autentická a jazykově obvykle i správná (co je v korpusu v daném registru či žánru frekventované, to zpravidla odpovídá zavedenému úzu/normě, a tudíž odráží kodifikaci).

2.2. Žákovský korpus a jazyk pro specifické účely

Jazykem pro specifické účely označujeme jazyk, který je užíván například v pracovním či akademickém prostředí a který se řídí určitými pravidly příslušného funkčního stylu (např. styl odborný, publicistický a administrativní), nezřídka je charakterizován i specifickou slovní zásobou. Jeho znalost umožňuje uživatelům začlenit se do skupiny lidí, odborníků či profesionálů, kteří takový jazyk typicky používají a kteří mají specifické komunikační potřeby a zvyklosti.

Jazyk pro specifické účely je od 60. let 20. století intenzivně zkoumán, což má mj. dopad i na jazykové vyučování, a to obzvláště pro angličtinu, která plní roli jazyka mezinárodní komunikace právě v profesní oblasti. Je také v tomto ohledu nejlépe popsána. Existují tak četné učebnice odborného jazyka, slovníky, aplikace a specializované jazykové kurzy. Rozlišuje se přitom mezi angličtinou pro pracovní účely (English for Occupational Purposes) a angličtinou pro účely akademické (English for Academic Purposes). Angličtina pro pracovní účely se nezaměřuje pouze na specifickou slovní zásobu (např. terminologii), ale především na to, jakým způsobem se v těchto oblastech lidské činnosti komunikuje, aby komunikace byla efektivní a pro dané účely vyhovující. To se týká jak jazyka psaného (např. obchodní korespondence, instruktážní texty), tak mluveného (např. komunikace lékařů s pacienty, komunikace v leteckém provozu).

Výzkum angličtiny pro akademické účely se zabývá celou škálou užití angličtiny v akademickém prostředí od interakcí mluvených (např. prezentace na konferencích) až po psaní akademických textů. Tato oblast je natolik široká, že se dále rozlišuje mezi angličtinou pro obecné a pro specificky oborové akademické účely. Do této oblasti se nezřídka zahrnují i specifické jazykové a studijní dovednosti a návyky (např. pořizování poznámek, práce s literaturou, organizace textu), ale například i prezentační, rétorické techniky. Pedagogické úsilí v této nesmírně stratifikované oblasti pak směřuje především k tomu, aby byly naplněny potřeby studentů a uživatelů jazyka v té které specializaci.

Korpusová lingvistika hraje v popisu jazyka pro specifické účely nezastupitelnou roli. Disponuje totiž technikami, jak shromáždit dostatečně velké vzorky reprezentativních ukázek odborného jazyka, na jejichž základě je možné popsat jeho typické rysy. To je důležité jak pro tvorbu pedagogických materiálů, tak pro zkoumání projevu těch, kdo si tento jazyk osvojují v rámci jiného jazyka, než je jejich mateřština. Zkoumají se např. frekvenční vzorce v jednotlivých stylech a je tak možné velmi přesně určit, která specifická slovní zásoba je pro daný styl charakteristická. Na základě korpusu *The Academic Corpus*[12] tak například vznikl seznam akademické slovní zásoby pro angličtinu (*Academic Word List*[13]), který je pro mezinárodní studenty neocenitelnou pomůckou při studiu akademické angličtiny a při psaní akademických textů. Využití pak samozřejmě nachází i při výuce a sestavování jazykových testů.

Vedle textů napsaných rodilými mluvčími se však analyzují i akademické texty nerodilých mluvčích. Vznikají tak speciální žákovské korpusy akademického žákovského jazyka. I zde zcela převládají korpusy angličtiny. Jde obvykle o databáze psaných úloh (esejů, písemných zkoušek atp.) vzniklých na univerzitách. Jazyk, který v daném kontextu studenti používají, je charakteristický jak užitím terminologie, tak užitím specifických typů argumentace a postupů při výkladu. Jedním z prvních takových korpusů je *International Corpus of Learner English* (*ICLE*), který začal vznikat v Belgii, na *Université catholique de Louvain*, v 80. letech 20. st. Práce na něm stále pokračuje, ač korpus v současné době obsahuje eseje v rozsahu 3,7 milionu slov od pokročilých studentů angličtiny z 16 různých zemí, respektive s 16 rozdílnými mateřskými jazyky.

Na stejném pracovišti se nyní rodí nový korpus akademické angličtiny s názvem *Varieties of English for Academic Purposes* (*VESPA*). Korpus je tvořen národními subkorpusy mluvčích s rozdílnou mateřštinou. Je tak možné nejen zkoumat specifika vyvíjejícího se akademického jazyka u studentů, ale zároveň srovnávat, do jaké míry může být jeho podoba ovlivněna mateřštinou. Dalšími významnými korpusy v této oblasti jsou *British Academic Written English* (*BAWE*) a americký *Michigan Corpus of Upper-level Student Papers* (*MICUSP*) a jejich mluvené protějšky *British Academic Spoken English* (*BASE*) a *Michigan Corpus of Academic Spoken English* (*MICASE*). Tyto čtyři korpusy jsou přitom ze zhruba 80 % tvořeny psaným projevem rodilých mluvčích, což umožňuje srovnávat jazyk cizojazyčných studentů s jazykem rodilých mluvčích. To je v žákovské korpusové lingvistice běžný přístup a nezřídka jsou žákovské korpusy doplňovány o kontrolní korpusy textů rodilých mluvčích, které jsou koncipovány tak, aby měly stejnou strukturu

12 https://www.victoria.ac.nz/lals/resources/academicwordlist/information/corpus

13 Pro češtinu je dostupný nástroj *Akalex*, viz www.korpus.cz/akalex.

a podobaly se kvůli srovnatelnosti v co největší možné míře. Korpus *ICLE* je takto doplněn o *Louvain Corpus of Native English Essays* (*LOCNESS*), tedy korpus stejně zadaných esejů psaných rodilými mluvčími. Takový přístup však skrývá i jedno zásadní úskalí, a tím je implicitní předpoklad, že text rodilého mluvčího je z podstaty věci dokonalejší. Stranou přitom zůstává, zda jsou vybraní autoři z řad rodilých mluvčích kvalitními autory.

2.3. Analýza žákovského jazyka

V 50. letech 20. století se někteří lingvisté[14] zabývali výzkumem jazyka za účelem zdokonalení metod jeho výuky. Vycházeli přitom z předpokladu, že je snazší naučit se jazyk, který je podobnější žákově mateřštině, a že podrobný popis podobností a rozdílů mezi mateřským jazykem a jazykem cizím je užitečnou pomůckou pro výuku a pro tvorbu jazykových učebnic. Předpokládali, že žáci budou chybovat v těch aspektech, které jsou rozdílné, a naopak snazší pro ně budou ty rysy, které jsou si mezi jazyky podobné. Domnívali se, že když se výuka bude zaměřovat především na tyto odlišnosti, žáci v nich nebudou chybovat. Skutečnost se ukázala být složitější a postupným výzkumem se ukázalo, že výuka vycházející ze srovnávání jazyků není zárukou bezchybného projevu žáků.

Toto období bylo rovněž počátkem zájmu o žákovský jazyk, a to především o chyby, kterých se žáci dopouštějí při psaní či mluvení. Chyby byly analyzovány ze všech stran a lingvisté se je pokoušeli pochopit v naději, že objeví důvod, proč žáci chybují, a budou tak moci vyvinout dokonalejší výukové metody, které žákům umožní nechybovat. Stále více lingvistů však upozorňovalo, že chyby jsou nedílnou součástí učení a že představují doklad o vývoji pokročilosti a o tom, že učení se cizímu jazyku je systematický proces. Tuto teorii završil v roce 1972 americký lingvista Larry Selinker formulací teorie tzv. mezijazyka.

Selinkerův mezijazyk je dynamický, proměnlivý systém, který se v žákově mysli formuje na základě jeho zkušeností s jazykem a který mu umožňuje cizí jazyk používat. To je možné díky pravidlům, která si vytváří sám žák, a to na základě výuky, učení se, kontaktu s osvojovaným jazykem a dalších faktorů. Existence mezijazyka je základním principem, z něhož vychází výzkum jazykové akvizice. Data pro tento výzkum pak pocházejí stále častěji z žákovských korpusů. Ty umožňují analyzovat osvojování gramatiky, slovní zásoby, frazeologie, diskurzu, stylistiky, pragmatiky a s nástupem videokorpusů v budoucnosti i výzkum nonverbální komunikace.

14 Např. Robert Lado, Charles Fries, Pitt Corder.

2.3.1. Gramatika

V centru zájmu při zkoumání žákovského jazyka stojí gramatika a slovní zásoba, které jsou vnímány jako dva základní komponenty znalosti cizího jazyka. Cílem žákovské korpusové lingvistiky v této oblasti je popsat, jakým způsobem studenti používají gramatiku cizího jazyka, který se učí, a stanovit příslušné pedagogické postupy, které z těchto informací vyplývají. Žákovská korpusová lingvistika má tedy velmi blízko k pedagogice a didaktice a přispívá k vytváření tzv. pedagogických gramatik, tzn. popisů cizojazyčných gramatických systémů při současném zohlednění možných postupů při jejich výuce.

K tomu, aby mohla být zkoumána gramatika v žákovském jazyce za použití žákovských korpusů, musí být tyto korpusy doplněny o anotace. Anotace jsou značky, které lingvisté doplňují do korpusu proto, aby se v něm dal vyhledávat nejen konkrétní řetězec znaků (např. slovo), ale i specifická jazyková informace např. o slovním druhu či přítomnosti chyby. Rozlišujeme tak především značkování morfologické a chybové. Zatímco morfologické značkování je možné provést víceméně automaticky s využitím počítačových programů, značkování chybové se provádí ručně a je značně náročné na čas. Při morfologickém značkování je každému slovu přiřazena informace o slovním druhu a příslušném jazykovém tvaru. Díky tomu můžeme při vyhledávání v korpusu zadat např. české slovo *bez* a stanovit přitom, chceme-li, aby výsledky vyhledávání zahrnuly slovo *bez* jako podstatné jméno, či jako předložku. Pro žákovské korpusy jsou pak typické značky chybové, které se přiřazují všem výskytům chyb v jazykovém projevu, ať už jsou na úrovni gramatické, pravopisné, lexikální, či jiné.

Chybová anotace se často vkládá přímo do textu. Označuje se přitom chybný tvar a často se doplňuje i správná varianta (tzv. emendace). Jako příklad uveďme systém užívaný pro chybové značkování korpusu *ICLE* (viz výše). Tento systém využívá značky složené z písmen. Ty se vkládají před chybně užitý tvar. Na prvním místě bývá označení typu chyby (tj. jde-li o pravopis, gramatiku, slovní zásobu, lexikálně-gramatický jev či slovosled). Ve větě označené Příklad 1 je ukázáno, jakým způsobem je možné označkovat chybné užití gramatického času v anglické větě. Student ve větě chybně užil přítomný čas, proto je před chybným tvarem slovesa *live* vložena značka (GVT). Ta označuje, že jde o gramatickou chybu (G) týkající se třídy sloves (V, anglicky *verb*) a kategorie času (T, anglicky *tense*). Po slovesu je vložena emendace, tedy návrh anotátora na to, jak by mohla znít správná věta. V tomto případě anotátor navrhuje, aby místo tvaru přítomného času slovesa *live* bylo užito předpřítomného průběhového času.

původní věta: *I live in Prague for ten years*
anotovaná věta: *I (GVT) live $have been living$ in Prague for ten years*

Příklad 1: Ukázka chybové anotace anglické věty – gramatika

V takto označkovaném korpusu je pak možné jednotlivé chybové značky vyhledávat a nechat si zobrazit třeba právě všechny věty s chybou v použití časů. Ty se pak dají dále analyzovat, případně je možné jich využít pro tvorbu cvičení pro výuku. Specifické uplatnění tyto postupy nacházejí při tzv. kontrastivní analýze mezijazyka, při níž se porovnává psaný nebo mluvený projev žáků s odlišnými mateřskými jazyky (např. angličtina Čechů a Španělů). Dají se tak odhalit podobnosti či rozdíly a zkoumat, z čeho tyto rozdíly pramení. Častým zdrojem chyb přitom bývá tzv. jazykový transfer, kdy si žák přenáší do svého cizojazyčného projevu prvky ze své mateřštiny. Jako ukázka nám znovu může posloužit Příklad 1, u něhož se můžeme domnívat, že chybné užití přítomného času v angličtině pramení z toho, že je v odpovídající větě v češtině používán právě přítomný čas (*Bydlím v Praze už deset let.*).

Je-li označkován celý korpus, je možno přistoupit k analýze chyb. Chyby se vyhledají a roztřídí do kategorií podle různých kritérií. S použitím statistických nástrojů je pak možné ověřovat hypotézy o souvislosti mezi chybovostí a nejrůznějšími žákovskými proměnnými, jako jsou např. jazyková úroveň, věk, mateřský jazyk, délka studia nebo pobyt v zahraničí.

Na základě chybové analýzy je pak možné sestavit katalogy chyb, které mohou být návodné pro výuku nebo se dají využít při psaní výukových materiálů. Jazykové učebnice a slovníky určené pro žáky angličtiny tak dnes nezřídka obsahují informace o častých chybách. Jsou návodné nejen pro žáky, ale i pro učitele. Existuje dokonce i slovník chyb v angličtině,[15] který obsahuje více než 2500 příkladů běžných chyb, vysvětlení daných jevů a ukázky správného užití z *British National Corpus*.

Zkoumat se ale nemusí pouze chyby. Srovnáváním jazyka rodilých mluvčích s jazykem žákovským je možné zjistit, co je pro žákovský jazyk charakteristické například frekvenčně, tj. které jevy se objevují častěji a které méně často. Mluví se pak o nadužívání či naopak menší míře používání určitých gramatických jevů. Takové analýzy jsou pak užitečné například pro výuku jazyka pro specifické účely (např. akademická angličtina, viz výše).

Jedním z nejzajímavějších a pro učitele angličtiny zřejmě i nejužitečnějších využití žákovských korpusů v oblasti gramatiky je projekt English

15 Turton, N.D. and J.B. Heaton (1996). Longman Dictionary of Common Errors. Harwich: Longman.

Grammar Profile.[16] Jde o volně přístupný internetový zdroj, který je nadstavbou Společného evropského referenčního rámce (Council of Europe, 2001). Umožňuje vyhledávat gramatické jevy podle jazykové úrovně žáků. Projekt vznikl sestavením žákovského korpusu textů z jazykových zkoušek pro jednotlivé úrovně angličtiny dle Společného evropského referenčního rámce (*Cambridge Learner Corpus)* a anotací gramatických jevů, které se v něm vyskytují. Ty pak byly rozřazeny podle pokročilosti mluvčích a bylo stanoveno, které gramatické jevy jsou charakteristické pro danou úroveň pokročilosti. Prostřednictvím projektu tak lze zkoumat nejen vývoj osvojování anglické gramatiky od začátečníků až po velmi pokročilé, ale projekt lze využít i prakticky, například pro sestavování osnov pro výuku či pro tvorbu testů, učebnic a jiných jazykových materiálů. Pomocí uživatelsky přívětivého a přehledného internetového rozhraní je možné si vybrat buď konkrétní jazykovou úroveň a prozkoumat, které gramatické jevy jsou pro ni charakteristické, nebo je možné naopak vyhledat konkrétní gramatický jev a zjistit, od kterých úrovní se v žákovské angličtině začíná typicky objevovat. Vše je doprovozeno množstvím autentických ukázek, a to v neopravené i opravené verzi, takže je možné sledovat, v čem se v daném jevu typicky chybuje.

2.3.2. Slovní zásoba

Korpusové studie slovní zásoby patří v korpusové lingvistice mezi nejčastější. Často přitom pracují s frekvencí výskytu jednotlivých slov, což umožňuje sestavit seznamy slov, která jsou více či méně běžná v určitých typech textů či stylech. Toho lze v kontextu vyučování využít například pro sestavování seznamů klíčové slovní zásoby,[17] která – jak označení napovídá – obsahuje běžná slova, jež jsou nepostradatelná pro komunikaci a jsou známa většině rodilých mluvčích. Přitažlivost tohoto konceptu spočívá v myšlence, že žák, který ovládá toto lexikální minimum, by měl znát podstatnou slovní zásobu potřebnou pro většinu běžných situací. Příkladem takového lexikálního minima je třeba definiční slovník využívaný v *Oxford Advanced Learner's Dictionary*. V tomto slovníku pro pokročilé studenty angličtiny jsou všechna slova definována na základě klíčové slovní zásoby čítající 3000 slov. Pokud student tato slova ovládá, měl by být schopen – podobně jako tento slovník – vyjádřit vlastně cokoliv. Skutečnost je však naneštěstí o něco složitější, což je dáno mimo jiné tím, že řada běžných, každodenních slov, která jsou pro uživatele jazyka důležitá (např. židle, stůl, kalhoty), ve skutečnosti v textech nepatří mezi ta nejčastější, a neobjeví se tak v seznamech vytvářených z korpusu právě na základě frekvence.

16 http://www.englishprofile.org/english-grammar-profile/egp-online

17 Angl. *core vocabulary*.

Na tento nedostatek výborně reaguje jeden z nejužitečnějších korpusových projektů pro učitele i studenty angličtiny, *English Vocabulary Profile*.[18] Pochází od stejných tvůrců jako *English Grammar Profile* a také využívá jako zdroj dat *Cambridge Learner Corpus*. Jde o unikátní projekt, který doplňuje Společný evropský referenční rámec o slovní zásobu, která je typicky využívána na jednotlivých úrovních pokročilosti. Jde přitom nejen o slova, ale i o slovní obraty, fráze a idiomy. Jak však sami tvůrci projektu tvrdí, nejde jen o seznam slov a slovních spojení. Uživatelsky přívětivé prostředí disponuje řadou funkcí. Pro každou úroveň pokročilosti je možné si vygenerovat seznam slovní zásoby, kterou tito mluvčí typicky ovládají. Je však možné nechat si vygenerovat i velice specifické seznamy slovní zásoby podle slovního druhu, typu jednotky (slovo, idiom, fráze, frázové sloveso), některých gramatických kategorií (např. počitatelnost), rejstříku (např. hovorová slova), témat (např. cestování), ale i použitých předpon a přípon. Můžeme tak snadno zjistit, jakou slovní zásobu k tématu cestování by měl znát student na úrovni např. A2. Při rozkliknutí vygenerovaných slov rovněž vidíme základní definice a příklady užití jednak ze slovníku, jednak ze samotného žákovského korpusu. Můžeme si rovněž slovo v databázi vyhledat, chceme-li zjistit, k jaké úrovni přináleží. Vzhledem k tomu, že slova mají často řadu více či méně příbuzných významů, poskytuje *English Vocabulary Profile* i údaje o tom, které významy jsou typicky součástí slovní zásoby studentů na dané úrovni. Výhodou je i to, že lze přepínat mezi britskou a americkou angličtinou, a v neposlední řadě i to, že ke slovům jsou přidány i nahrávky jejich výslovnosti. Projekt *English Profile* se svou gramatickou a lexikální částí je v současnosti jednou z nejpřesvědčivějších ukázek, jak je možné bezprostředně propojit žákovskou korpusovou lingvistiku s vyučováním a učením se cizího jazyka.

Seznamy slov podle úrovně pokročilosti nacházejí široké uplatnění při psaní zjednodušených čítanek a slovníků pro různě pokročilé žáky. Zjednodušené čítanky využívají omezeného množství slov pro převyprávění celých knih a umožňují tak, aby se při četbě žák opakovaně setkával se stejnou slovní zásobou, což – jak dokládá výzkum osvojování slovní zásoby – výrazně napomáhá učení. Zjednodušené čítanky pak mají i tu výhodu, že neodrazují žáka tím, že jsou po stránce slovní zásoby příliš složité. Uvádí se, že by na stránce zjednodušené knihy měla být maximálně 3 až 4 neznámá slova.

Seznamy slov se vytvářejí i pro potřeby výuky jazyka pro specifické účely (viz výše např. *Academic Word List*). Bývají také založeny na frekvenci výskytu, ale odstraňují se obecná frekventovaná slova, tak aby zůstala jen slova odborná. Obecně lze tedy říci, že seznamy slov vytvořené na základě různých korpusů významně přispívají k doplňování učebních sylabů

18 http://www.englishprofile.org/wordlists

o složku slovní zásoby na základě pokročilosti či specifických lexikálních potřeb studentů. V kontextu vyučování žáků s odlišným mateřským jazykem (dále jen OMJ) je toto jednou z cest, jak těmto žákům zprostředkovat slovní zásobu, kterou potřebují znát v jednotlivých školních předmětech.

Prostřednictvím žákovských korpusů je možné studovat užití slovní zásoby v žákovském jazyce. Podobně jako u gramatiky i zde hrají roli jazykové chyby. Ty jsou označeny speciální značkou a z takto vybaveného korpusu je následně možné vygenerovat seznam všech chyb, které se v korpusu v užití slovní zásoby objevují. Stejně jako u značkování gramatických chyb se u chyb lexikálních mohou značky vkládat přímo do textu, a to včetně emendace (tj. uvedení správné varianty). Ve větě uvedené jako Příklad 2 je ukázka značkování chyby v užití anglických sloves *borrow* a *lend*, která studenti často zaměňují zřejmě kvůli tomu, že čeština v obou případech využívá stejné sloveso a význam rozlišuje vyjádření zvratnosti (tj. půjčit komu a půjčit si). Před chybně užitým tvarem slovesa *borrowed* je značka (LS), která znamená, že jde o chybu lexikální (L, z angl. *lexical*) týkající se jednoho slova (S, z angl. *single*). Po slovesu je vložena emendace vyjadřující domněnku anotátora, že student chtěl ve skutečnosti použít sloveso *lend*.

původní věta:	*she borrowed me some money*
anotovaná věta:	*she (LS) borrowed $lent$ me some money*

Příklad 2: Ukázka chybové anotace anglické věty – slovní zásoba

Na základě chybové anotace slovní zásoby je tak možné zjistit, kde v užití slov žáci nejčastěji chybují. Na tyto chyby se lze posléze soustředit při výuce nebo tvorbě cvičení či jazykových materiálů. I zde se nezřídka ukazuje, že chyby jsou často důsledkem jazykového transferu, při němž student použije v cizím jazyce slovo, o němž se domnívá, že je užíváno stejně jako v jeho mateřštině. Tato slova označujeme jako zrádná, často též termínem faux amis (tj. falešní přátelé). Příkladem zrádného slova pro Čechy studující angličtinu může být slovo *klimatizace*, kterému v angličtině neodpovídá výraz *climatization*, jak se studenti někdy domnívají, ale *air-conditioning*.

Slovní zásoba se však v žákovských korpusech nezkoumá jen s ohledem na chybovost. Mezi často sledované parametry patří například frekvence, lexikální hustota, lexikální šíře či lexikální složitost. Frekvenční studie popisují, jak často studenti používají jednotlivá slova a jak se liší frekvence slov v žákovském jazyce a v jazyce rodilých mluvčích. Je tak možné například odhalit, která běžně užívaná slova se v žákovském jazyce objevují méně a naopak, která slova žáci užívají příliš. I to lze následně využít při výuce a tvorbě materiálů. Studie lexikální hustoty a lexikální šíře se pak zaměřují například na to, jaký je v textu poměr lexikálních (obsahových) slov

a slov funkčních a nakolik se slova v textu opakují. K tomu je možné využít nejrůznější nástroje volně dostupné na internetu. Lexikální složitost pak popisuje, do jaké míry žáci využívají méně frekventovaných slov. Prostřednictvím těchto testů je pak možné posuzovat pokročilost žáka s ohledem na jeho znalost slovní zásoby.

2.3.3. Frazeologie

Frazeologie zkoumá ustálená slovní spojení. Ta mohou být klasifikována na základě různých kritérií. Ve výzkumu slovní zásoby za použití korpusů se zřejmě nejčastěji zkoumají tzv. kolokace. Jde o spojení zpravidla dvou, ale i více slov, která se v korpusu často vyskytují společně a která spolu významově souvisejí a vytvářejí dohromady lexikální jednotku (např. *domácí zvíře, šedé vlasy, hustá mlha, temná noc* atp.). Některá slova mají přitom tendenci spojovat se jen s určitými jinými slovy, např. je běžné označit něco jako *krajně nevhodné*, ale spojení *krajně vhodné* rodilí mluvčí hodnotí jako neobvyklé. Rodilí mluvčí si tak zjevně na základě své jazykové zkušenosti vytvářejí mezi určitými slovy asociace, které jsou pak v jazyce opakovaným používáním upevňovány. To do značné míry charakterizuje, nakolik je text z pohledu rodilého mluvčího vnímán jako přirozený, jakkoliv ani intuice rodilého mluvčího zde nemusí být přesná a pro spolehlivé určení kolokací a jejich frekvence je nutný výzkum na základě rozsáhlých korpusů.

Ve výuce cizích jazyků hrají kolokace značnou roli. Výzkum ukázal, že žákům pomáhají při učení slovní zásoby a že díky užívání víceslovných spojení, která však produkují jako jednu naučenou lexikální jednotku, získávají čas potřebný pro plánování své promluvy. To má pozitivní dopad na plynulost jejich projevu. V neposlední řadě si při učení kolokací osvojují spojení, která jsou v jazyce rodilých mluvčích přirozená, což přispívá ke zvyšování jejich jazykové úrovně. Z těchto důvodů se informace o kolokacích stále častěji zařazují do jazykových učebnic a žákovských slovníků.

Užitečným nástrojem pro studenty a učitele, kteří chtějí pracovat s kolokacemi v anglickém jazyce, je nástroj *Just the Word,*[19] který umožňuje zadat jednotlivé slovo a nechat vypsat všechny jeho kolokace. Ty se generují z Britského národního korpusu (*British National Corpus*) a jsou rozděleny do kategorií dle typu spojení (např. adjektivum + substantivum, adverbium + adjektivum, sloveso + substantivum atp.). Nástroj je možné použít i jako pomůcku při psaní, jejímž prostřednictvím si v případě nejistoty můžeme ověřit, jaké spojení je časté, a tudíž přirozené, a můžeme si i vybrat z alternativ. Dozvídáme se přitom, kolikrát se zjištěná spojení v korpusu objevují, a pomocí odkazů si můžeme zobrazit i příslušné konkordanční (výsledkové)

19 www.just-the-word.com

stránky a prohlédnout si konkrétní příklady užití v korpusu. Nástroj tak lze velmi dobře využít i pro sestavování kolokačních cvičení a testů.

Mezi víceslovná spojení patří i frazémy (často též v kontextu vyučování angličtiny označované jako idiomy), jimiž rozumíme ustálená víceslovná spojení, jejichž význam není možné si odvodit z významů slov, ze kterých se frazém skládá (např. *kápnout božskou*, *spadnout z višně*, *natáhnout bačkory* atp.). Pro studenty cizích jazyků jde o náročnou součást slovní zásoby, kterou si mohou osvojit pouze přesným zapamatováním a musí navíc i vědět, kdy a v jakém kontextu lze tato spojení použít. Podobně neodvoditelná jsou v angličtině tzv. frázová slovesa, u nichž je význam slovesa v základu změněn či upraven částicí, která se s ním pojí (např. *take* = vzít, *take off* = *vzlétnout*).

Důležitost frazeologie ve výuce angličtiny zohledňuje i výše zmíněný internetový zdroj *English Vocabulary Profile*, který ve svém pokročilém vyhledávání umožňuje pracovat zvlášť s frázemi, frazémy a frázovými slovesy. Je možné si vygenerovat seznam všech slovních spojení z těchto skupin pro jednotlivé úrovně pokročilosti nebo je možné zadat libovolné slovo a nechat si vypsat, v jakých víceslovných spojeních se na příslušné úrovni vyskytuje. Této funkce mohou využít nejen žáci, ale i učitelé při plánování výuky či sestavování jazykových cvičení a testů.

I u kolokací a frazémů se setkáváme s tzv. faux amis, tedy zrádnými kolokacemi a zrádnými frazémy. Žáci je nejčastěji používají proto, že se domnívají, že stejné spojení jako v jejich mateřštině funguje i v jazyce cizím. Mohou tak doslova přeložit frazém, který v cílovém jazyce neexistuje (např. nelze doslovně přeložit do angličtiny český frazém *šplouchá ti na maják* jako *it splashes on your lighthouse*). Příkladem zrádné kolokace může být třeba pokus o překlad českého *silný kuřák* do angličtiny jako *strong smoker*; správné je totiž *heavy smoker*. Je pravděpodobné, že zrádnost by fungovala i v opačném směru, tedy pro Angličana učícího se češtinu, který by spojení *těžký kuřák* v češtině při neznalosti správné kolokace zřejmě použil.

Kromě kolokací a frazémů lze text rozparcelovat i na tzv. n-gramy, tedy sekvence zpravidla dvou a více slov, mezi nimiž nemusí být nutně nějaký lingvistický vztah. Takové n-gramy, které se v textech nejčastěji opakují, obvykle zahrnují kromě gramatických slov i slova plnovýznamová a vypovídají něco o daném jazyce, např. o jeho typických ustálených jednotkách nebo tendenci k formulaickému vyjadřování.

V žákovských korpusech se stejně jako u gramatiky a slovní zásoby chyby ve frazeologii značkují. V již zmíněném systému by pak, jak vidíme z věty označené Příklad 3, byla kolokační chyba označena značkou (LP), kde L označuje *lexical* a P pak *phrase* (fráze).

původní věta:	*he is a strong smoker*
anotovaná věta:	*he is a (LP) strong $heavy$ smoker*

Příklad 3: Ukázka chybové anotace anglické věty – frazeologie

Jeden typ výzkumu frazeologie v žákovském jazyce vychází z anotací chybovosti, při níž podobně jako u gramatiky a slovní zásoby lze s použitím anotovaných korpusů vyhledat všechny chybné výskyty. Řada anotátorů však chybné kolokace nechápe vždy jako chyby, ale spíše jako odchylky od běžné normy. Vycházejí z toho, že např. spojení *těžký kuřák* je spíše nepřirozené nežli úplně chybné. Výzkum frazeologie tak přispívá k pochopení vývoje přirozenosti žákovského jazyka a jeho přiblížení se k jazyku rodilých mluvčích. Taková srovnávání se uplatňují i v jiném typu výzkumu, na jehož základě je možné sledovat, jak časté jsou v žákovském textu a v textu rodilého mluvčího kolokace, případně n-gramy, a v čem jsou si tyto dvě jazykové variety blízké či naopak vzdálené. Je též nasnadě, že pokročilejší studenti si osvojili frazeologii cizího jazyka ve větší míře a s menší chybovostí, a výzkum tak umožňuje sledovat vývoj frazeologie u jednotlivých úrovní pokročilosti. Dalším specifickým typem výzkumu s pedagogickým dopadem je pak zkoumání frazeologie v různých žánrech. Toho se často využívá při výuce akademického jazyka a sestavování příslušných materiálů pro výuku.

2.3.4. Diskurz

Diskurzem rozumíme souvislý textový útvar, který přesahuje rámec jedné věty. U mluveného jazyka hovoříme obvykle o promluvách, u jazyka psaného pak o textech. Výzkum diskurzu je motivován přesvědčením lingvistů především v druhé polovině 20. století, že jazyk není dostatečné zkoumat jen s ohledem na nejnižší jazykové jednotky (např. hlásky, slabiky, morfémy, slova, fráze, věty), ale že je nutné vzít v potaz i takové aspekty, jakými jsou například struktura textu, význam v kontextu, role sociálního a kulturního kontextu na podobu textu, prostředky textové soudržnosti, informační struktura a obecně tedy veškeré jazykové i mimojazykové (nonverbální) prostředky, které s těmito aspekty souvisejí.

Zkoumání diskurzu v rámci korpusové lingvistiky je metodologicky zajímavé. Na jedné straně je komplikované tím, že výsledky vyhledávání v korpusu jsou prezentovány v úsečné podobě prostřednictvím konkordančních řádků, na druhé straně však korpusy díky své velikosti umožňují pracovat s obrovskými počty textů najednou a vyhledávat v nich jednotlivé diskurzní prvky v takovém rozsahu, že je možné výsledky zobecňovat a hledat tendence platné napříč texty.

Ukázkovým příkladem využití korpusu v tomto směru je významná gramatika anglického jazyka *Longman Grammar of Spoken and Written English.*[20] Vychází z korpusu *Longman Spoken and Written Corpus*, který obsahuje 40 milionů slov z 37 244 textů čtyř základních typů: konverzace, beletrie, žurnalistika a akademický text. Na jeho základě byl velmi přesně popsán například jazyk konverzace: bylo shledáno, že je bohatý na nevětné struktury, neúplné věty (vynechávají se například slova s nízkou informační hodnotou) a ustálená víceslovná spojení, která plní různé funkce (např. mluvčí indikuje, že bude pokračovat v promluvě, že se vrací k původnímu tématu atp.). Obsahuje zároveň i tzv. prostředky řečového managementu, které mluvčímu umožňují hovořit plynule. Sem patří např. výplňková slova a výplňkové pauzy, jejichž funkcí může být získání času pro plánování další promluvy nebo indikace, že mluvčí ještě současnou promluvu nedokončil. Tento výzkum inspiroval didaktiku cizích jazyků k tomu, aby se zaměřila nejen na tradiční výuku gramatiky založenou na psaném jazyce, ale i na jazykové prostředky charakteristické právě pro jazyk mluvený. Žákovská korpusová lingvistika pak v této oblasti zkoumá, do jaké míry tyto prostředky žáci využívají a jaké prostředky se naopak v jejich projevech vyskytují omezeně.

Dalším z ústředních témat je výzkum textové koherence. Jako koherenci označujeme sémantické, obsahové vztahy uvnitř diskurzu, které zajišťují jeho soudržnost, smysluplnost a plynulost. Mezi jednotlivými částmi koherentního textu totiž musí být patrná souvislost. Toho je možné dosáhnout používáním kohezních prvků, které explicitně naznačují, jak jsou jednotlivé úseky textu propojeny. Mezi tyto prvky patří například reference (odkazování na to, co již bylo řečeno, nebo na to, co bude v textu následovat), konektory (spojovací výrazy vyjadřující logický vztah mezi jednotlivými částmi), substituce (použití jiného slova, které jasně odkazuje na to, co již bylo vyřčeno) a tematicko-rematická struktura textu (tzv. aktuální členění větné). Výzkum koherence za použití korpusů je komplikovaný, neboť korpus neumožňuje zobrazit dlouhé úseky textu. Nejčastěji se proto pracuje právě s explicitními kohezními prvky (obzvláště konektory), které je možno v korpusu vyhledávat. V žákovském jazyce se pak často zjišťuje, jak žáci tyto prostředky dokážou využít a v jakých funkcích a jak jejich prostřednictvím vytvářejí strukturu textu. To má význam především při studiu a výuce akademického psaní. I v této oblasti se využívají kontrastivní studie, které srovnávají texty žákovské s texty rodilých mluvčích.

Pro vědecký text je též typické využívání rétorických prostředků pro zeslabování jistoty tvrzení (tzv. hedge) nebo jeho zesilování (tzv. booster). Ty se realizují použitím různých prostředků, např. modálních sloves (např. *vý-*

20 Biber, D., Stigg, J., Leech, G., Conrad, C. a E. Finegan. (1999) Longman Grammar of Spoken and Written English. Harlow: Pearson Education Limited.

sledky mohou naznačovat, že…), adverbií (např. *tento výsledek nejspíš vypovídá o…*) aj. I toto je oblast, kterou si nerodilý mluvčí píšící vědecký text v cizím jazyce musí osvojit. V žákovských korpusech akademického jazyka pak můžeme zkoumat, do jaké míry se to daří a v čem spočívají odlišnosti od textů rodilých mluvčích či kompetentnějších autorů. Častým zjištěním je, že méně zkušení autoři nadužívají určitých prostředků modalizace vědeckého textu a nevyužívají dostatečně jiných možností.

V neposlední řadě lze žákovský text korpusově analyzovat i z pohledu zastoupení osobních zájmen a činného či trpného rodu a zjistit tak, do jaké míry dokáže žák prezentovat objektivní, neosobní vědecký text a jaké prostředky dokáže využít, když jako autor do textu potřebuje vstoupit například prezentováním svého postoje.

2.3.5. Pragmatika

Pragmatika studuje jazyk z pohledu mluvčího a příjemce a také to, jakým způsobem tito uživatelé jazyka dosahují svých komunikačních záměrů. Zkoumá tedy, jaký záměr má mluvčí, když něco sděluje, a jak je jeho sdělení interpretováno příjemcem. To závisí na takových faktorech, jako je například kontext promluvy či vztah mezi mluvčím a příjemcem, roli hraje i příslušná jazyková kultura. Tyto faktory pak spolu s významem slov ovlivňují význam a dopad celého sdělení. Podle teorie řečových aktů, která stojí v centru pozornosti pragmatiky, je možné promluvou nejen něco prostě sdělit, ale i něco chtít, k něčemu se zavázat, k něčemu vyjádřit svůj postoj, případně vyřčením způsobit nějakou změnu. V žákovském jazyce je pak možné tyto řečové akty identifikovat a zjistit, jaké prostředky pro jejich vyjádření mluvčí používá. I zde je výhoda korpusů především ve zpřístupnění velkého objemu dat k výzkumu, který může být navíc prohledáván automatickými algoritmy za účelem identifikovat vzorce, které se v textu objevují. K diskurzní a pragmatické analýze korpusů musejí být korpusy anotovány. To je však značně časově náročný a ne vždy příliš snadný úkol, především vzhledem k nutnosti správně vystihnout a interpretovat původní záměr mluvčího.[21] Zkoumají se tak například funkce tzv. diskurzních částic, což jsou prostředky, které umožňují v rozhovoru udržovat kontakt mezi mluvčími (např. angl. *you know* nebo české *víte*) a jsou v korpusech snadno vyhledatelné. Výzkum též dokládá, že užívání diskurzních částic přispívá ke zvýšení plynulosti mluveného projevu a zvyšuje jeho přirozenost.

Z hlediska funkčního se značkují například korpusy zabývající se výzkumem specifických textových žánrů a útvarů. Existují např. korpusy

21 Jeden z příkladů pragmatické anotace korpusu je nástroj DART, více viz např. Weisser, M. (2018) How to Do Corpus Pragmatics on Pragmatically Annotated Data Speech acts and beyond. Studies in Corpus Linguistics: John Benjamins.

abstraktů (tj. krátkých shrnutí akademického textu). V nich jsou identifikovány a označkovány příslušné rétorické kroky (např. zasazení tématu do kontextu, stanovení cíle studie atp.), následně pak mohou být tyto rétorické postupy z korpusu extrahovány a je možné zkoumat, jakými jazykovými prostředky jsou realizovány. Žákovský jazyk se zde znovu obvykle srovnává s jazykem rodilých mluvčích, ale provádějí se i srovnání mezi různými vědeckými obory ve snaze zefektivnit tento typ psané komunikace.

Studium pragmatiky ukazuje, že žák cizího jazyka musí kromě gramatiky a slovní zásoby ovládnout i pravidla komunikace, která jsou běžná pro sociokulturní prostředí, v němž se tento jazyk používá. Nejde jen o to, aby žák dosáhl svého komunikačního cíle, ale také aby svým jazykovým projevem neporušoval běžné (např. zdvořilostní) normy a dokázal pro svá sdělení volit prostředky, které odpovídají konkrétní situaci, v níž se komunikace odehrává. Jakkoliv je tato oblast důležitá pro výzkum i výuku, korpusové studie v této oblasti jsou zatím omezené.

2.3.6. Plynulost

S nástupem korpusů zachycujících mluvený jazyk, jež jsou s ohledem na nutnost nahrát a přepsat texty náročnější na sestavení než korpusy psané, se korpusově začínají zkoumat i charakteristiky řeči. Jednou z takových oblastí je výzkum plynulosti. Plynulost je obtížně definovatelný pojem, pod kterým si však nejsnáze představíme produkci řeči v dostatečném tempu a bez velkého množství prvků, které bezprostředně nesouvisejí s významem sdělení, jako jsou např. různé projevy váhání.

Tempo mluvy lze v korpusech měřit pouze tehdy, pokud jsou dostupné i původní nahrávky. Uvádí se obvykle ve slovech za minutu či slabikách za vteřinu. U pokročilých studentů angličtiny zjišťujeme průměrné tempo mluvy asi 150 slov za minutu, u rodilých mluvčích pak více než 180 slov za minutu.

Mezi jevy, které tzv. narušují plynulost projevu, se řadí kupříkladu vyplněné a nevyplněné (tiché) pauzy (např. *byl jsem eh unavený*), opakované segmenty (např. *nevím co se co se přesně stalo*), nedokončené segmenty (např. *byl jsem šel jsem tam pěšky*), opravy (např. *vzala mě nás na výlet*), fragmentovaná slova (např. *to nemu nemuselo být*) a prodlužování slabik (např. *řekl mi žeee*). Také tyto jevy se v korpusech značkují. Pak je možné zjistit nejen jejich frekvenci, ale i to, na jakém místě se v promluvě vyskytují. Jedním z rozdílů mezi žákovským jazykem a jazykem rodilého mluvčího je to, že žáci častěji umísťují tyto prvky mezi jednotlivé části frází, které rodilí mluvčí zpravidla vyřknou najednou (např. *velký eh význam má*). Vypovídá to zřejmě o tom, že rodilí mluvčí jsou schopni rychleji plánovat promluvu a produkují tak delší celky a zároveň váhají spíše na jejich začátcích než v jejich průběhu.

Všechny tyto prvky lze srovnávat na různých úrovních pokročilosti a také mezi žáky a rodilými mluvčími. Zkoumat lze i to, jaký efekt má na plynulost a její součásti typ úlohy, který žáci řešili (např. šlo-li o monolog, dialog, projev s předchozí přípravou, popis obrázku atp.). Tento výzkum pak krom studia pokročilosti a rozdílů mezi žákovským a rodilým jazykem přispívá i k poznání, jakým způsobem je produkována řeč. V didaktice má tento výzkum význam především pro oblast testování, neboť poskytuje popis ukazatelů charakteristických pro plynulost projevu, což se u ústních jazykových zkoušek nezřídka hodnotí a zároveň je i jedním z parametrů popisovaných Společným evropským referenčním rámcem.

V této kapitole jsme se pokusili nastínit, co jsou akviziční korpusy a jak se využívají při výzkumu žákovského jazyka za účelem jeho poznání a navržení možných didaktických postupů. Akviziční korpusy umožňují zkoumat žákovský jazyk z různých pohledů, srovnávat, jak se mění na jednotlivých úrovních pokročilosti a jaké jazykové prostředky a projevy tyto úrovně charakterizují. Dále umožňují srovnávání s jazykem rodilých mluvčích jako přirozeným cílem velké části žáků cizích jazyků a v neposlední řadě poskytují i empiricky ověřená zjištění využitelná při tvorbě jazykových materiálů (učebnic, slovníků, aplikací), jazykových testů a didaktických postupů. Jakkoliv je žákovská korpusová lingvistika oborem mladým, přinesl již výzkum ve všech těchto oblastech cenné výsledky a především prokazuje, že akviziční korpusy mají velký výzkumný i pedagogický potenciál, z něhož mohou čerpat nejen badatelé, ale i učitelé, studenti a tvůrci jazykových testů a sylabů.

3. Významné parametry korpusů

U korpusů obecných se předpokládá, že směřují k naplnění určitých závazných parametrů. Jde zejména o velikost a vyváženost korpusu a o autentičnost jazykových dat, která korpus zahrnuje. Uvedené tři parametry zajišťují jeho reprezentativnost ve vztahu k normálnímu, obvyklému jazykovému úzu.

Dalším důležitým parametrem je soustavné zaznamenávání informací o původu jednotlivých složek (textů), které jsou v něm zařazeny, o jejich autorech, žánru a dalších údajů relevantních pro práci s korpusem. Tyto metatextové informace jsou přiřazeny všem strukturním jednotkám korpusu formou strukturních atributů.

K uvedeným parametrům přistupuje lingvistické značkování (lingvistická anotace), tj. připojování lingvistické interpretace k jednotlivým tokenům (výskytům, textovým pozicím) formou pozičních značek. Každé slovo tedy mívá přidělený základní slovníkový tvar a dále značku, obvykle s morfologickými či méně často se syntaktickými informacemi.

Korpusy speciální se vyznačují tím, že některý či některé z těchto parametrů porušují, nenaplňují, případně ve snaze o jejich naplnění postupují specifickým způsobem, některé parametry rozšiřují (strukturní atributy, lingvistické značkování) apod.[22]

3.1. Velikost

Obecné korpusy usilují o maximální možnou velikost. Čím větší objem jazykového materiálu zahrnují, tím větší je pravděpodobnost, že odrážejí reálný jazykový úzus a že budou využitelné i u jevů méně frekventovaných. Větší velikost rovněž umožňuje, aby byly v korpusu zahrnuty všechny relevantní variety daného jazyka v odpovídajících proporcích, tedy aby byl vyvážený a aby bylo možné považovat ho za reprezentativní vůči normálnímu, ob-

22 K tematice obecných korpusů a korpusové lingvistiky podrobněji např. Mc Enery - Hardie (2012), k akvizičním korpusům, zvl. pro češtinu, Šebesta (2010), Šebesta - Škodová (2012), Štindlová (2011), k aktuálním datům o nabídce korpusů pro češtinu a korpusů žákovských https://wiki.korpus.cz/doku.php/cnk:uvod, resp. Learner corpora around the world; z těchto zdrojů tato i první kapitola převážně čerpají.

vyklému užívání jazyka. Díky tomu mohou tyto korpusy dobře sloužit pro tvorbu slovníků, gramatik, thesaurů a dalších referenčních knih.

Velikost synchronních korpusů v posledních desetiletích rychle roste. Zatímco v počátcích korpusové lingvistiky, v 60. a 70. letech minulého století, se velikost korpusů uváděla v milionech tokenů (konkrétní výskyt každého slovního tvaru v textu),[23] u dnešních korpusů se uvádí v miliardách. Např. Český národní korpus jako zastřešující projekt zajišťující budování korpusů českého jazyka na Univerzitě Karlově zahrnuje k datu vzniku tohoto textu relativně nově český synchronní korpus *SYN* verze 6 o velikosti cca 4 miliardy slov, patří tedy mezi tzv. korpusy velké (pohybující se v řádu stovek milionů až v miliardách textových slov). Vedle toho rozlišujeme dnes ještě korpusy střední (obsahující řádově desítky milionů textových slov) a korpusy malé (pohybující se v řádu stovek tisíc textových slov).

I sebevětší korpus je ovšem menší než web – největší existující a volně dostupný soubor textů. I web lze využívat jako korpus; byla vyvinuta rozhraní, která k využití webu jako korpusu slouží – např. *WebCorp*.[24] Web jako korpus má některá specifika, která se mohou jevit jako výhody, nebo nevýhody, v závislosti na cíli, který při vyhledávání sledujeme. Většinou je nepochybnou výhodou webu jako korpusu jeho velikost. Některé výrazy či kolokace, které jsou málo frekventované, mohou v obyčejném korpusu úplně chybět nebo mít jen mizivé zastoupení, totéž platí i o některých druzích málo frekventovaných pravopisných či jiných formálních odchylek či variant, málo frekventovaných typů užití apod., které na webu – vzhledem k jeho velikosti – můžeme najít s vyšší úspěšností než v korpusu.

Hlavním problémem webu jako korpusu je skutečnost, že nám podává jazykový materiál v nerozlišené směsi textů různých typů, není tedy možné posoudit ani jeho skladbu, ani kontext užití hledaných jevů. To se může jevit jako značná potíž např. v případech, kdy hledáme velmi frekventované jazykové jevy a potřebujeme počet nalezených řádků redukovat podle nějakých lingvisticky relevantních kritérií. Dalším problémem je nestálost webu – web se neustále a průběžně mění, výzkum, který na něm byl založen, tedy není možné replikovat nebo jen velmi obtížně. Výhodou standardního obecného korpusu ve srovnání s webem je tedy jeho promyšlené složení z různých typů textů v náležitých proporcích, větší nebo menší míra jeho vyváženosti, jeho relativní ustálenost a zaznamenávání metatextových informací.

Žákovské korpusy se co do velikosti od korpusů obecných výrazně liší, přesněji řečeno, jsou v tomto ohledu podstatně pestřejší. Žákovské korpusy

23 Už zmiňovaný *Brown Corpus* obsahoval jeden milion slov, stejnou velikost měl i rovněž zmiňovaný korpus britské angličtiny *LOB corpus*.

24 http://www.webcorp.org.uk/live/

s objemem několika (desítek) milionů textových slov najdeme v současnosti pouze u angličtiny jako cizího jazyka. Dosud největší známý korpus je *Chungdahm Corpus*, tvořený eseji, které napsali korejští studenti angličtiny na Chungdahm institutu, národním řetězci škol anglického jazyka, a chybově anotovaný tutory. Eseje jsou uloženy jako stále doplňovaná databáze, nikoli korpus; z této databáze byl vyčleněn pro výzkumné účely velký objem dat jako *Chungdahm Corpus* o velikosti přibližně 131 milionů slov (přesněji 130 754 tisíc slov) ve více než 860 esejích o průměrné délce 152 slov (Han et al. 2010).

Korpus této velikosti je však i u angličtiny spíše výjimkou. Co do velikosti následují *The Cambridge Learner Corpus* obsahující cca 50 milionů textových slov (také ten je založen na esejích napsaných v rámci zkoušek angličtiny) a *HKUST* (*The Hong Kong University of Science & Technology Learner Corpus*) o velikosti cca 25 milionů slov. Ostatní korpusy angličtiny i korpusy dalších jazyků se pohybují v podstatně nižších hodnotách – jen výjimečně dosahují hodnoty 2 milionů slov nebo ji překračují.

Mezi velikostně nejmenšími uváděnými korpusy jsou např. *The Pilot Arabic Learner Corpus*, korpus arabštiny užívané mluvčími angličtiny jako prvního jazyka, obsahující psané vyprávěcí texty mírně pokročilých a pokročilých studentů arabštiny o celkovém rozsahu 9 000 slov, nebo The *PIKUST*, korpus slovinštiny jako cizího jazyka, který obsahuje texty psané mluvčími 18 různých prvních jazyků, většinou chorvatštiny, srbštiny a ruštiny, o celkovém rozsahu 35 tisíc textových slov.

Velikost žákovských korpusů sice postupně roste, podobně jako velikost korpusů obecných, ale podstatně nižším tempem. Menší objem žákovských korpusů je dán obtížností sběru dat a finanční i časovou náročností jejich přepisu a dalšího zpracování.

První problém je spojen se skutečností, že je objem řečové produkce nerodilých mluvčích neporovnatelně menší než objem dat produkovaných mluvčími rodilými. Navíc se jedná o produkci obtížně dostupnou – pokud jde např. o písemné práce psané pro školu, disponuje jimi škola a nemusí je pro potřeby korpusu zpřístupnit. U mluvených projevů se sběry setkávají s podobnými problémy jako sběry pro mluvené korpusy jazyka rodilých mluvčích, jen s tím rozdílem, že projev nerodilého mluvčího bývá mnohonásobně kratší než projev rodilého mluvčího a celkový objem mluvených projevů nerodilých mluvčích je rovněž neporovnatelně menší.

K obtížnostem sběru přistupuje fakt, že projevy nerodilých mluvčích se sbírají v podobě nahrávek mluvených projevů nebo v podobě rukopisů, v obou případech se tedy musí texty přepisovat. Zatímco u mluvených projevů lze dnes, jde-li o kvalitní nahrávky standardní mluvy, využít automatického přepisu, u nerodilých mluvčích jde většinou o nestandardní jazyková data, která automatické zpracování v dobré kvalitě neumožňují, přepisuje se tedy manuálně, příp. se automatický přepis následně upravuje.

Je příznačné, že větších velikostí dosahují ty žákovské korpusy, které se mohou opřít o spolupráci se školami nebo které jsou přímo s jejich činností spojeny – srov. výše uvedený *Chungdahm Corpus*, zpracovávající produkci žáků řetězce škol Chungdahm, *Cambridge Learner Corpus*, rovněž obsahující produkci žáků při zkouškách, nebo *HKUST*, obsahující písemné práce vzniklé v rámci maturitních zkoušek z angličtiny. Totéž platí i o větších žákovských korpusech mluvených, jako je např. mluvená složka korpusu *BICCEL* (*Bilingual Corpus of Chinese English Learners*), která vychází z nahrávek při národním testu mluvené angličtiny, korpus *NICT JLE* (*Japanese Learner English*), také čerpající z testů mluvené angličtiny, nebo korpus *SWECCL* (*Spoken and Written English Corpus of Chinese Learners*).

České žákovské korpusy řady *CZESL* patří svou velikostí v porovnání s jinými jazyky kromě angličtiny, která má v tomto ohledu mimořádné postavení, ke korpusům standardním, či spíše větším.

Pro badatelské i pedagogické účely se z korpusů této řady nejspíše nabízí automaticky emendovaný, anotovaný (také chybově) a lemmatizovaný korpus *CZESL-SGT*. Zahrnuje 8617 textů od 1965 různých autorů (to znamená, že od jednoho autora jsou v korpusu zařazeny v průměru zhruba čtyři texty). Objem jazykových dat činí 1148 tisíc tokenů (pozic), z toho 958 tisíc slov a 111 tisíc vět. Druhým korpusem téže řady s provedenou emendací, anotací a lemmatizací je *CZESL-MAN*; od *CZESL-SGT* se liší tím, že byl emendován manuálně a chybová anotace byla provedena poloautomaticky s vysokým podílem manuálního zpracování. Manuální emendace a anotace je spolehlivější než emendace a anotace automatická, ale také je časově výrazně náročnější. *CZESL-MAN* je proto zatím podstatně menší než korpus *CZSL-SGT* (obsahuje cca 124 tisíc slov) a není volně přístupný.

3.2. Vyváženost

Vyváženost obecných korpusů je vlastnost, o niž se tvůrci korpusů v nějaké míře snaží, míra jejího dosažení záleží ale na tom, jaká kritéria zvolí a jak se vyrovnají s obtížemi, které jsou s tím spojeny. Vzhledem k velmi variabilním možnostem využití korpusů není možné vyváženost korpusu opírat o vnitřní, lingvistická hlediska, ale o kritéria vnější.

Obecné korpusy v Českém národním korpusu se opíraly zpočátku o zřetel k recepci – ty typy textů, které byly častěji čteny větším počtem mluvčích daného jazyka, byly v korpusu zastoupeny ve větší proporci než typy textů, které se četly méně. Struktura typů textů a proporce jejich zastoupení v synchronních korpusech současné psané češtiny byly tedy zpočátku stanoveny na základě výzkumů čtenářské recepce.

V prvním korpusu řady *SYN*, *SYN2000*, tvořila většinu (dvě třetiny) textů publicistika, přibližně čtvrtinu texty odborné, zbytek texty imaginativní

(beletrie, především próza). V korpusu *SYN2005* se tyto proporce na základě nových výzkumů recepce změnily – publicistika představovala pouze 1/3 velikosti korpusu, odborná literatura 27 %, podíl beletrie vzrostl na 40 %; podobné proporce byly zvoleny pro korpus *SYN2010*. Korpus *SYN2015* toto rozdělení opustil, tři textové makrotypy (publicistika, odborná literatura a beletrie) jsou v něm zastoupeny rovným dílem. Všechny uvedené korpusy (včetně korpusu *SYN2015*) se pokládají za reprezentativní v tom smyslu, že zahrnují co nejširší spektrum různých typů veřejných psaných (tištěných) komunikátů, které jako celek reprezentují současnou psanou češtinu, nezachycují ji však v proporcích odpovídajících přesně jazykové populaci.[25]

V řadě *SYN* byly kromě toho zveřejněny rovněž korpusy české publicistiky (*SYN2006PUB, SYN2009PUB, SYN2013PUB*), které reprezentativní ve vztahu k obecnému jazykovému úzu nejsou (zahrnují pouze texty z oblasti publicistiky). Zatím poslední korpus *SYN* verze 6 zahrnuje všechny starší synchronní psané korpusy řady *SYN* včetně korpusů české publicistiky a také nový publicistický materiál, vzhledem k velké převaze publicistické složky ho tedy rovněž nelze označit za vyvážený a reprezentativní.

Vyváženost (a také velikost) a s nimi spojená reprezentativnost se řeší samostatně a dosahuje jiných hodnot u korpusů psaného a korpusů mluveného jazyka. Materiál pro psané korpusy se snáze získává (pro korpusy řady SYN získává tvůrce naprostou většinu materiálu, přibližně 90 %, přímo v elektronické podobě od nakladatelů a vydavatelů na základě smluv), zatímco získat materiál pro mluvené korpusy je obtížnější (data se sbírají formou audio či videonahrávek; větší podíl má běžná mluva) a náročné je i jeho další zpracování, především přepis. Proto se u mluvených korpusů pohybujeme v objemech podstatně nižších.

Pro češtinu je zatím největším mluveným korpusem *ORAL*, zahrnující dřívější korpusy řady *ORAL* (*ORAL2006, ORAL2008, ORAL2013*) a přepisy nových nahrávek (*ORAL-Z*). Přepisy nahrávek převážně neformálních rozhovorů, které ho tvoří, mají celkový objem 5,4 milionu textových slov, v porovnání se stamilionovými objemy psaných korpusů tedy podstatně menší. Korpus také není vyvážený z teritoriálního hlediska (podíl rozhovorů z Čech je vyšší).

Vyvážený co do relevantních sociologických kritérií (pohlaví, věk, vzdělání, oblast pobytu v dětství) a z tohoto hlediska reprezentativní je nejnovější mluvený korpus češtiny *ORTOFON* o objemu 1 milion textových slov.

S vyvážeností volně souvisí i volba komponent, z nichž se korpus skládá; někdejší *Brown Corpus* a další korpusy jím v tomto ohledu inspirované (*Lancaster-Oslo-Bergen Corpus*, *British National Corpus*) se skládaly ze vzorků textů o standardním rozsahu (*Brown Corpus* např. obsahoval 500 vzorků

25 http://wiki.korpus.cz/doku.php/cnk:syn2015

textů o 2000 slovech). Soudobé korpusy, včetně obecných korpusů ČNK, se skládají z celých textů.

Lze říci, že reprezentativnost korpusu ve vztahu k běžnému, normálnímu jazykovému úzu je hodnota, k níž obecné korpusy snahou o velikost a vyváženost směřují, i když jí nikdy plně nedosahují. Při práci s korpusem ovšem sledujeme často jen určitou specifickou oblast jazyka a jeho užívání, posuzujeme tedy přirozeně i reprezentativnost příslušného korpusu z hlediska jeho výzkumného cíle a podle populace textů, na niž je náš výzkum zaměřen.

U žákovských korpusů o vyváženost ve smyslu úplného a proporcionálního zastoupení jednotlivých variet jazyka nerodilých mluvčích neusilujeme. Je to dáno tím, že u cizího jazyka, jeho osvojování a užívání se setkáváme s neporovnatelně vyšší variabilitou než u jazyka prvního a tato variabilita je vázána na mimořádně vysoký počet významných proměnných, které při výzkumu i při výuce potřebujeme brát v úvahu. Přinejmenším za současných technických podmínek a při současné velikosti žákovských korpusů není možné uvažovat o vyváženosti jakéhokoli žákovského korpusu ve vztahu k celé populaci.

Není to ani cílem jejich tvůrců. Hlavním cílem žákovských korpusů v tomto směru je umožnit studium variability žákovského jazyka (mezijazyka) a jeho užívání a studium vztahu této variability k co největšímu počtu sledovaných a pečlivě zaznamenávaných proměnných. Při práci s žákovským korpusem tedy neočekáváme, že reprezentuje mezijazyk nerodilých mluvčích jako celek, to není dost dobře možné, ale že spolehlivě zaznamenává a umožňuje studovat užívání některé jeho variety ve vztahu k co největšímu počtu faktorů. Při posuzování kvality žákovského korpusu tak přihlížíme nejen k jeho velikosti, ale k tomu, zda obsahuje tu varietu mezijazyka žáků, která nás zajímá, a také k tomu, kolik faktorů (a které) ovlivňujících její užívání zaznamenává a jak podrobně. Nejčastěji to bývá první jazyk žáků (např. čeština žáků s prvním jazykem čínským) a úroveň ovládání cílového jazyka (zřídka jde o začátečníky, častěji o mírně pokročilé nebo pokročilé), téma (blízké, známé, citově zabarvené vs. vzdálené, neznámé, citově neutrální), žánr, způsob sběru apod.

Korpusy řady *CZESL*[26] byly budovány se snahou získat jazyková data v takové skladbě, aby v nich byly zastoupeny všechny úrovně ovládání jazyka, od začátečníků po pokročilé, a aby v repertoáru prvních jazyků studentů byly dostatečně zastoupeny tři jejich skupiny: jazyky slovanské, indoevropské neslovanské a jazyky neindoevropské, češtině vzdálené. Jednotlivé úrovně ovládání jazyka (podle SERR) a skupiny prvních jazyků nejsou v těchto korpusech zastoupeny rovnoměrně, korpusy tedy nejsou z těchto

26 *CZESL-PLAIN, CZESL-SGT, CZESL-MAN*.

hledisek vyvážené; pro studium variability užívání jazyka nerodilými mluvčími ve vazbě na tyto i další parametry je však lze velmi dobře využívat.

3.3. Autentičnost jazyka

Třetím významným parametrem obecných korpusů je autentičnost jazykových dat, která jsou v nich uložena. Jde – vedle vyváženého zastoupení všech relevantních variet daného jazyka a vedle dostatečné velikosti korpusu – o třetí důležitou podmínku pro to, abychom mohli s korpusem pracovat jako s reprezentativním vzorkem normálního, obecného jazykového úzu.

Za autentické se pokládají takové projevy, které vznikly z autentické komunikační potřeby rodilého mluvčího v obvyklé, standardní situaci, nikoli projevy, které byly uměle elicitovány např. pro potřeby lingvistického výzkumu. U obecných korpusů je autentičnost dat zajištěna tím, že jsou do nich zařazovány projevy vzniklé v reálných komunikačních situacích.

Trochu složitější je situace u korpusů mluveného jazyka. Diskutovaná je především hranice toho, co máme považovat za jazyk mluvený. Prototypicky mluvené jsou nepochybně neformální nepřipravené, spontánní projevy rodilých mluvčích, v praxi se však někdy do mluvených korpusů zařazují i projevy připravené a často stylizované, šířené v médiích.[27]

České mluvené korpusy řady *ORAL* i nový korpus *ORTOFON* se drží přísnějšího vymezení mluveného jazyka a mluvených korpusů – obsahují neformální spontánní projevy rodilých mluvčích. Od mluvených korpusů je účelné odlišovat, i když se to někdy nečiní, korpusy řečové, které jsou zaměřeny na zachycení zvukové stránky jazyka. Současné mluvené korpusy fungují částečně i jako korpusy řečové – nabízejí vedle ortografického přepisu mluveného projevu rovněž jeho přepis fonetický, původní zvukový záznam, někdy k tomu přistupuje i videozáznam, případně přepis gest.

Ke studiu zvukové stránky jazyka ovšem nestačí pouze záznam autentického (neformálního) mluveného projevu. Důležitou složkou korpusů řečových (speech corpora) jsou např. záznamy čtení speciálně pro tyto potřeby připravených textů, čtení izolovaných slov, jejich párů a skupin sestavených s ohledem na některé významné fonetické jevy. Zpravidla jde o záznamy nahrávané ve studiu, v experimentálních podmínkách. Řečové korpusy tohoto typu jsou tedy typické korpusy speciální, nikoli obecné.

Korpusy žákovské pracují s pojmem autentičnosti poněkud odlišně od toho, jak se chápe v korpusech obecných. Především nevnímají autentičnost jako hodnotu diskrétní, nýbrž skalární. Prosté rozlišení textů vzniklých ze skutečné komunikační potřeby ve standardní situaci a textů ostatních je pro potřeby výzkumu osvojování, vyučování i užívání jazyka žáky příliš

27 Pro jejich zařazování se vyslovuje např. J. Sinclair (Sinclair, 1996).

hrubé, protože nebere ohled na různé situační faktory, za nichž text vzniká a které jeho podobu i stupeň autenticity ovlivňují a které mohou být z badatelského i didaktického hlediska zajímavé.

Pracujeme-li se škálovým pojetím autenticity, máme na jednom pólu škály projevy autentické, tj. žákovo přirozené, učitelem neřízené a zřetelem k výuce neovlivňované vyjadřování, vzniklé z komunikační potřeby v autentické komunikační situaci (např. při nákupu, objednávání hotelu, při rozhovoru s kamarády v restauraci apod.). Na druhém pólu jsou projevy žáka zřejmě a výrazně řízené jinou osobou, většinou učitelem, a to s pozorností zaměřenou primárně na jazykovou formu sdělení (projevy experimentálně elicitované). Může jít např. o překlad do druhého jazyka, převádění vět do minulého času nebo množného čísla podle zadání apod.

Mezi oběma póly je velmi rozsáhlá oblast projevů, jejichž vznik byl rovněž podnícen a jeho zpracování ovlivněno jinou osobou než mluvčím (zpravidla učitelem, ale také rodičem, vychovatelem, badatelem, sběračem dat apod.), ale které vznikaly alespoň zčásti se zřetelem k funkci či obsahu sdělení (např. s cílem vyprávět kamarádovi obsah filmu, popsat obrázek tak, aby ho kamarád rozpoznal apod.), nikoli k jazykové formě. Tento typ projevů bývá označován jako projevy klinicky elicitované (Ellis – Barkhuizen 2005: 23). Za klinicky elicitované můžeme pokládat většinu prací vzniklých ve škole pro školní potřeby, především slohové práce (eseje) různého druhu, které bývají předmětem sběru pro korpusy.

Míra řízenosti u prací vzniklých ve školním kontextu je velmi různá. Typicky experimentální povahu mají různé gramatické transformace, ale také výsledky mluvních cvičení zaměřených na zvukovou stránku řeči. Experimentální povahu mají také korpusy (či části korpusů) obsahující přepisy nahrávek čtených textů, čtených seznamů izolovaných slov apod., které bývají do akvizičních, a zvláště žákovských korpusů mluveného jazyka zařazovány. Korpusy tohoto typu jsou přirozeně velmi užitečné pro výzkum i pro pedagogické využití. Tak např. řečový korpus *ISLE* (*Interactive Spoken Language Education*) obsahuje materiál z čtení jednoduchých vět a minimálních párů; řečový korpus angličtiny čínských žáků *ESCCL* (*English Speech Corpus of Chinese Learners*) zahrnuje čtené dialogy; korpus *LeaP* (*Learning Prosody in a Foreign Language*) zahrnuje jak materiál z čtení (izolovaných slov, krátkého příběhu), tak z převyprávění i volného rozhovoru; čtení (nepřipravené čtení většinou úryvků z beletrie) je základem korpusu *Learners' Corpus of Reading Texts*.

U klinicky elicitovaných prací hraje roli řada faktorů: vztah mezi iniciátorem a žákem (učitel, kamarád, neznámá osoba; věkový rozdíl; vzájemná známost apod.), povaha úkolu (součást jazykového testu, úkol zadaný se zřetelem k obsahu a funkci), přípravné aktivity, možnost využívat jazykové pomůcky, téma (blízké, známé, důvěrné vs. vzdálené, neznámé; žánr atd.).

Autentické projevy můžeme zaznamenávat např. v mimoškolním prostředí, v rodině. Pro korpusy žákovské je získání autentické jazykové produkce značně obtížné, zejména pokud sledujeme i nižší úrovně ovládání jazyka. Autentická produkce žáků na nejnižších úrovních ovládání jazyka je velmi omezená nejen kvantitativně, ale také funkčně, tematicky a žánrově. Většina známých žákovských korpusů je proto založena na sběrech dat klinicky nebo experimentálně elicitovaných.

Akviziční korpusy zaměřené na první jazyk, tedy korpusy jazyka dětí a mládeže předškolního a školního věku, se od korpusů žákovských v tomto ohledu poněkud odlišují. Pro studium jazyka mládeže je nezbytné získat její autentické neformální a spontánní projevy, vernakulární jazyk. Jeho sběr není snadný, proto se vedle nahrávek neformální komunikace v autentickém prostředí volí většinou různé jiné cesty, jak získat projevy blízké autentickým – např. stimulováním dialogu mezi vrstevníky o tématu, které je jim blízké a emotivní, aranžováním situace, která má typické rysy situace neformální, apod.

I u školní mládeže je ovšem podstatné studovat, jak si žáci osvojují tu varietu rodného jazyka, která je vyžadována a zprostředkovávána školou, jak zvládají písemný projev, jak si osvojují tzv. intelektuální lexikum apod., ani korpusy jazyka mládeže se tedy neomezují jen na projevy vernakulární.

Žákovská data v korpusech řady *CZESL* i data v akvizičních korpusech češtiny jako prvního jazyka (*SKRIPT2012*) jsou rovněž založena na klinicky elicitované produkci. Jde pravidelně o písemné práce psané ve školním kontextu, ale za různých podmínek, s různou mírou přípravy apod. Autentickým se blíží data v obou akvizičních korpusech češtiny mluvené, tedy *SCHOLA2010*[28] a zčásti *ROMi_1.0*[29].

Obecně lze říci, že převážná většina akvizičních a žákovských korpusů je téměř zcela založena na jazykovém materiálu získaném z projevů klinicky elicitovaných, v omezenější míře elicitovaných experimentálně a jen velmi omezeně (především u korpusů zachycujících vernakulární jazyk mládeže) z projevů autentických nebo autentickým projevům se blížících. Aby bylo možné posoudit míru a povahu řízenosti těchto projevů, je nezbytné při budování akvizičních korpusů zaznamenávat všechny relevantní charakteristiky textu, mluvčích, situace produkce a způsobu sběru, které by mohly charakter jazyka v korpusu ovlivnit, a uvádět je v korpusu jako strukturní atributy mnohem podrobněji, než je obvyklé u korpusů obecných.

28 Jde o přepisy autentických vyučovacích hodin, které zahrnují záznam autentické mluvené produkce žáků ve škole.

29 Nahrávky promluv dětí ze sociokulturně znevýhodňujícího prostředí; ani ty ovšem nejsou autentické plně.

3.4. Strukturní atributy

Korpusy jsou vnitřně strukturovány – člení se na menší jednotky; u psaného korpusu *SYN2015* a psaných korpusů pozdějších to jsou dokumenty, dále texty (pokud je dokumentem např. číslo novin či sbírka fejetonů), ty se dělí na odstavce a odstavce na jednotlivé věty, resp. větné celky (od počátečního do koncového signálu). (Při tokenizaci se věty dále člení na jednotlivé tokeny – výskyty, pozice.) K strukturním jednotkám jsou v korpusu přiřazeny strukturní atributy, které nesou důležité metatextové informace (metadata); podle nich můžeme v korpusu cíleně vyhledávat a vyhodnocovat výsledky.

Příkladem strukturních atributů psaných korpusů pro češtinu nám může být korpus *SYN2015*. U tohoto korpusu se pracuje kromě jednoznačného identifikátoru dokumentu celkem s 15 strukturními atributy. Většinou slouží:

a. identifikaci původu dokumentu: název dokumentu nebo periodika, autor, vydání (u periodik), vydavatel, místo vydání, rok vydání, překladatel, zdrojový jazyk;
b. podání informace o druhu dokumentu: skupina textových typů (beletrie, oborová literatura, publicistika), textový typ (u beletrie např. próza, kratší próza, drama, poezie; u oborové literatury odborná, populárně naučná, profesní, memoáry a autobiografie, administrativa; u publicistiky publicistika tradiční a volnočasová), skupina oborů/témat, žánr/oblast, médium dokumentu (kniha, časopis, noviny, jiná tiskovina, referenční příručka, učební materiál), periodicita;
c. podání informace o cílovém adresátovi – jde o odlišení dokumentů určených dětskému čtenáři od ostatních.

Strukturní členění korpusů mluvených je poněkud odlišné a odlišné jsou i strukturní atributy.

Pro akviziční korpusy, prvního i druhého/cizího jazyka, jsou metatextové informace mimořádně důležité. Dovolují totiž sledovat, jak jednotlivé vnější faktory (spojené se vznikem textu, s jeho autorem, se způsobem sběru atd.) působí na volbu variety jazyka a na jeho užívání, jak ovlivňují řečové chování mluvčích.

Strukturní atributy u akvizičních korpusů jsou proto výrazně bohatší a detailnější. Bohatství této dokumentace je podstatným měřítkem kvality žákovského, resp. šíře akvizičního korpusu a zvyšuje či limituje jeho využitelnost.

Zajímavostí akvizičních korpusů je to, že obvykle obsahují metadata spojená s žákem (autorem textu). Např. v korpusu *CZESL-SGT* najdeme údaje trojího typu: (a) o osobě mluvčího/pisatele, např. o jeho věku, pohlaví, (b) o je-

ho prvním jazyce a jazykovém zázemí vůbec (první jazyk, skupina jazyků, další jazyky, úroveň ovládání češtiny podle SERR), (c) o podmínkách jeho studia češtiny jako cizího jazyka (možnost seznámení s jazykem v rodině, pobyt v ČR, délka, intenzita a povaha kurzu češtiny až po údaj o užívané učebnici).

Strukturní atributy korpusů *CZESL-SGT* i dalších žákovských a akvizičních korpusů umožňují sledovat vliv velmi širokého spektra faktorů na osvojování a užívání jazyka; dosavadní výzkumy tyto možnosti ještě ani zdaleka nevyužívají.[30]

3.5. Poziční atributy – lingvistická (a chybová) anotace

Lingvistické či jazykové značkování spočívá v připojování lingvistické interpretace formou tzv. pozičních atributů k existujícímu korpusu psaného nebo mluveného jazyka, přesněji k jednotlivým korpusovým pozicím, tokenům. Korpusy se v rozsahu a povaze této anotace odlišují a přirozeně se tato anotace také vyvíjí.

Vlastnímu značkování předchází rozčlenění textu, a sice: (a) explicitní segmentace textu do vět (značených počáteční značkou <s> a koncovou </s>); jde o větné celky vymezené počátečním a koncovým signálem, tedy věty jednoduché i souvětí; (b) tokenizace, rozdělení textu na tokeny (korpusové pozice).

Značkování probíhá tak, že je každému tokenu přiřazeno lemma (reprezentativní slovníková podoba) a morfologická značka, tag. Slouží k tomu dva kroky: (a) morfologická analýza (přiřazení všech teoreticky možných morfologických a slovnědruhových interpretací každé korpusové pozici bez ohledu na kontext), (b) disambiguace (výběr správné morfologické a slovnědruhové interpretace na základě kontextu a vyloučení ostatních, pokud ovšem situace nevyžaduje zachování většího počtu možností).

Například ve větě *Hráli házenou* dospějeme u tokenu *házenou* morfologickou analýzou ke čtyřem alternativním možnostem (s přiřazenými lemmaty, slovnědruhovou a morfologickou charakteristikou; zde uvádíme zkráceně pouze základní určení):

- substantivum v akuzativu singuláru,
- substantivum v instrumentálu singuláru,
- adjektivum v akuzativu singuláru,
- adjektivum v instrumentálu singuláru.

30 Srov. vyjádření S. Grangerové (Granger 2009: 17): *One must admit, however, that this facility is still seldom used and LC researchers (myself included) have had a tendency to base their analysis on the whole corpus or on subcorpora distinguished only on the basis of the learners' mother tongue. In fact, a properly coded learner corpus makes it possible for researchers to study the effect of a much wider range of variables.*

Disambiguace pak tři z těchto možností vyloučí, ponechá pouze správnou variantu (substantivum v akuzativu singuláru). Morfologická analýza i disambiguace se provádí s využitím speciálních programů; jde ale o proces velmi náročný, zejména u syntetických jazyků s bohatou flexí a homonymií, včetně homonymie gramatické, jako je čeština.

Lemmatizace a morfologické značkování představují základní typy lingvistické výbavy korpusů. K nim přistupuje dnes jako třetí anotace syntaktická, která byla pro češtinu vyvinuta v rámci *Pražského závislostního korpusu* s využitím domácí, české syntaktické tradice. Díky tomu se mohou na bázi syntakticky anotovaných korpusů vyvíjet i programy pro potřeby školní výuky. V současné době využívají syntaktickou anotaci rovněž synchronní korpusy budované v Českém národním korpusu; prvním syntakticky anotovaným je korpus *SYN2015*.

Všechny uvedené typy lingvistické anotace (i další, zde nezmíněné, jako je anotace fonetická, fonologická či sémantická) mohou být velmi užitečné i v akvizičních korpusech. Zejména u žákovských korpusů je ovšem lingvistické značkování velmi svízelné. Žákovský jazyk totiž zahrnuje ve vysoké míře nestandardní jazyková data, nelze tedy na ně bez předběžné úpravy aplikovat nástroje pro automatické zpracování. Před lingvistickou anotací je tudíž potřeba korpusy upravit tak, aby bylo automatické zpracování umožněno, tj. opravit odchylky od obvyklého jazykového úzu (emendovat je). Emendace se provádí manuálně (taková úprava je spolehlivá, ale časově i finančně velmi náročná), automaticky nebo semiautomaticky.

Na emendaci pak někdy navazuje kromě standardní lingvistické anotace rovněž anotace, která byla vyvinuta specificky pro účely akvizičních, zvl. žákovských korpusů, totiž anotace chybová, tedy přiřazení informací o typu chyby.[31] Termínem chyba se zde označuje odchylka od obvyklého úzu rodilého mluvčího nebo od jeho jazykové intuice. V literatuře i v praxi se setkáváme s různými typologiemi jazykových chyb, založenými na různých kritériích o různé jemnosti. Jen relativně malý počet žákovských korpusů je dnes chybovým značkováním vybaven; volí se přitom různě detailní strategie – někdy jsou chyby pouze vyznačeny, emendovány, jindy též pojmenovány (zařazeny do určité kategorie), případně hodnoceny atd.

Z korpusů řady *CZESL* jsou emendovány, lemmatizovány a otagovány, včetně tagování chybového, *CZESL-SGT* a *CZESL-MAN*. *CZESL-SGT* byl emendován automaticky s pomocí specifických programů pro kontrolu pravopisu a gramatiky; vznikly tak dvě vrstvy, původní a emendovaná, slovům v obou těchto vrstvách byla přiřazena, opět automaticky, lemmata a tagy (slovnědruhové a morfologické značky), dále pak údaje o chybě. *CZESL-MAN* pro-

31 Podrobněji k chybové anotaci žákovských korpusů např. Štindlová, 2011.

šel emendací manuální a návazně semiautomatickou anotací, rovněž včetně anotace chybové.

Při hledání v korpusu *CZESL-SGT* můžeme vyjít z původního tvaru (word, k němu je vázáno lemma, tag; k typům vyhledávání viz dále) nebo z tvaru emendovaného (word1, lemma1, tag1), můžeme hledat typy chyb (err) apod. Zadáme-li např. word1[32] slova *miluje*, zobrazí se nám na pozici word jak užité tvary standardní, tak i tvary chybové (míluje, milujút).

Je potřeba upozornit, že u automatické emendace a anotace musíme počítat s nepřesnostmi a omyly, a to jak ve vlastní emendaci, tak v anotaci. Pokud jsme si toho ale vědomi, může nám korpus *CZESL-SGT* posloužit jako mimořádně bohatý zdroj dat pro výzkum i výuku.

32 Viz dále.

4. Práce s korpusem

V následující kapitole se seznámíme s několika korpusy češtiny a se základními způsoby vyhledávání v nich.

4.1. Český národní korpus (ČNK)

4.1.1. Materiál/Obsah

Český národní korpus je projekt, který má za úkol vytváření dílčích korpusů češtiny. Obsahuje především korpusy psaného jazyka, ale také několik korpusů mluvené češtiny. V souhrnu se jedná o nejobsáhlejší jazykový materiál, který je pro češtinu zpracován. Psané korpusy lze dělit na korpusy synchronní, které dokládají současný jazykový úzus, a korpusy diachronní, které zachycují jazyk v jeho vývoji. Nezanedbatelný je také mnohojazyčný paralelní korpus *InterCorp*, který nabízí zarovnané texty (překlady textů) z téměř 40 jazyků. Kompletní přehled všech dostupných korpusů vč. jejich charakteristiky najdete na webu Českého národního korpusu.[33]

Pro výuku žáků s OMJ jsou relevantní především reprezentativní korpusy současné češtiny, např. SYN2015, nebo akviziční korpusy, např. žákovský korpus CZESL. V závislosti na cíli výuky a po zvážení jazykového původu žáků lze využít také texty paralelního korpusu InterCorp.

4.1.2. Vyhledávač neboli korpusový manažer

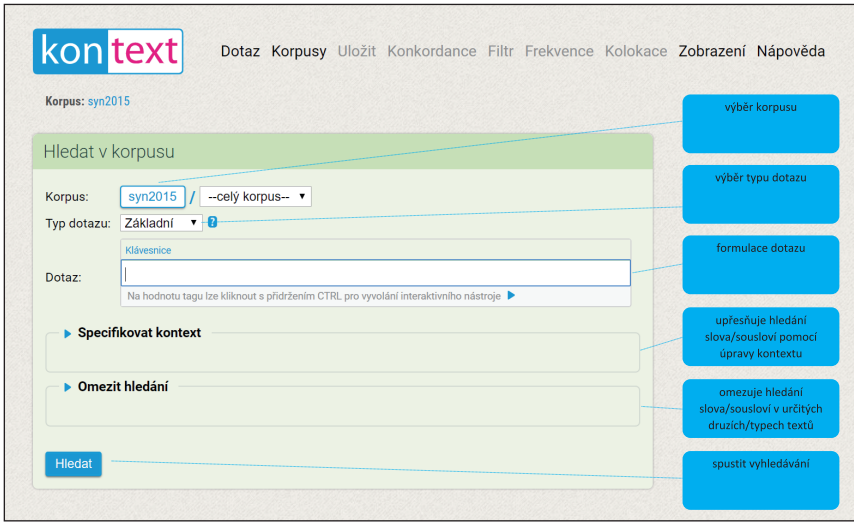
ČNK využívá korpusový manažer KonText, dostupný z webové stránky ČNK www.korpus.cz. KonText je umístěn v horním menu vlevo. Lze jej využívat i bez registrace, po bezplatném zaregistrování je ale k dispozici více funkcí.

33 https://wiki.korpus.cz/doku.php/cnk:uvod

Obrázek 1: Úvodní stránka manažeru KonText

4.1.3. Vyhledávání

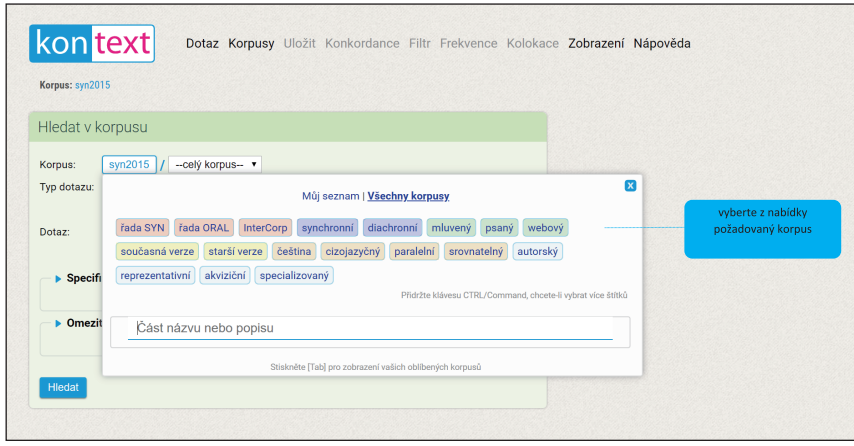
Před vyhledáváním je třeba si ujasnit základní parametry, které ovlivňují jeho výsledek. Jedná se o volbu korpusu, výběr typu dotazu, zadání dotazu, příp. specifikaci kontextu a omezení hledání. První tři jmenované kategorie jsou klíčové a nelze je vynechat. Na volbě těchto tří kategorií také přímo závisí výsledek korpusové analýzy. Poslední dvě kategorie jsou fakultativní a umožňují přesnější zadání dotazu.



Obrázek 2: Zadání dotazu v KonTextu

4.1.4. Volba korpusu

V první řadě je třeba zvolit korpus, v němž bude hledání probíhat. Výběrem pole s přednastaveným korpusem se otevře tabulka se štítky, podle nichž lze v korpusech hledat. Pro lepší přehlednost je možné se orientovat podle seznamu korpusů na encyklopedii wiki.[34] Každý ze štítků pod sebou skrývá nabídku všech relevantních korpusů, z nichž si lze následně vybrat ten nejvhodnější.



Obrázek 3: Výběr zdrojového korpusu

4.1.5. Typ dotazu

Před zadáním vlastního dotazu je třeba vybrat jeho typ. Přednastavený je *základní* typ dotazu. Jedná se o intuitivní způsob zadání dotazu. Pokud je do pole dotazu zadáno slovo v základním tvaru (u podstatných jmen první pád jednotného čísla, u sloves infinitiv, u přídavných jmen nestupňovaný tvar mužského rodu apod.), zobrazí se výsledky obsahující všechny tvary hledaného slova (od slova *mladý* doklady obsahující všechny slovní tvary – *mladými*, *mladým*, *mladého* apod.). Pokud v *základním* typu dotazu zadáme konkrétní slovní tvar (např. *mladými*), budou výsledky analýzy zahrnovat pouze texty obsahující tento slovní tvar. V *základním* dotazu lze zadávat i slovní spojení, přičemž platí stejná pravidla jako u jednoslovných zadání. U sousloví v konkrétním tvaru (např. *s velkou pravděpodobností*) budou výsledky obsahovat texty s tímto konkrétním tvarem. Pro vyhledání všech tvarů sousloví je třeba zadat jeho komponenty v základním tvaru, tj. např. pro spojení *s velkou pravděpodobností* zadáme dotaz ve tvaru *s velký pravděpodobnost*.

34 https://wiki.korpus.cz/doku.php/cnk:uvod

Výsledky pak obsahují sousloví *s velkou pravděpodobností, s větší pravděpodobností, s největší pravděpodobností*. Využití základního typu dotazu je tedy vhodné pro jednodušší dotazy a doporučuje se pro korpusové začátečníky. Jeho nevýhodou je, že nemůže obsahovat tzv. regulární znaky, tj. znaky mající zástupnou funkci (více o regulárních znacích obsahuje wiki[35]).

Druhým typem dotazu je *lemma*. Do pole dotazu se zadává základní tvar slova a ve výsledcích se zobrazují všechny tvary zvoleného slova. Výhodou dotazu typu *lemma* je, že umožňuje doplnit dotaz o specifikaci slovního druhu, a tím u některých slov zamezit vysoké chybovosti z důvodu homonymie. Příkladem může být slovo *stát*, které v češtině existuje jako podstatné jméno i sloveso.

Třetím typem dotazu je *fráze*. Při volbě *fráze* se do pole dotazu zadává slovní spojení v konkrétním tvaru, které je pak také v témže tvaru zobrazeno ve výsledcích analýzy.

Čtvrtým typem dotazu je *slovní tvar*. Do pole dotazu se zadává konkrétní slovní tvar, který je pak zobrazen ve výsledcích. Pokud je při tomto typu zadán neutrální slovní tvar (např. infinitiv), je na rozdíl od *základního* typu dotazu ve výsledcích zobrazen jen infinitiv, nikoli další slovní tvary. Slovní tvar je možné upřesnit volbou slovního druhu (např. *mezi* – předložka vs. podstatné jméno v 6. pádě jednotného čísla).

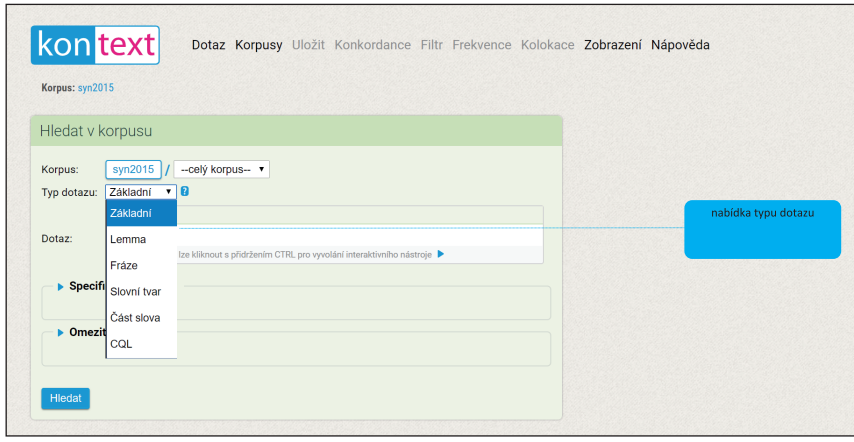
U typů dotazů *lemma* a *slovní tvar* je na rozdíl od *základního* typu dotazu možné volit slovní druhy, a tím eliminovat tvarovou homonymii.

Předposledním nabízeným typem dotazu je *část slova*. Umožňuje vyhledat slovní tvary, které obsahují zadaný řetězec znaků, a zahrnout do dotazu též regulární (tj. zástupné) znaky. Řetězcem znaků je myšlen pouhý sled znaků/písmen bez nároku na morfologické dělení (kořen slova, předpony, přípony atd.). Hledáme-li tedy pomocí tohoto typu dotazu slova s předponou *před-*, bude ve výsledcích analýzy zobrazen též slovesný tvar *přede* od slovesa *přist*.

Posledním typem dotazu je *CQL* (= *corpus query language*), což je dotazovací jazyk umožňující komplikovanější vyhledávání. Více k práci s CQL najdete na wiki.[36]

Pro jednoduchou a rychlou orientaci v rozdílech mezi typy dotazů slouží nápověda ve formě modrého pole s otazníkem, která uživateli poskytne rychlou charakteristiku zvoleného typu dotazu.

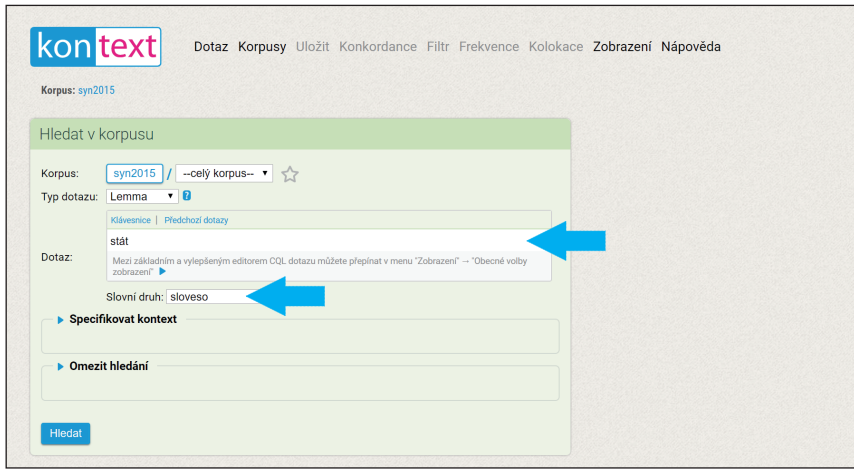
35 https://wiki.korpus.cz/doku.php/kurz:regularni_vyrazy
36 https://wiki.korpus.cz/doku.php/kurz:pokrocile_dotazy



Obrázek 4: Volba typu dotazu

4.1.6. Zadání dotazu

Po výběru typu dotazu je třeba dotaz formulovat. V závislosti na vybraném typu dotazu lze zadat základní nebo konkrétní slovní tvar, příp. použít v dotazu regulární znaky. Po výběru některých typů dotazů je zpřístupněno doplňkové okno umožňující definovat slovní druh. U níže uvedeného příkladu je hledáno sloveso *stát* ve všech jeho tvarech (je tedy vyloučena homonymie s podstatným jménem *stát*).



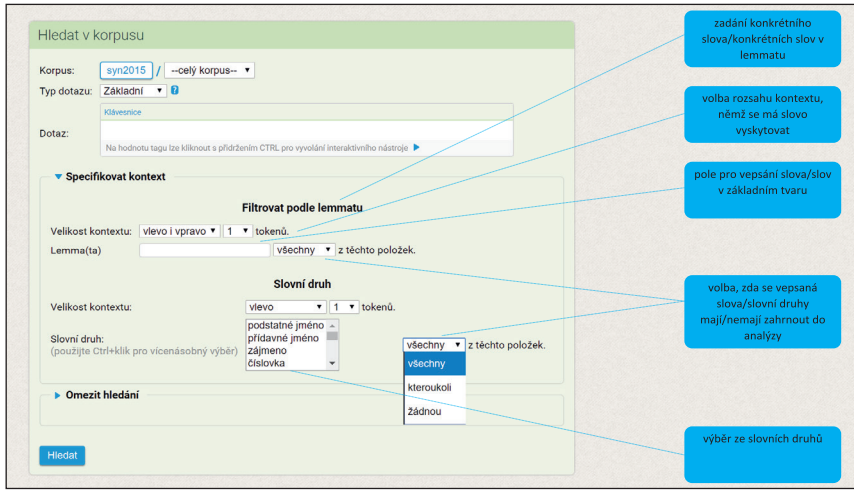
Obrázek 5: Zadání dotazu a specifikace slovního druhu

4.1.7. Specifikace kontextu

Oddíl *specifikace kontextu* se využívá tehdy, pokud uživatel potřebuje ovlivnit kontext vyhledávaného výrazu nad rámec možností, které mu poskytuje zadání dotazu. Využívá se tehdy, pokud hledané slovo chceme v jeho nejbližším okolí rozšířit o další slovo/slova, příp. pokud chceme další slovo/slova v okolí z vyhledávání vyloučit. Zadávat lze jak konkrétní slova/slovní tvary, tak kategorie slovních druhů. Specifikace kontextu také slouží tam, kde je třeba vyhledat víceslovné spojení, jehož komponenty nestojí pevně za sebou (k tomu slouží typ dotazu *fráze*).

Po volbě *specifikace kontextu* se zadávací tabulka rozšíří o několik částí. V oddílu *Filtrovat podle lemmatu* je třeba nejprve určit, jaký kontext má být do vyhledávání zahrnut. Kontext se určuje v tokenech (jeden token = zjednodušeně jeden slovní tvar na daném místě), jejichž počet si uživatel volí v tabulce. Před údajem k počtu tokenů uživatel vybírá, zda se počet zahrnutých tokenů týká jen části textu před hledaným slovem, tj. *vlevo*, nebo po hledaném slově, tj. *vpravo*, nebo na obě strany od hledaného slova, tj. *vlevo i vpravo*. Je možné zadávat slova v jejich základních tvarech, tj. lemmatech. Pole *Lemma(ta)* umožňuje vepsat jedno nebo více slov. V tabulce na konci řádku je třeba vybrat, zda mají být zadaná slova obsažena *všechna* v zadaném kontextu, nebo zda může být vybráno *kterékoli* ze zadaných slov, příp. zda nemá být vybráno *žádné*. Při zadání jednoho slova znamená výběr *všechny* a *kteroukoli* totéž. U zadání *všechna* se ve výsledcích hledání zobrazí všechny úryvky textů, které obsahují hledaný výraz (v poli *Dotaz*) a zároveň také všechna zadaná slova v rámci nastaveného kontextu. Jako příklad uveďme podstatné jméno *telefon* zadané v poli *Dotaz*, k němuž jsou ve specifikaci kontextu uvedeny atributy *chytrý* a *mobilní*. Výsledky pak budou zobrazovat jen ty úryvky textů, které obsahují všechna tři slova, tj. *chytrý*, *mobilní*, *telefon*. U zadání *kteroukoli* se ve výsledcích hledání zobrazí všechny úryvky textů, které obsahují hledaný výraz (v poli *Dotaz*) a zároveň vždy alespoň jedno ze zadaných slov, ve výše uvedeném příkladu to tedy budou úryvky obsahující slovní spojení *chytrý telefon* a *mobilní telefon*. U zadání *žádnou* se ve výsledcích hledání zobrazí všechny úryvky textů, které obsahují hledaný výraz (v poli *Dotaz*) a zároveň neobsahují ani jedno ze zadaných slov, obsahují tedy *telefon*, ale nikoli *chytrý* ani *mobilní*. Volba *žádnou* má tedy funkci negativního filtru, jímž definujeme, s jakými slovy se slovo zadané v dotazu nemá vyskytovat.

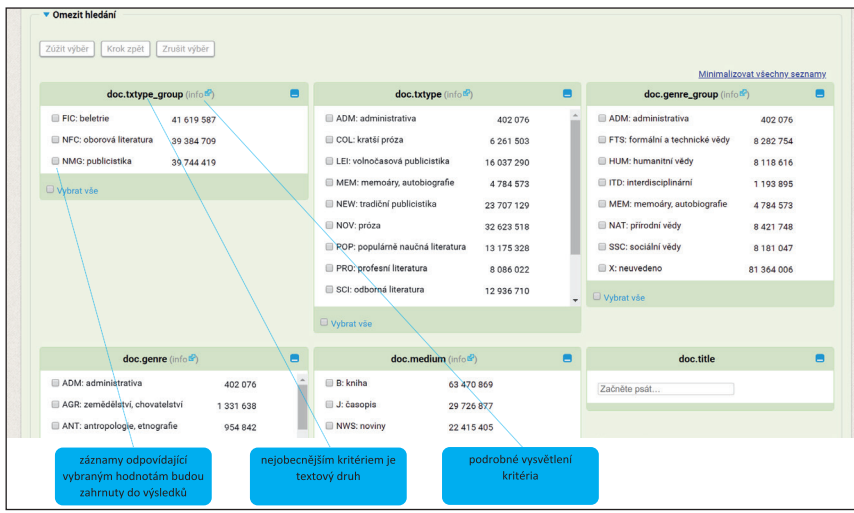
V oddílu *Slovní druh* se postupuje při zadávání stejně jako v oddílu *Filtrovat podle lemmatu*, jen s tím rozdílem, že se nevypisují konkrétní slova v základním tvaru, ale volí se jen slovní druh/druhy jako kategorie.



Obrázek 6: Specifikace kontextu

4.1.8. Omezení hledání

Pokud se uživatel rozhodne, že nechce vyhledávat v celém zvoleném korpusu, může své vyhledávání omezit. Texty jsou v korpusu tříděny podle různých kritérií (podle textového druhu, žánru, typu média, pohlaví autora apod.) a na základě těchto kritérií lze vybrat skupiny textů, které budou z analýzy vyloučeny.



Obrázek 7: Omezení hledání

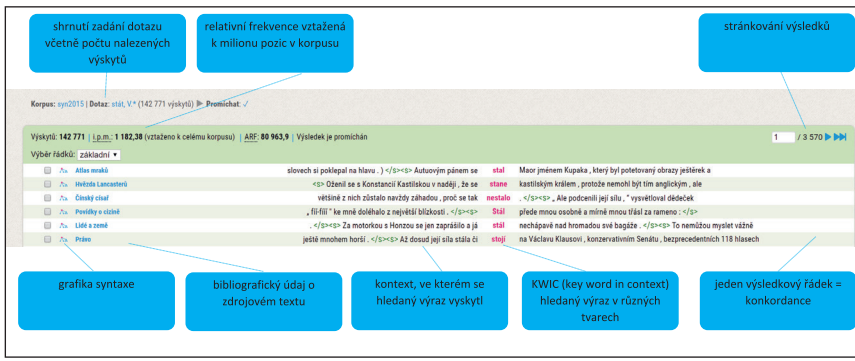
Po vyplnění úvodní zadávací tabulky (ne všechny oddíly je třeba vždy vyplnit) a zvolení tlačítka *Hledat* se spustí analýza.

4.1.9. Výsledky hledání

V následujícím textu bude popsána výsledková tabulka a možnosti dalšího zpracování zobrazených výsledků. Pro zde použité vzorové vyhledávání byl zvolen výchozí korpus *SYN2015*, typ dotazu *Lemma*, dotaz *stát* se specifikací slovního druhu *sloveso*.

Ve výsledkové tabulce je zobrazeno shrnutí zadání dotazu a počet nalezených výsledků, tj. počet textů, v nichž se objevil slovní tvar odpovídající zadání. Kromě absolutního počtu výskytů slovního tvaru je uváděn také údaj i.p. m., což je relativní frekvence přepočítaná na celkovou velikost korpusu. Pro kvantifikaci se jedná o relevantnější údaj, než je frekvence absolutní.

Výsledková tabulka obsahuje několik základních částí. Každý řádek (konkordance) je úryvkem textu, v němž se hledané slovo vyskytuje. V každém řádku lze odlišit několik typů údajů, které napříč řádky tvoří sloupce. Zleva se jedná o ikonu umožňující nahlédnout do syntaktické struktury zvolené věty. Následuje bibliografický údaj o zdrojovém textu, po jehož výběru se zobrazí kompletní bibliografické informace o textu. Nejdůležitější je střední část obrazovky, která nabízí úryvky textů s vycentrovaným klíčovým/hledaným slovem (KWIC). Pokud je potřeba zobrazení delšího kontextu, je to umožněno kliknutím na vybrané klíčové slovo. Defaultní nastavení korpusu zobrazuje 40 konkordancí. Pro zobrazení dalších je třeba se přesunout na další stránku výsledků.



Obrázek 8: Výsledky hledání

Zobrazením výsledkové tabulky ale práce s korpusem nekončí/nemusí končit. Výsledky lze ještě třídit a filtrovat nebo v nich dále hledat. Všechny nástroje pro zpracování výsledků a ovládání vyhledávání jsou k nalezení v horním menu. V následujícím textu si představíme ty nejdůležitější z nich.

Pomocí záložky *Dotaz* je možné vrátit se do úvodní zadávací tabulky a zahájit tak nové hledání. Nabídka *Předchozí dotazy* zobrazí historii dosavadních dotazů (v rámci jednoho sezení) a umožní se vrátit ke starším vyhledáváním a jejich výsledkům.

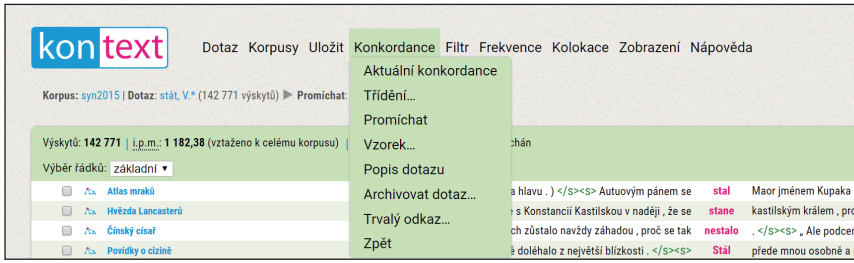
V záložce *Korpusy – Dostupné korpusy* je nabízen přehled všech korpusů ČNK, které je možné zvolit jako zdroje pro vyhledávání (obdobně v úvodní zadávací tabulce při volbě korpusu). ČNK umožňuje svým uživatelům vytvoření vlastních subkorpusů, tj. korpusů, které si uživatel definuje z dostupného textového materiálu ČNK sám. Pro vytvoření vlastního subkorpusu musí být relevantní důvod, protože výběrem textů do korpusu jsou významně ovlivněny výsledky hledání a tím i interpretace těchto výsledků. Pro vytvoření a správu vlastních subkorpusů slouží nabídka *Mé subkorpusy*, *Vytvořit vlastní subkorpus* a *Vytvořit veřejný subkorpus*. Vlastní subkorpus bude k dispozici jen danému uživateli, veřejný subkorpus lze nasdílet vybraným uživatelům, např. studentům ve třídě.

Obrázek 9: Vytvoření vlastního subkorpusu

Výsledky hledání lze uložit na vlastní zařízení, např. na disk počítače. Je tím umožněna práce s daty i v režimu offline. Ukládat lze v různých formátech, např.csv.xlsx.xml či.txt.

Obrázek 10: Uložení výsledků hledání

S výsledky hledání, resp. jednotlivými konkordancemi, lze dále pracovat. Relevantní může být jejich promíchání (volba *Promíchat*) či výběr náhodného vzorku, jehož velikost určuje uživatel sám (volba *Vzorek*). Tuto volbu lze provést v případech, kdy je počet nalezených konkordancí příliš vysoký na to, aby mohlo dojít k jeho manuálnímu zpracování.

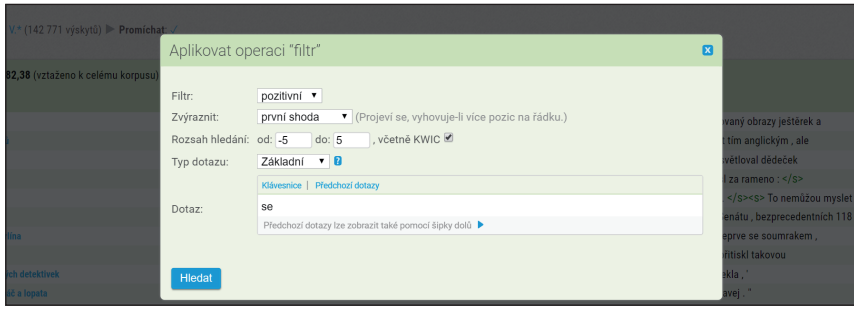


Obrázek 11: Promíchání a výběr vzorku

Výsledky lze různým způsobem také filtrovat. V nabídce je *filtr pozitivní* a *negativní*. V *pozitivním* filtru dochází k přidání další podmínky, např. dalšího slova, se kterým se má hledaný výraz vyskytovat. *Negativní* filtr naopak zaručuje odfiltrování např. nechtěných slovních spojení.

Obrázek 12: Filtrování výsledků

Ukázkou pozitivního filtru v hledání lemmatizovaného slovesa *stát* je přidání podmínky, která zajistí kombinaci slovesa *stát* se zvratným zájmenem *se*. Zvolený filtr je pozitivní, rozsah hledání -5/5 značí, že zvratné zájmeno se bude vyskytovat na pozicích pět slovních tvarů vlevo a pět vpravo od klíčového slovesa *stát*. Typ dotazu je zvolen jako základní, čímž je zajištěna flexe zájmena (*se*; *si*). Ve výsledné vyfiltrované tabulce se zvratná zájmena zobrazují barevně a kurzívou. Obdobným způsobem lze užít i filtr negativní. Příkladem může být hledané přídavné jméno *mobilní*, přičemž za použití negativního filtru jsou ignorovány všechny výskyty s podstatným jménem *telefon* na první pozici vpravo.



Obrázek 13: Využití filtru – výběr užití zvratného *se* u slovesa *stát*

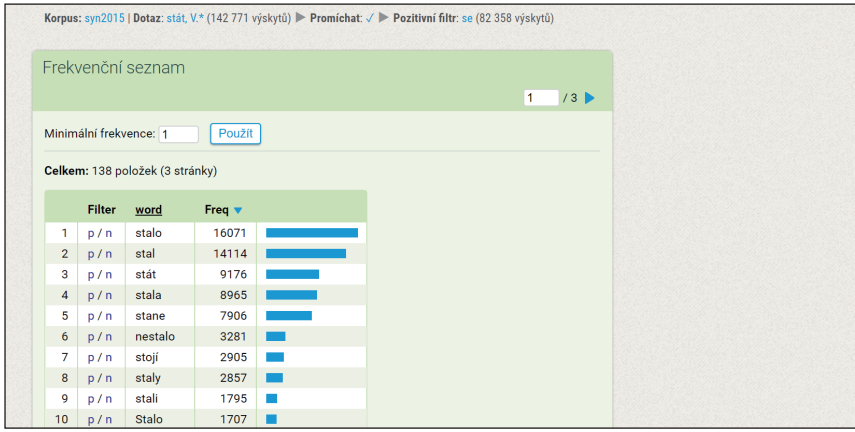
Obrázek 14: Využití filtru – výsledky

Další z nabídky operací je *Frekvence*. Nabízí statistické údaje o lemmatech, slovních tvarech, rozložení hledaného slova v dokumentech a typech textů a umožňuje provést tutéž statistickou analýzu i na základě vlastních kritérií. Je-li předmětem hledání jedno lemma nebo jeden slovní tvar, nemá smysl zadávat statistickou analýzu lemmatu či slovního tvaru, protože výsledkem je jen zadané lemma či slovní tvar sám. U lemmatu je ovšem možné analyzovat jednotlivé slovní tvary a dozvědět se tak, které slovní tvary z celého paradigmatu převládají.

Obrázek 15: Frekvence – výběr řazení

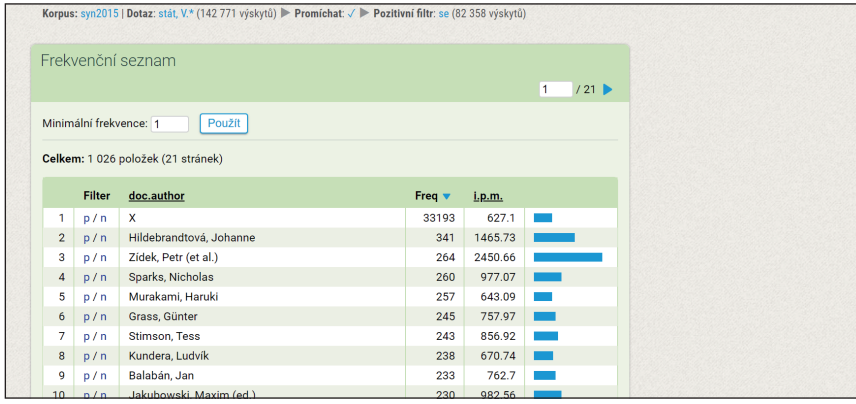
Ve zvoleném příkladu lemmatu *stát* vypadá analýza slovních tvarů takto: nejfrekventovanějším slovním tvarem lemmatu *stát* je *stalo*, následované tvary *stal* a *stát* atd. Frekvenční seznam lze řadit abecedně podle nalezených

tvarů (kliknutím na sloupec *word*) anebo u vybraného slovního tvaru použít pozitivní/negativní filtr.



Obrázek 16: Frekvenční seznam – absolutní frekvence

Ve frekvenční analýze dle vlastních kritérií mohou výsledky vypadat následovně (frekvenční seznam byl vytvořen na základě Kritéria *Vlastní – autor*). Seznam je seřazen dle absolutní frekvence. Pro řazení dle relativní frekvence na milion pozic je nutné kliknout na *i.p.m.*

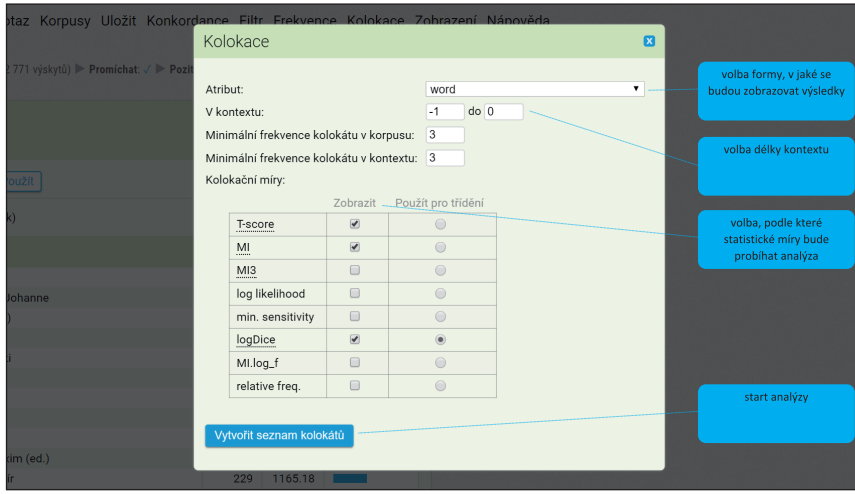


Obrázek 17: Frekvenční seznam

V korpusu lze také nechat analyzovat kontext vyhledaného slova a tím zjišťovat častá slovní spojení (kookurence či kolokace). Slouží k tomu volba *Kolokace*. V tabulce nastavovacích parametrů je třeba zvolit, v jaké formě budou zobrazeny výsledky (položka *Atribut*). Nejčastěji se volí *word* (slovní

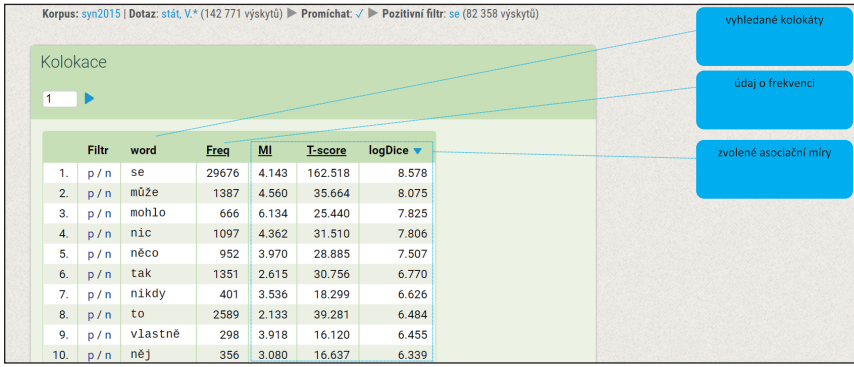
tvar) nebo *lemma*. Poté je třeba zvolit prohledávaný kontext. Pozice s minusem označují počet slov vlevo, bez minusu slova vpravo. Výběrem asociačních/statistických měr se určuje, podle kterého algoritmu bude analýza provedena. Každá z asociačních měr bere v úvahu různé parametry ovlivňující spojitost mezi vyhledaným slovem a jeho partnerem/kolokátem (více k různým asociačním mírám najdete na wiki).[37] Výsledná tabulka nabízí přehled analyzovaných slovních tvarů či lemmat (dle volby), se kterými se hledané slovo nejčastěji spojuje. Výsledky jsou tříděny podle asociační míry zvolené při zadávání, ale lze je nechat přeskupit podle ostatních asociačních měr (jejich zvolením) nebo podle frekvence.

Obrázek 18: Vlastní nastavení kolokace



Obrázek 19: Volba parametrů analýzy kolokací

37 https://wiki.korpus.cz/doku.php/pojmy:asociacni_miry



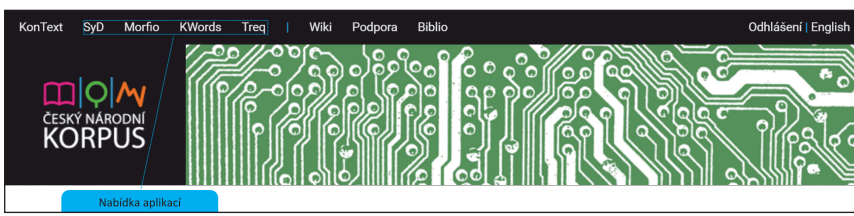
Obrázek 20: Kolokace – výsledky hledání

Poslední dvě volby *Zobrazení* a *Nápověda* slouží k nastavení uživatelského prostředí KonTextu a nabízejí odpovědi na nejčastěji kladené otázky. Podrobnější návody k práci s korpusy ČNK a s prohlížečem KonText jsou k nalezení také na wiki.[38]

Obrázek 21: Výběr zobrazení

4.2. Korpusové aplikace

Vedle vlastní přímé práce s korpusovým vyhledávačem je možné pro práci s korpusovými daty využít také aplikace, které jsou volně dostupné z hlavní webové stránky *korpus.cz*. Jedná se o aplikace SyD, Morfio, KWords a Treq. Všechny jsou dostupné bez registrace a přihlášení, a může je tak využívat kdokoli s přístupem na internet.



Obrázek 22: Náhled menu aplikací

38 http://wiki.korpus.cz/doku.php

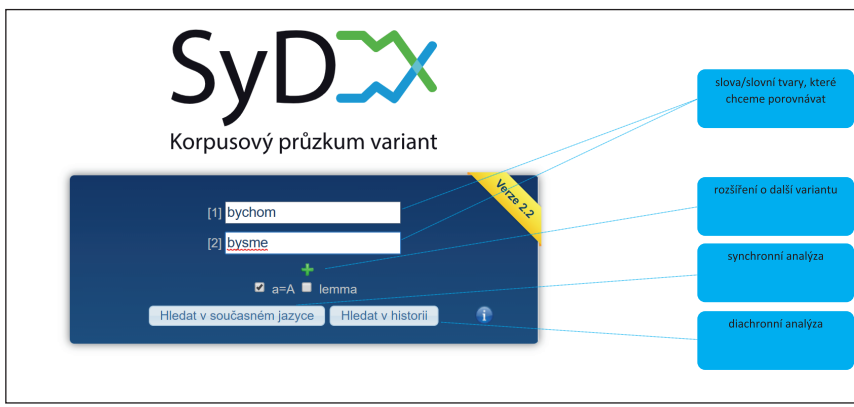
4.2.1. Aplikace SyD

Tato aplikace umožňuje analýzu a porovnání alespoň dvou jevů, které si v jazyce mohou konkurovat. Může se jednat o varianty pravopisné (*gymnasium* vs. *gymnázium*), morfologické (*mohou* vs. *můžou*), lexikologické (*již* vs. *už*) či syntaktické (na *kurzu* vs. v *kurzu*) a slovosledné (*sebe sama* vs. *sama sebe*). Výsledky je možné nahlížet ze dvou hledisek, synchronního (tedy z hlediska současného jazyka) a diachronního (tedy z hlediska vývoje jazyka v čase). Protože diachronní hledisko není pro výuku žáků s odlišným mateřským jazykem relevantní, věnuje se následující text pouze práci s částí synchronní.

Synchronní náhled nabízí porovnání vybraných variant v textech současného českého jazyka ve formě psané veřejné (analyzován je korpus *SYN2010*), dále ve formě psané neveřejné (analyzován je korpus *KSK-dopisy*) a ve formě mluvené (analyzovány jsou korpusy *ORAL2006*, *ORAL2008* a *ORAL2013*). Výsledky porovnání v této synchronní části nabízejí relativizovaná data o frekvenci vybraných jevů a jejich distribuci v textech (z hlediska textových typů a žánrů) a u mluvených textů data o pohlaví, věku, vzdělání a regionální příslušnosti mluvčích. Zadávat je možné konkrétní slovní tvary nebo lemmata, pro složitější zadání lze využít dotazovací jazyk CQL.

Pro potřeby výuky se jeví využitelnější analýza synchronní. Žáci na ní ověřují rozdíly mezi variantami pravopisnými, morfologickými i lexikálními. Následující obrázky zobrazují a popisují postup práce s aplikací SyD.

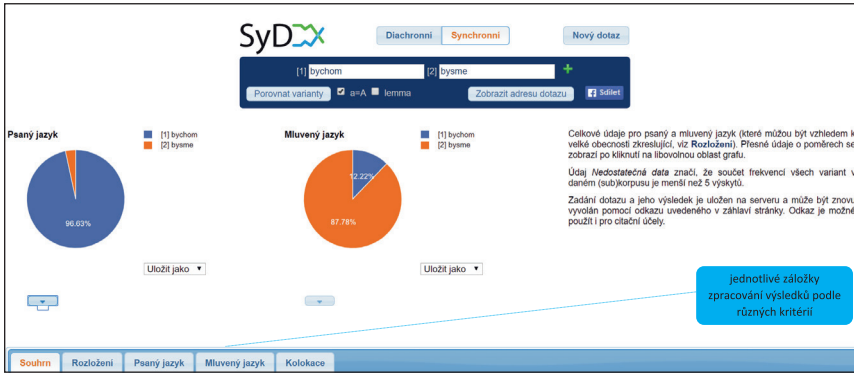
Nejprve je třeba najít alespoň dvě slova/slovní tvary, které budou porovnávány. Aplikace umožňuje porovnat i více slovních tvarů (znak +). Volbou *Hledat v současném jazyce* se spustí synchronní analýza.



Obrázek 23: SyD – zadání porovnání tvaru *bychom* a *bysme*

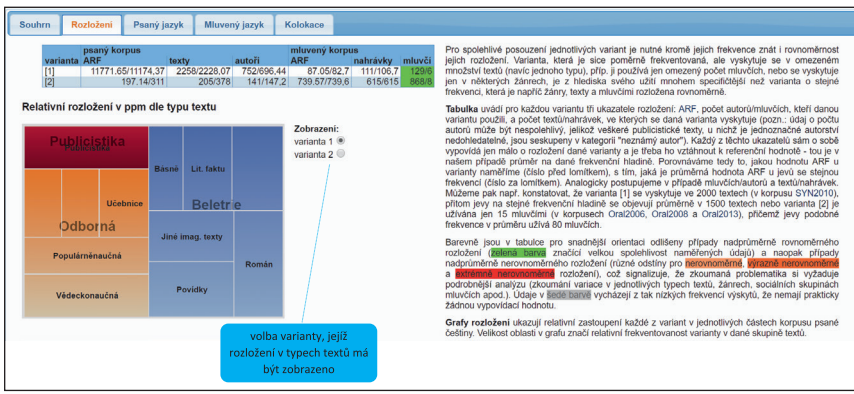
Výsledky se zobrazují v několika různých zpracováních. První výsledky jsou zobrazené v koláčovém grafu a napovídají zastoupení obou zvolených va-

riant v psaném a mluveném jazyce. Ostatní výsledky jsou roztříděné podle dalších kritérií a jsou k nahlédnutí v záložkách pod koláčovými grafy.



Obrázek 24: SyD – výsledek vyhledávání porovnání tvarů *bychom* a *bysme*

V záložce *Rozložení* je možné nahlédnout do grafu znázorňujícího, v jakých textových typech a žánrech se zvolená varianta nachází. Po zvolení varianty 2 se graf přeskupí. Výsledky jsou opatřeny podrobným komentářem umožňujícím orientaci v problematice interpretace těchto výsledků.

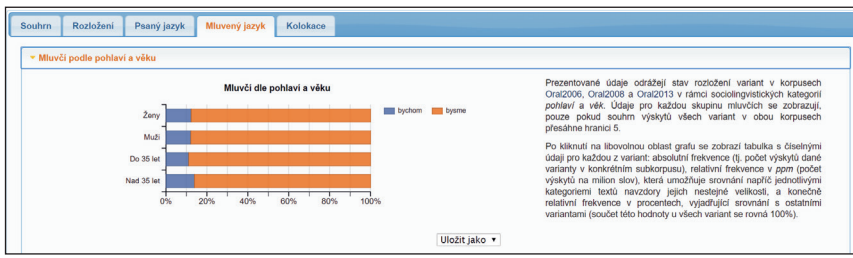


Obrázek 25: SyD – typy textů

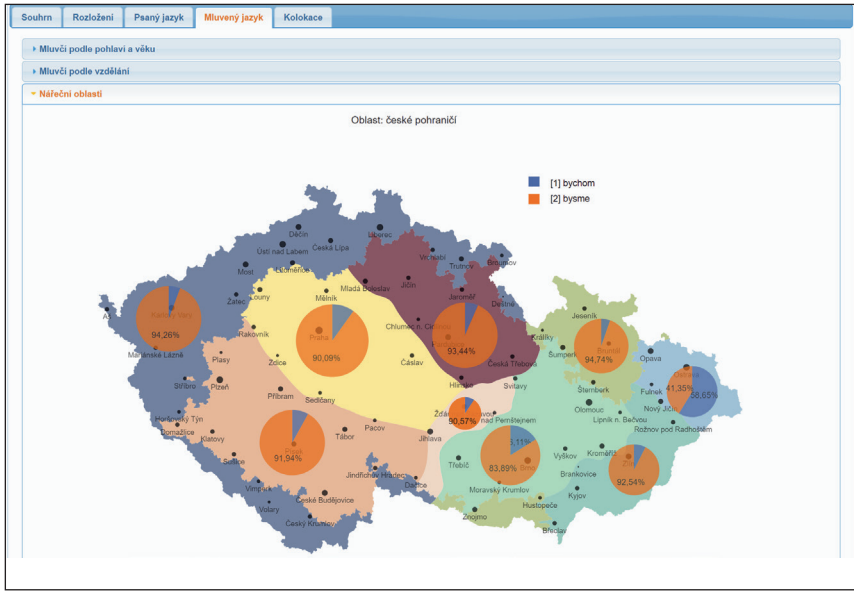
V následující záložce jsou graf týkající se psaného jazyka a rozložení zvolených variant v typech textů a textových žánrech zobrazeny podrobněji než v záložce *Rozložení*. Pro ilustraci je v následujícím obrázku znázorněno rozložení podle žánrů.

Obrázek 26: SyD – rozložení podle žánrů

Záložka pro mluvený jazyk nabízí jiné kategorie třídění, např. podle pohlaví a věku mluvčího, podle jeho vzdělání či nářeční příslušnosti.



Obrázek 27: SyD – třídění podle pohlaví a věku



Obrázek 28: SyD – třídění podle nářeční příslušnosti

Poslední záložka nabízí nejčastější kolokace k oběma variantám. I zde jsou výsledky opatřeny doprovodným vysvětlujícím textem. Více informací k práci s aplikací *SyD* je k nalezení ve wiki.[39]

Obrázek 29: SyD – kolokace

39 https://wiki.korpus.cz/doku.php/manualy:syd

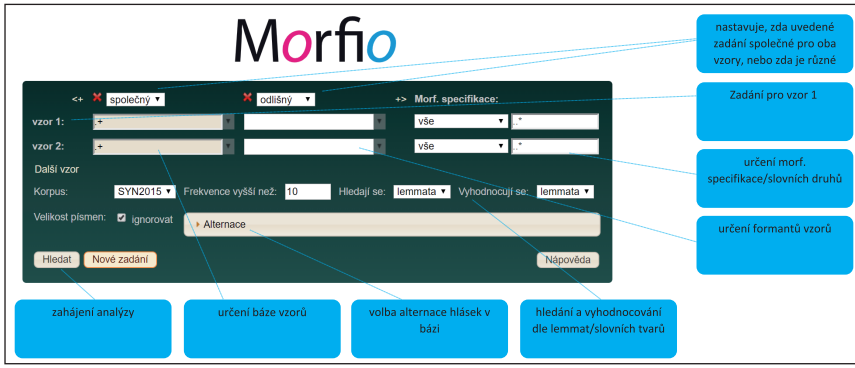
4.2.2. Aplikace Morfio

Aplikace *Morfio* je navržena tak, aby umožňovala odhalovat a pozorovat slovotvorné pravidelnosti češtiny. Využívá korpusových dat k odhadování rozsahu a produktivity slovotvorných modelů se zaměřením na odvozování. Vychází z poznatků slovotvorby češtiny a vyhledává dvojice (nebo vícečetné kombinace) slov, která se chovají stejně jako zadaný model.

Ve výuce žáků s OMJ lze *Morfio* využít pro odhalování slovotvorných pravidel češtiny, pro uvědomění si jazykového systému a osvojení si nových slovotvorných postupů. Znalosti ze slovotvorby jsou při učení se každému jazyku mocným nástrojem pro rychlé obohacení slovní zásoby o nová slova, a to na základě analogie, kterou slovotvorba disponuje. V případě slovotvorných produktů tak mluvíme o potenciální slovní zásobě, tj. slovní zásobě, kterou žák sice sám nikdy nepoužil ani se s ní nesetkal, ale na základě znalosti slovotvorného pravidla dokáže slovo vytvořit a použít. Práce s *Morfiem* tak představuje induktivní metodu výuky slovotvorby češtiny, která pracuje s jazykovým materiálem zábavnou formou.

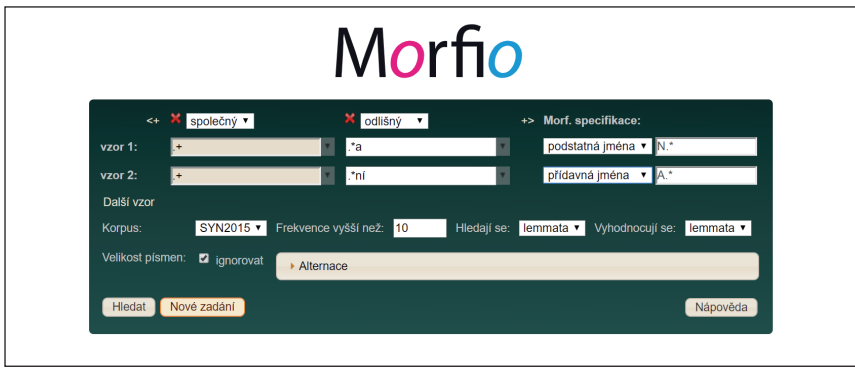
Východiskem práce s *Morfiem* je stanovení *báze* (základu slova) a *formantů* (částí slov, které podléhají změně). Při zadávání do vstupní tabulky lze využít jak kombinace písmen, tak regulárních znaků. Základem pro oba vzory může být tatáž konkrétní báze, např. *škol-*, nebo báze libovolná, tvořená jen regulárními znaky.+, přičemž uživatel volí, zda tato báze, příp. zvolené formanty mají být stejné nebo odlišné pro oba vzory. Oběma vzorům lze přiřknout ještě zvolený slovní druh, je-li tento údaj relevantní (např. slovo končící na *-í* může příslušet k různým slovním druhům, a proto lze slovní druh vybrat). Výběr slovního druhu se doporučuje především proto, aby se eliminovaly nežádoucí výsledky. U některých bází dochází při spojení s různými formanty ke změnám, s nimiž Morfio počítá a nazývá je *alternace*. Jedná se o změny typu *knih-a* X *kniž-ní*; *psá-t* X *psa-ní* apod. Důležité je, že ve vstupní tabulce není třeba vždy vyplňovat všechna zadávací pole, ale vždy jen taková, která jsou opodstatněná vzhledem k cíli analýzy.

Východiskem je stanovení báze (základu slova) a formantů (částí slov, které podléhají změně). Základem slova *škola* (tj. bází) je *škol-*, formantem je *-a*, přičemž i bez korpusu jsme sami schopni vytvořit další slova s jinými formanty od téže báze, např. školní, školský, školník, školení apod. Díky Morfiu můžeme rychle odhalit, která slova jsou schopna stejné změny jako slovo vyhledávané. Drobnou nevýhodou aplikace *Morfio* je, že nedokáže rozeznat hranice morfémů (např. u předpony *před-* jsou generována slova *předevčírem*, ale také *předobrý*).



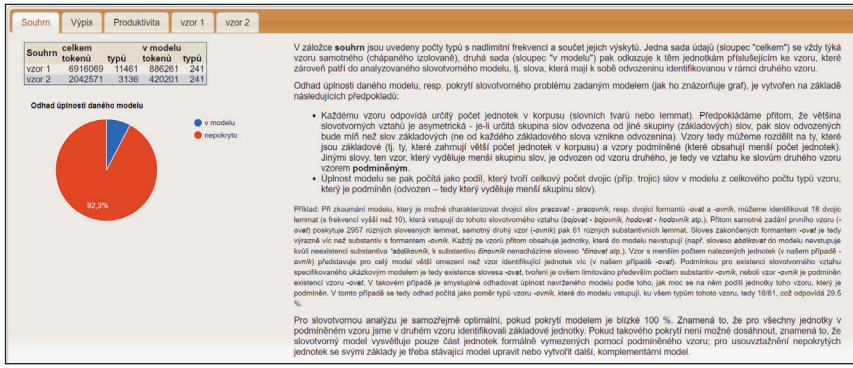
Obrázek 30: Morfio

Následující zadání umožňuje hledat dvojice slov, jejichž báze může být různá, ale která tvoří vždy dvojici, např. substantivum končící na *-a* a tvořící adjektivum končící na *-ní*. Hledají se a ve výsledcích se zobrazují lemmata. Alternace nebyly určeny.

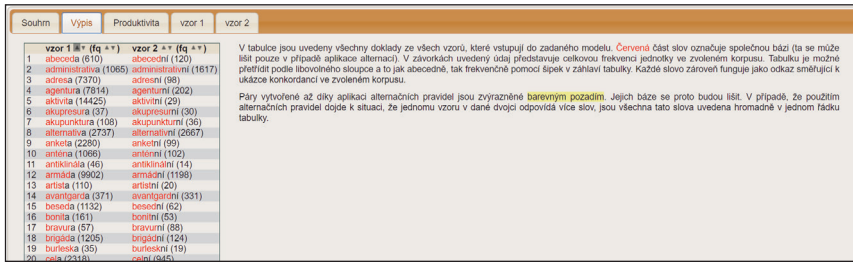


Obrázek 31: Morfio – zadání dotazu

Výsledky analýzy se zobrazí ve formě několika záložek a mohou mít různou míru využití pro různé účely. Každá z výsledkových tabulek je opatřena doprovodným komentářem. Úvodní tabulka a graf představují souhrnné výsledky analýzy. Pro účely výuky se jeví nejvyužitelnější tabulka *Výpis*, nabízející nalezené páry, které jsou kandidáty na hledaný slovotvorný produkt (např. abeceda – abecední). Jednotlivé slovní tvary lze kliknutím aktivovat a dostat se tak do korpusových textů, v nichž lze pozorovat konkrétní užití daného slovního tvaru.



Obrázek 32: Morfio – souhrnné výsledky



Obrázek 33: Morfio – záložka Výpis

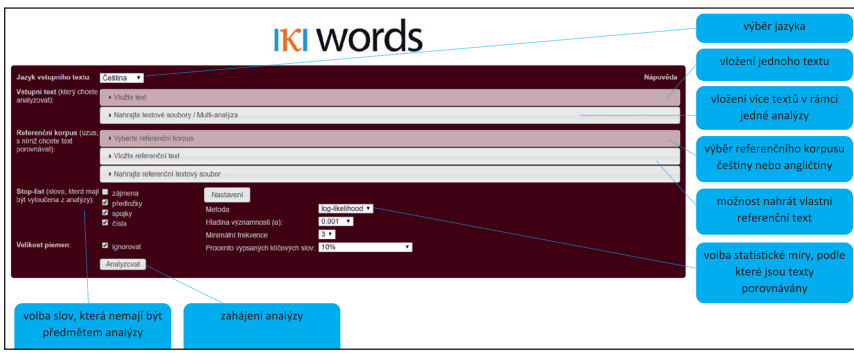
4.2.3. Aplikace KWords

KWords je aplikace umožňující analýzu až 20 vybraných textů, např. stejného textového druhu či žánru, při níž probíhá porovnání těchto vložených textů s referenčním korpusem. Při analýze jsou v textu vyhledávána klíčová slova (měří se relativní frekvence každého slovního tvaru a porovnává se s relativní frekvencí téhož slovního tvaru v referenčním korpusu, čímž lze získat seznam slovních tvarů, které jsou ve vybraných textech užity výrazně častěji a lze o nich uvažovat jako o klíčových, resp. majících spojitost s daným textových druhem, resp. žánrem).

Texty lze v současné době nahrávat v češtině nebo angličtině. Pro relevantní výsledky je však dobré zvážit volbu vhodného referenčního korpusu, s nímž budou vložené texty porovnávány. Nabídku referenčních korpusů lze nalézt na wiki.[40] Pro výuku žáků s OMJ však není angličtina relevantní a z nabízených českých korpusů doporučujeme využít přednastavený referenční korpus *SYN2015*.

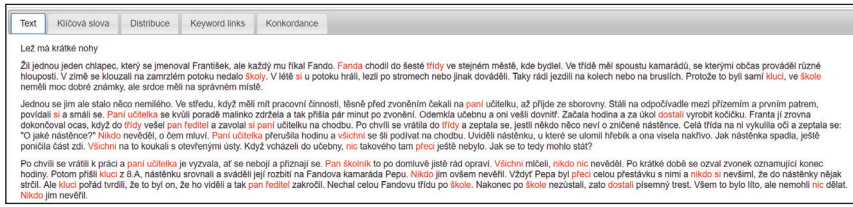
40 https://wiki.korpus.cz/doku.php/manualy:kwords

Práce s aplikací je intuitivní. Po otevření aplikace na webové stránce je třeba vyplnit zadávací tabulku. Klíčové je zadání vlastního textu. Analyzovaný text by měl mít od několika set do několika tisíc znaků a je možné ho vložit dvěma způsoby. Text (nebo i více textů) lze přímo kopírovat do připraveného okna, nebo lze v případě vkládání více textů užít tzv. multianalýzy, při níž se do korpusu jednotlivé texty nahrají. Druhým klíčovým krokem je zvolení referenčního korpusu, tedy korpusu, s nímž se vlastní nahrané texty budou porovnávat. Na výběr je několik referenčních korpusů. Pro analýzu současných textů doporučujeme využívat např. *SYN2015*. Před zahájením analýzy je vhodné zvážit, zda z ní nevyloučit některé/všechny nabízené slovní druhy, které jsou většinou vysokofrekvenční a mohou výsledky analýzy znehodnotit, např. předložky, zájmena, spojky apod. Vyloučení z analýzy lze provést v tzv. stop-listu. Vlastní analýza probíhá na základě statistických měr; je možné zvolit jednu ze statistických měr nebo obě dvě zároveň. Analýza začne probíhat po stisku příkazu *Analyzovat* v levé dolní části stránky.



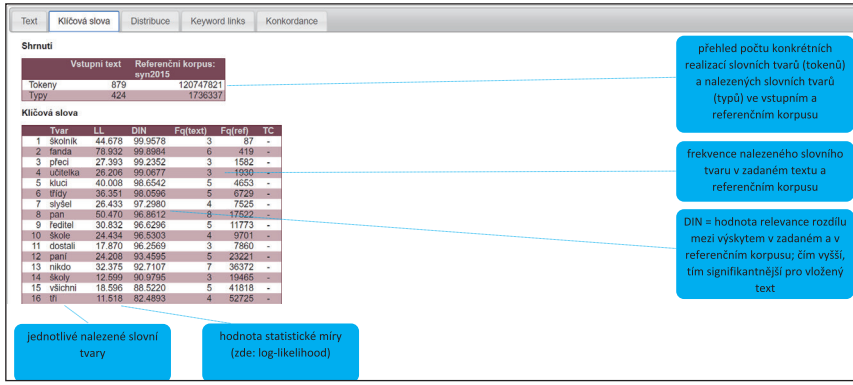
Obrázek 34: KWords – Zahájení analýzy

Výsledky analýzy se zobrazují několika způsoby. V záložce *Text* je zobrazen analyzovaný text a všechna nalezená signifikantní slova jsou zbarvena červeně. Text lze tedy pročítat a soustředit se přitom na vyhledané výrazy.



Obrázek 35: KWords – ukázka analýzy[41]

Výsledky analýzy jsou nabízeny také v jiných formách. V záložce *Klíčová slova* uživatel nalezne souhrnnou tabulku slovních tvarů vyhodnocených jako signifikantní pro zadaný text. Kromě seznamu je zde nabídnuta také hodnota statistické míry, podle které byl výpočet proveden, dále hodnota DIN, která uvádí míru relevance rozdílu mezi výskytem v zadaném textu a v referenčním korpusu (zde platí, že relevantní jsou slovní tvary s hodnotou 75–100, přičemž čím je hodnota vyšší, tím je relevantnější). U vygenerovaných slovních tvarů je také uvedena jejich frekvence v obou textových zdrojích.

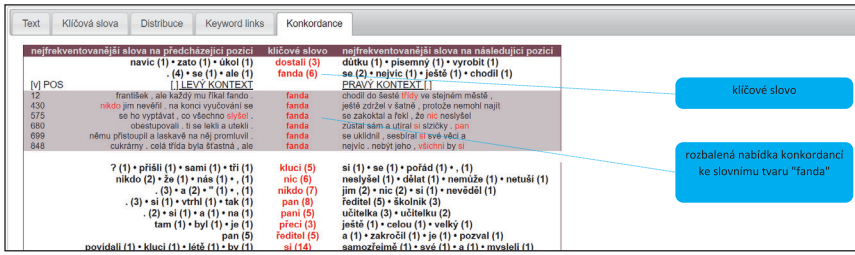


Obrázek 36: KWords – výsledek analýzy

Další možností náhledu je distribuce zobrazující rozložení nalezených signifikantních slov v textu a *Keyword links*, jenž zobrazuje vztahy mezi jednotlivými klíčovými slovy. Interpretace těchto dat je určena spíše pro odborníky a zkušené korpusové uživatele, proto se těmto doplňkovým informacím nebudeme dále věnovat. Bližší popis pro zájemce lze nalézt na wiki.

41 Analyzovaný text pochází z: http://www.cesky-jazyk.cz/slohovky/pohadky/lez-ma-kratke-nohy.html#axzz5IU96O5h7.

Poslední z řady je přehled konkordancí, tj. konkrétních míst v textu v zobrazení KWIC. Po výběru konkordance se soupis rozšíří o všechny konkordance s vybraným slovním tvarem. Tento náhled je výhodný především proto, že shromáždí všechny výskyty vybraného slovního tvaru pod jednu konkordanci a lze na něm, na rozdíl od náhledu lineárního textu, analyzovat chování vybraného slovního tvaru ve všech jeho výskytech najednou.



Obrázek 37: KWords – přehled konkordancí

Využití aplikace *KWords* ve výuce jazyka může být různorodé. Nabízí se možnost analýzy textových druhů a žánrů, na jejímž podkladu žáci samostatně vytvářejí vybraný textový druh či žánr. Zajímavou možností je analýza vlastních vytvořených textů a tím reflexe vlastního idiolektu nebo porovnání textů stejného zadání od žáků s češtinou jako mateřštinou a žáků s OMJ.

4.2.4. Aplikace Treq

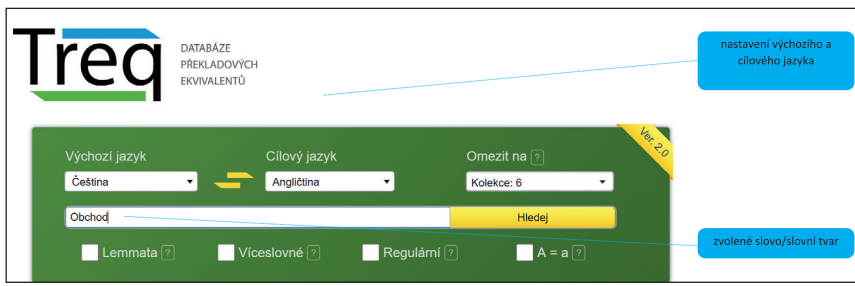
Aplikace *Treq* je databází překladových ekvivalentů. Protože je založena na textech paralelního korpusu *InterCorp*, funguje obousměrně pro jazykové páry čeština–cizí jazyk a angličtina–cizí jazyk. Cizí jazyky jsou zde nabízeny v rozsahu odpovídajícím jazykům zastoupeným v *InterCorpu*.

Samotné vyhledávání probíhá v prvním kroku zvolením výchozího a cílového jazyka. V druhém kroku musí být zvolen hledaný výraz. Vyhledávané slovo musí být slovo z jazyka označeného jako výchozí a lze ho formulovat jako určitý slovní tvar, jako lemma, případně lze vyhledávat víceslovná spojení nebo užít při hledání regulárních výrazů či ignorovat malá a velká písmena (až na vyhledávání konkrétního slovního tvaru je u všech výše jmenovaných potřeba danou kategorii zaškrtnout). Hledání lze provést ve všech textech *InterCorpu* nebo lze omezit zdrojové texty na beletristické jádro (texty ručně zkontrolované a zarovnané) nebo některou z kolekcí (texty automaticky zarovnané). Výsledkem hledání je seznam překladových ekvivalentů. Proklikem jednotlivých ekvivalentů se uživatel dostane do korpusu *InterCorp*, který mu zobrazí zarovnané texty obsahující jak výraz jazyka výchozího, tak nabídnutý ekvivalent jazyka cílového.

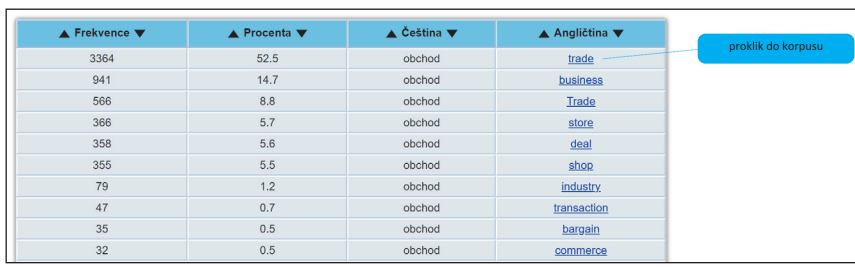
Na těchto kontextech lze ověřit, zda zvolený ekvivalent odpovídá nárokům hledaného ekvivalentu.

Své využití najde *Treq* především tam, kde se žáci učí pracovat s cizojazyčnými ekvivalenty a jejich výběrem a také se synonymy. *Treq* sice nelze označit za slovník, protože mu chybí celá řada nezbytných lexikografických informací, ale na rozdíl od běžného slovníku disponuje velkou výhodou, kterou je propojení s autentickými texty ve formě konkordancí obsahujících vyhledaný ekvivalent. Z kontextů tak lze mnohdy vyčíst stejné nebo ještě bohatší informace o významové a pragmatické stránce jazykového ekvivalentu než ze slovníku. Uživatel si užíváním *Trequ* cvičí nejen své jazykové znalosti, ale i svůj jazykový cit a senzibilizuje se na práci s textovým materiálem a jeho výpovědní hodnotou.

Následující obrázky nabízejí náhled úvodní zadávací stránky aplikace *Treq*, stránky s výsledky (nabídkou překladových ekvivalentů) a stránky zobrazující konkordance *InterCorpu* u zvoleného ekvivalentu. Více informací o aplikaci *Treq* je k nalezení na stránkách wiki.[42]



Obrázek 38: Treq – zadávací tabulka



Frekvence	Procenta	Čeština	Angličtina
3364	52.5	obchod	trade
941	14.7	obchod	business
566	8.8	obchod	Trade
366	5.7	obchod	store
358	5.6	obchod	deal
355	5.5	obchod	shop
79	1.2	obchod	industry
47	0.7	obchod	transaction
35	0.5	obchod	bargain
32	0.5	obchod	commerce

Obrázek 39: Treq – výsledek frekvenční analýzy

42 https://wiki.korpus.cz/doku.php/manualy:treq

Obrázek 40: Treq – náhled korpusu InterCorp – české slovo *obchod* a anglický ekvivalent *trade*

Využití korpusů ve výuce

5. Náměty pro strategie vyučování a pro tvorbu pracovních listů

5.1. Frekvenční analýza

Korpusová data je možné mj. využít také pro různé strategie vyučování, např. pro výuku slovní zásoby. Pro tento účel je dobře využitelná frekvenční analýza. Ve slovní zásobě jazyka existuje souvislost mezi frekvencí výskytu slova a jeho pořadím. Jde o tzv. Zipfův zákon. Mluvnice současné češtiny (2010) uvádí, že když seřadíme různá slova v určitém korpusu od nejfrekventovanějšího po nejméně časté a přiřadíme ke každému slovu číslo označující jeho pořadí, můžeme si všimnout, že pořadí slova krát jeho frekvence je víceméně konstantní. Z tohoto vztahu vyplývá, že v jazyce existuje velmi málo hodně frekventovaných slov – to jsou zejména gramatická slova jako předložky a spojky. Většinu lexikonu ale tvoří velké množství slov, která mají velmi nízkou frekvenci. Z uvedeného zákona plynou praktické důsledky pro výuku slovní zásoby ve výuce cizího jazyka. Nejfrekventovanějších 100 slov totiž pokrývá necelých 40 % textu, 1000 slov tvoří 62 % textu a se znalostí 3000 slov jsme schopni rozumět více než 75 % textu (Cvrček a kol 2010: 78). Jako ukázku využití frekvenční distribuce slovní zásoby můžeme použít následující tabulku. Data v tabulce pocházejí z korpusu *SYN2015*. Při práci s korpusovým manažerem KonText jsme vybrali typ dotazu lemma, tedy základní (slovníkový) tvar, který není ovlivněn různou frekvencí výskytů např. různých pádových tvarů. Do vyhledávacího řádku jsme zadali symbol.*, tedy symbol pro vyhledání všech slov o délce alespoň jednoho písmena. Při vyhledávání jsme omezili vyhledávání pouze na beletrii a publicistiku, oborovou literaturu jsme nezahrnuli.

Výsledkem je tabulka zahrnující sto nejvíce frekventovaných slov v daném korpusu. Podle vztahů uvedených výše by mělo platit, že těchto sto slov pokrývá 40 % textů v korpusu SYN2015.

Pořadí	Vše	Substantiva	Adjektiva	Slovesa	Předložky	Spojky
1	být	rok	velký	být	v	a
2	se	člověk	celý	mít	na	že
3	a	den	další	moci	s	ale

Pořadí	Vše	Substantiva	Adjektiva	Slovesa	Předložky	Spojky
4	v	ruka	dobrý	chtít	z	i
5	ten	dítě	nový	říci	do	jako
6	na	život	jiný	muset	o	když
7	on	místo	malý	vědět	k	aby
8	že	hlava	český	jít	za	nebo
9	s	doba	poslední	stát	po	než
10	z	oko	starý	dát	pro	ani
11	mít	muž	rád	říkat	od	protože
12	do	žena	vysoký	vidět	u	však
13	který	město	mladý	začít	před	pokud
14	já	svět	vlastní	myslet	při	kdyby
15	ale	čas	jediný	přijít	podle	či
16	o	dům	dlouhý	dostat	mezi	jestli
17	i	země	možný	dělat	bez	takže
18	k	strana	hlavní	udělat	pod	až
19	jako	chvíle	stejný	nechat	nad	tedy
20	co	práce	ostatní	najít	přes	ovšem
21	za	cesta	bílý	dokázat	kolem	jenže
22	tak	věc	špatný	čekat	proti	zatímco
23	moci	hodina	plný	hrát	kvůli	zda
24	svůj	voda	černý	mluvit	během	li
25	když	koruna	jistý	vypadat	díky	vždyť
26	po	dveře	evropský	vrátit	kromě	ať
27	oni	slovo	důležitý	vzít	vedle	dokud
28	už	Praha	různý	snažit	místo	jakmile
29	rok	případ	známý	zůstat	mimo	přestože
30	jak	konec	těžký	cítit	včetně	ačkoli
31	pro	cena	krásný	zeptat	uprostřed	aniž
32	všechen	milión	domácí	potřebovat	vůči	neboť
33	aby	pohled	jasný	znát	vzhledem	anebo
34	od	peníze	podobný	slyšet	pomocí	proto
35	jen	tvář	silný	žít	okolo	nicméně
36	jeho	škola	americký	podívat	navzdory	buď
37	chtít	problém	bývalý	patřit	podél	jak
38	říci	část	státní	sedět	skrz	avšak
39	jeden	týden	pražský	rozhodnout	oproti	byť
40	člověk	společnost	současný	vést	prostřednictvím	jelikož
41	muset	firma	příští	chodit	uvnitř	zato

Pořadí	Vše	Substantiva	Adjektiva	Slovesa	Předložky	Spojky
42	nebo	otec	minulý	dojít	naproti	jestliže
43	my	auto	místní	získat	nedaleko	jenomže
44	stát	hlas	mrtvý	otevřít	dle	coby
45	vědět	tělo	obrovský	pomoci	poblíž	nýbrž
46	ještě	noha	světový	věřit	ohledně	tak
47	jít	měsíc	lidský	pracovat	von	tudíž
48	u	noc	příjemný	ukázat	blízko	jednak
49	pak	matka	zajímavý	stačit	van	ač
50	tento	minuta	červený	dívat	zpod	neboli
51	než	řada	pravý	objevit		
52	velký	rodina	společný	jet		
53	až	jméno	schopný	zdát		
54	ani	zápas	základní	skončit		
55	před	ulice	politický	znamenat		
56	vy	pravda	veřejný	odpovědět		
57	jenž	stůl	letošní	držet		
58	ty	pán	šťastný	uvést		
59	dva	stát	nízký	dávat		
60	můj	procento	zvláštní	pokračovat		
61	celý	tým	krátký	umět		
62	tam	kniha	správný	vyjít		
63	hodně	okno	národní	ležet		
64	její	pokoj	zelený	lze		
65	další	kraj	německý	hledat		
66	něco	pan	finanční	brát		
67	dát	otázka	otevřený	změnit		
68	dobrý	situace	blízký	psát		
69	den	film	sociální	dodat		
70	ne	Kč	vhodný	podařit		
71	nový	pocit	samotný	začínat		
72	říkat	hra	modrý	odejít		
73	nic	projekt	jednoduchý	zastavit		
74	také	smrt	skutečný	zjistit		
75	vidět	paní	městský	připravit		
76	kde	síla	bezpečný	představit		
77	první	způsob	zlatý	zvednout		
78	teď	vláda	životní	platit		
79	ruka	většina	rychlý	sledovat		

Pořadí	Vše	Substantiva	Adjektiva	Slovesa	Předložky	Spojky
80	místo	vlas	dětský	trvat		
81	při	skupina	osobní	bát		
82	začít	důvod	volný	tvrdit		
83	protože	světlo	slavný	postavit		
84	podle	možnost	skvělý	existovat		
85	jiný	hráč	pracovní	vydat		
86	každý	kolo	prázdný	líbit		
87	druhý	byt	široký	připadat		
88	myslet	výsledek	určitý	napsat		
89	tenhle	klub	střední	uvidět		
90	náš	rameno	nutný	zapomenout		
91	takový	Petr	moderní	ptát		
92	dítě	syn	oblíbený	poznat		
93	mezi	metr	dnešní	napadnout		
94	nějaký	srdce	ruský	koupit		
95	jejich	systém	čistý	působit		
96	přijít	přítel	rodinný	otočit		
97	život	začátek	úspěšný	milovat		
98	však	krok	hezký	číst		
99	dostat	akce	jednotlivý	chápat		
100	hlava	policie	krajský	pustit		

Tabulka 1: Nejfrekventovanější slova v korpusu SYN2015

5.2. Pracovní listy

Následující ukázkové pracovní listy obsahují čtyři části. Na první straně naleznete anotaci a popsaný postup práce s pracovním listem. Druhou stranu tvoří samotný pracovní list, který lze namnožit a rozdat žákům. Po pracovním listu následuje popis práce s korpusem s názornými snímky obrazovky. Na konec je zařazeno řešení úkolů. Pracovní listy obsahují jak cvičení, která vyžadují přímou práci s korpusem během výuky (hands-on), tak cvičení, která jsou založená na jazykových datech z korpusu, ale přímou práci s korpusem během výuky nevyžadují (hands-off). Pracovní listy mají sloužit jako inspirace pro tvorbu vlastních korpusových cvičení.[43]

43 Podobné pracovní listy naleznete také pod odkazem https://www.korpus.cz/proskoly.

5.2.1. Prostředí KonText

Anotace
Cvičení tohoto typu je vhodné zařadit do výuky, pokud dělá žákům problém rozlišovat význam podobných adjektiv a užívat je ve správných spojeních. Může jít jak o adjektiva, která mají jiný význam – zde např. *zdravý* a *zdravotní*, tak o adjektiva, která mají v některých kolokacích význam stejný, ale existují různá omezení – např. *šedý* a *šedivý*, kdy se některá substantiva mohou pojit jen s jedním adjektivem (*šedá eminence*). Cvičení také rozšiřují slovní zásobu. Pracovní list vyžaduje samostatnou práci s korpusovým rozhraním KonText a obsahuje tři úkoly, které vedou k tomu, aby si žák vhodné použití adjektiv osvojil sám pomocí vyhledávání typických kolokací, následně formulací pravidel jejich spojování se substantivy a nakonec jejich použitím v textu.

Postup práce
Před rozdáním pracovního listu je nutné žáky připravit na téma úkolů a vhodně je motivovat k jejich vypracování. V tomto případě může učitel s žáky mluvit o tom, co sami dělají pro své zdraví a jaké existují zdravotní problémy. K diskuzi může využít předem připravený text, který bude obsahovat tematizovaná adjektiva, nebo jen otázku v názvu pracovního listu.

Po tomto úvodu učitel obrátí pozornost žáků k adjektivům *zdravý* a *zdravotní* a zeptá se jich, jaký je v nich rozdíl a která substantiva se s nimi mohou pojit. Jako motivační prvek dobře slouží návodné obrázky (zdravé jídlo nebo tělesné aktivity X různá onemocnění, nemocniční personál). Odpovědi žáků lze zapisovat na tabuli formou myšlenkové mapy. Následně učitel rozdá žákům pracovní list. Ti pak splní první úkol dle schopnosti práce s korpusem sami, ve dvojicích nebo společně s učitelem. Před plněním je vhodné projít postup práce s korpusem. Žáci mohou první cvičení také dostat za domácí úkol. V případě, že žáci některým substantivům, která najdou, nerozumí, je třeba jejich význam osvětlit. Vhodné je využít předem připravené obrázky a příkladové věty. Příkladové věty mohou žáci sami najít kliknutím na pozitivní filtr (p/n) u vybraného substantiva v seznamu kolokátů, po kterém se jim objeví všechny věty s daným adjektivem a substantivem. Věty pocházejí z českých textů, je třeba proto počítat s tím, že mohou být pro žáky na nižší jazykové úrovni (asi do B2) nesrozumitelné a matoucí. Důležité je upozornit žáky také na gramatický rozdíl mezi měkkým adjektivem *zdravotní* a tvrdým adjektivem *zdravý*. Tvary adjektiv si opět mohou ověřit kliknutím na pozitivní filtr u daného substantiva. Nalezené kolokace by během kontroly vždy měli číst se správnou koncovkou adjektiva (tj. *zdravý rozum*, *zdravá výživa* X *zdravotní problém*, *zdravotní sestra*).

Až si žáci najdou substantiva a bude osvětlen jejich význam, měli by se sami nebo ve skupině pokusit zformulovat významovou definici. Mohou vymyslet příklady dalších kolokátů (některé jsou uvedeny v části řešení).

Druhý úkol žáci plní až po kontrole prvního. Opět je nutné, aby před začátkem práce rozuměli zadaným substantivům a aby adjektiva použili se správnou koncovkou.

Třetí úkol mohou žáci plnit samostatně nebo ve dvojicích. Mohou soutěžit o to, který text bude nejvtipnější nebo ve kterém bude použito nejvíce spojení adjektiv a substantiv. Učitel si může připravit začátek textu, na který budou žáci navazovat. Např. *Když byla Klára malá, měla hodně zdravotních problémů. Kvůli nim často chyběla ve škole, a tak ještě v deseti letech neuměla dobře číst a psát…* Texty poté mohou žáci přečíst v plénu nebo je učitel může vybrat a opravit. Texty si také žáci mohou opravit navzájem.

Pracovní list

ZDRAVÝ ŽIVOTNÍ STYL = ŽÁDNÉ ZDRAVOTNÍ PROBLÉMY?

1. Najděte v korpusu SYN2015 deset nejčastějších substantiv po adjektivech *zdravý* a *zdravotní*. Potom vysvětlete, co znamenají a kdy se používají.

	zdravý	zdravotní
1.		
2.		
3.		
4.		
5.		
6.		
7.		
8.		
9.		
10.		

2. Co ještě může být zdravé a co zdravotní? Přiřaďte a napište k substantivům adjektivum se správnou koncovkou:

handicap, holčička, noha, obtíž, prohlídka, zub

zdravý: .....

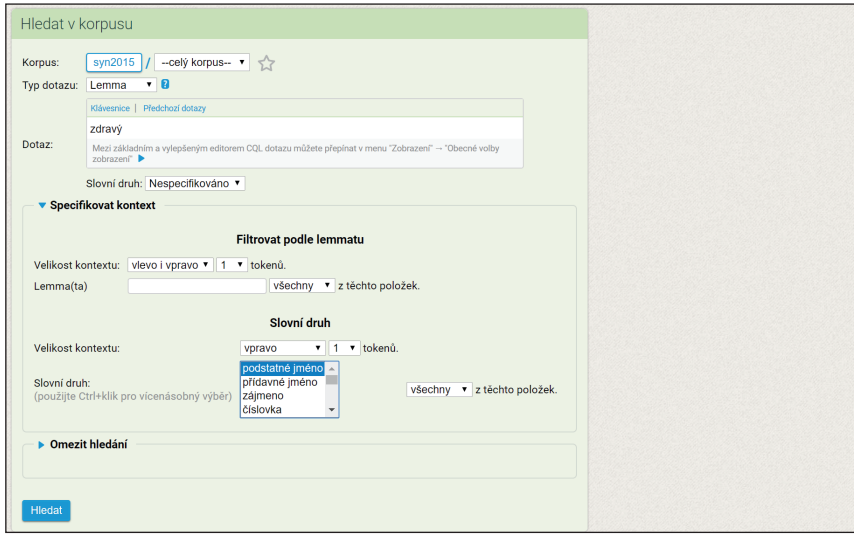
zdravotní: .....

3. Napište krátký text, ve kterém použijete vždy alespoň třikrát adjektivum *zdravý* a *zdravotní*. Ke každému adjektivu musíte použít jiné substantivum.

Postup práce s korpusem

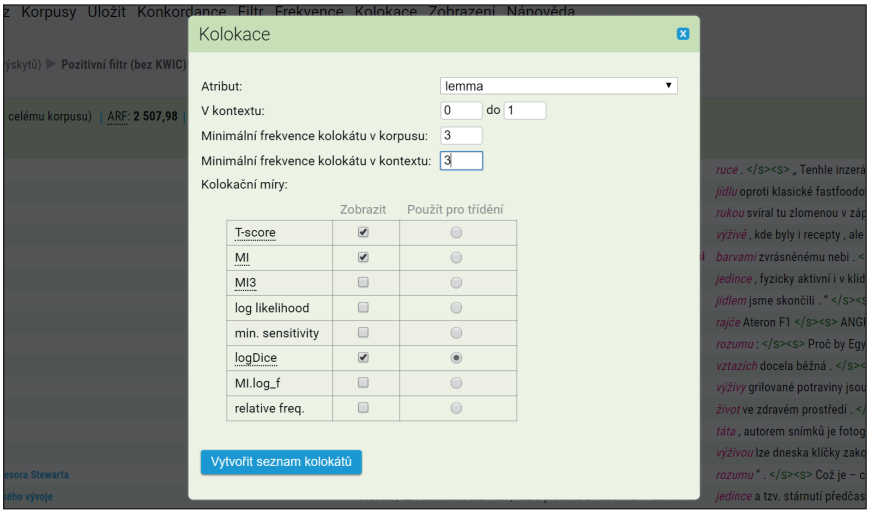
Postup 1

1. V korpusu *SYN2015* zvolte Typ dotazu lemma a do okénka *Dotaz* napište hledané adjektivum (*zdravý* nebo *zdravotní*). Dále rozbalte možnost *Specifikovat kontext*. Zde v části *Slovní druh* zadejte *Velikost kontextu* vpravo 1 token. Jako slovní druh zvolte podstatné jméno. Tím najdete všechny věty, ve kterých bude hned vpravo vedle daného adjektiva podstatné jméno.



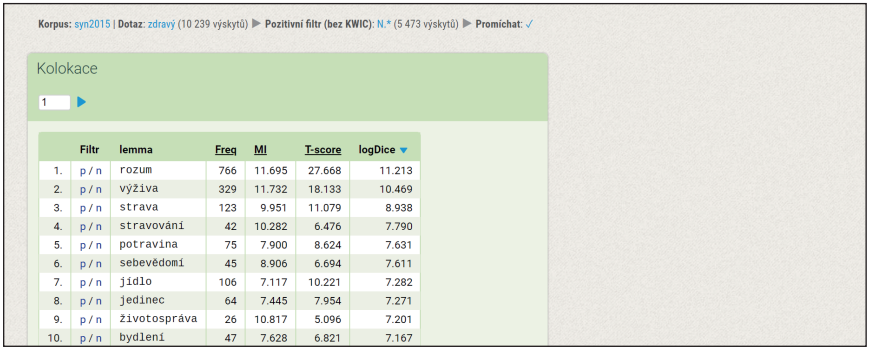
Obrázek 41

2. Až se vám zobrazí konkordance, vyberte možnost *Kolokace*, dále zvolte *Vlastní*. Jako *Atribut* si vyberte lemma, v kontextu od 0 do 1.



Obrázek 42

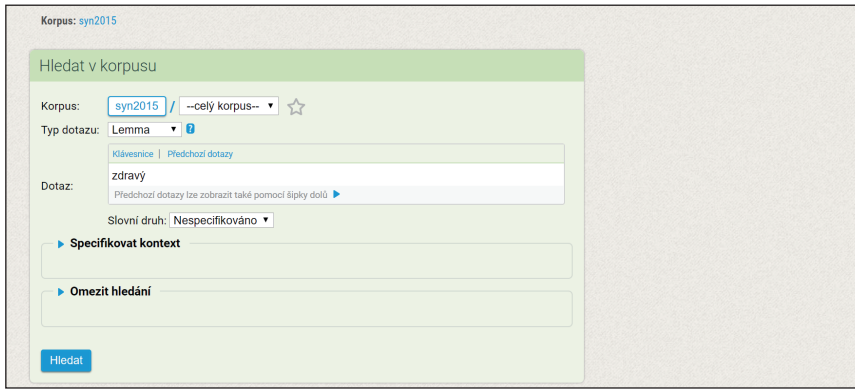
3. Zobrazí se frekvenční seznam nejčastějších substantiv, která se pojí s daným adjektivem a stojí ve větě vždy hned vedle něj.



Obrázek 43

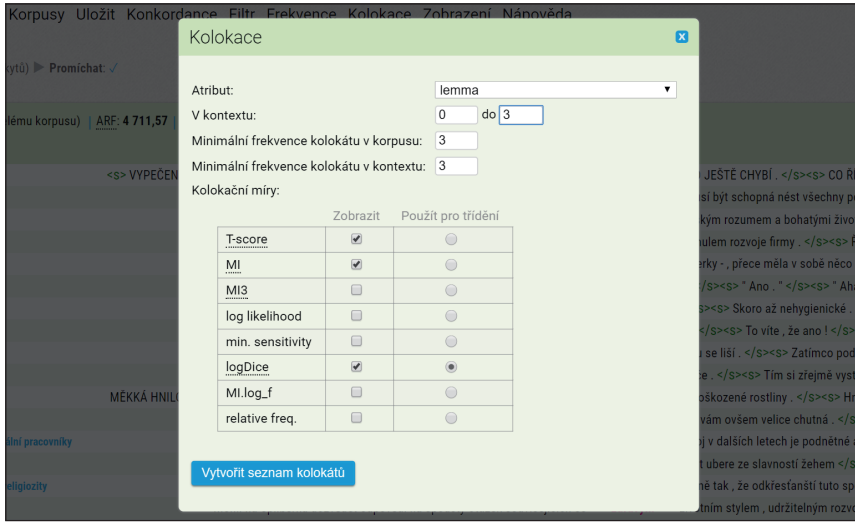
Postup 2

1. V korpusu *SYN2015* zvolte Typ dotazu lemma a do okénka *Dotaz* napište hledané adjektivum (*zdravý* nebo *zdravotní*).



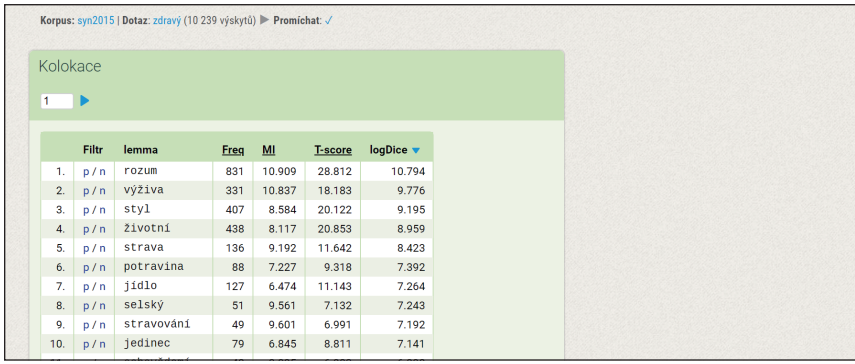
Obrázek 44

2. Až se vám zobrazí konkordance, vyberte možnost *Kolokace*, dále zvolte *Vlastní*. Jako *Atribut* si vyberte lemma, v kontextu od 0 do 3.



Obrázek 45

3. Zobrazí se vám frekvenční seznam nejčastějších slov následujících po daném adjektivu, a to až do třetí pozice od vlastního adjektiva. Protože jste na začátku nespecifikovali slovní druh, je třeba ze seznamu kolokátů ručně vybrat jen substantiva. Výhodou je, že najdete i kolokáty, které jsou dále od daného adjektiva (např. *zdravý životní styl*).



	Filtr	lemma	Freq	MI	T-score	logDice
1.	p / n	rozum	831	10.909	28.812	10.794
2.	p / n	výživa	331	10.837	18.183	9.776
3.	p / n	styl	407	8.584	20.122	9.195
4.	p / n	životní	438	8.117	20.853	8.959
5.	p / n	strava	136	9.192	11.642	8.423
6.	p / n	potravina	88	7.227	9.318	7.392
7.	p / n	jídlo	127	6.474	11.143	7.264
8.	p / n	selský	51	9.561	7.132	7.243
9.	p / n	stravování	49	9.601	6.991	7.192
10.	p / n	jedinec	79	6.845	8.811	7.141

Obrázek 46

4. Pokud chcete využít pozitivní filtr, klikněte na písmeno *p* ve sloupci *filtr*. Zde je příklad pozitivního filtru u substantiva *rozum* v seznamu nejčastějších kolokátů adjektiva *zdravý*.

Obrázek 47

Řešení:

1. Seznam deseti nejčastějších kolokátů

zdravý

první postup	druhý postup
rozum	rozum
výživa	výživa
strava	styl
stravování	strava
potravina	potravina
sebevědomí	stravování
jídlo	jídlo
jedinec	jedinec
životospráva	sebevědomí
bydlení	bydlení

zdravotní

první postup	druhý postup
pojišťovna	pojišťovna
postižení	postižení
pojištění	pojištění
stav	péče
péče	stav
sestra	sestra
potíž	potíž
problém	problém
komplikace	komplikace
služba	služba

Vysvětlením významu může být např. zjednodušená definice ze Slovníku spisovného jazyka českého (http://ssjc.ujc.cas.cz/):

zdravý = má bezproblémové zdraví (*zdravý jedinec, člověk*); přirozený (*zdravý rozum, zdravé sebevědomí*); dobrý pro zdraví (*zdravé jídlo, zdravé bydlení*)

zdravotní = týká se zdraví (*zdravotní problém, stav*); stará se o zdraví (*zdravotní sestra, pojišťovna*)

2. Spojení adjektiva a substantiva

zdravá holčička, zdravá noha, zdravý zub
zdravotní handicap, zdravotní obtíž, zdravotní prohlídka

5.2.2. Morfio

Anotace
Cvičení tohoto typu je vhodné zařadit pro znázornění různých slovotvorných modelů. Žáci si díky nim lépe uvědomí, že lze od známých slov odvozovat nová slova pomocí přípon. Cvičení lze libovolně variovat, je možné hledat např. slova odvozená předponami (např. *chod – vchod*) nebo slova odvozená předponami i příponami (např. *lov – úlovek*). V tomto pracovním listu půjde o adjektiva utvořená příponami *-ý* a *-otní* (např. *zdravý* a *zdravotní*).

Postup práce
Tento pracovní list může učitel zařadit po pracovním listu, ve kterém žáci v KonTextu vyhledávají podstatná jména, která se pojí s přídavnými jmény *zdravý* a *zdravotní*. V prvním úkolu si mají na tato podstatná jména vzpomenout a vhodně je spojit s adjektivy, ve druhém pak mohou opakovat definici z předchozího pracovního listu. Odpovědí na třetí otázku má být substantivum *zdraví*, ke kořenu *zdrav-* se připojují přípony *-ý* a *-otní*. Pokud žáci neznají názvosloví, může je s ním učitel tímto způsobem seznámit. Následně se učitel žáků zeptá, zda znají nějaká další slova, která vznikla stejným způsobem. Učitel si předem může připravit obrázky nebo věty s vynechanými slovy, která žáci hledají (viz řešení), a žáci se mohou sami pokusit zjistit, o která slova se jedná. Pravděpodobně se jim nepodaří určit všechna slova, a to by mělo sloužit jako motivace pro práci s korpusem. Žáci se sami nebo s učitelem podívají do aplikace Morfio a zjistí, která další slova jsou odvozena stejným způsobem. K některým přídavným jménům by měli být sami schopní utvořit příkladové kolokace. Pokud budou chtít, mohou si nejčastější kolokace najít pomocí rozhraní KontText stejným způsobem, kterým vyhledávali kolokace přídavných jmen *zdravý* a *zdravotní*. Pokud učitel s KonTextem v hodině pracovat nechce, může si nejčastější kolokáty najít předem a žákům je poté nabídnout v pátém úkolu (viz nabídka, zvolit lze samozřejmě i jiná slova). V pátém úkolu jsou úmyslně vynechána přídavná jména *prvý* a *prvotní*, která mají jinou sémantiku.

Pracovní list

ZDRAVÝ ŽIVOTNÍ STYL = ŽÁDNÉ ZDRAVOTNÍ PROBLÉMY?

1) Co může být *zdravé*? Co může být *zdravotní*?

2) Jak můžete definovat slova *zdravý* a *zdravotní*?

3) Od kterého podstatného jména tato přídavná jména vznikla? Jak?

4) Znáte další přídavná jména se stejnými příponami?

5) Použijte aplikaci Morfio a najděte další přídavná jména se stejnými příponami. Napište je a vymyslete k nim příklad.

6) Co může být *husté*, *teplé* a *živé*? Co *hustotní*, *teplotní* a *životní*? Dopište přídavná jména k podstatným jménům se správnou koncovkou.

prostředí, styl, cyklus, úroveň, pojištění, příběh

plot, organismus, tvor, bytost, sen, diskuse

koeficient, výkyv, rekord, extrém, stupnice

voda, počasí, vzduch, klima, jídlo, koupel

mlha, obočí, vlasy, srst, déšť, polévka, hranice, profil, kontrast

Postup práce s korpusem:

V první kolonce pro společnou bázi ponechte.+. U prvního vzoru doplňte do kolonky pro odlišný formant *-ý*, u druhého *-otní*. U obou zvolte morfologickou specifikaci přídavná jména. Vyberte, že se hledají a vyhodnocují lemmata. Velikost písmen můžete ignorovat. Nakonec klikněte na tlačítko *Hledat*. Pro zobrazení výsledků klikněte na *Výpis*.

Obrázek 48

Pro vyhledání kolokátů ve cvičení 5 použijte rozhraní KonText. Vyhledejte v něm v korpusu *SYN2015* příslušná lemmata (*teplý*, *teplotní*, *hustý*, *hustotní*, *zdravý*, *zdravotní*). Až se vám zobrazí konkordance, klikněte na *Kolokace*, zvolte *Vlastní* a jako *Atribut* si vyberte lemma, v kontextu od 0 do 3. Ze seznamu kolokátů si poté vyberte podstatná jména dle potřeb a schopností žáků.

Řešení:

1.
zdravý: např. *rozum*, *výživa*, *strava*, *stravování*, *potravina*, *sebevědomí*, *jídlo*, *jedinec*, *životospráva*, *bydlení*

zdravotní: např. *pojišťovna*, *postižení*, *pojištění*, *stav*, *péče*, *sestra*, *potíž*, *problém*, *komplikace*, *služba*

2.
Zjednodušená definice ze Slovníku spisovného jazyka českého (http://ssjc.ujc.cas.cz/):

zdravý = má bezproblémové zdraví (*zdravý jedinec, člověk*); přirozený (*zdravý rozum, zdravé sebevědomí*); dobrý pro zdraví (*zdravé jídlo, zdravé bydlení*)

zdravotní = týká se zdraví (*zdravotní problém, stav*); stará se o zdraví (*zdravotní sestra, pojišťovna*)

3.
Podstatné jméno zní *zdraví*. Vzniklo připojením přípon *-ý* a *-otní* ke kořenu *zdrav-*.

4.
hustý, hustotní; prvý, prvotní; teplý, teplotní; živý, životní

5.
životní: *prostředí, styl, cyklus, úroveň, pojištění, příběh*
živý*: plot, organismus, tvor, bytost, sen, diskuse*
teplotní*: koeficient, výkyv, rekord, extrém, stupnice*
teplý: *voda, počasí, vzduch, klima, jídlo, koupel*
hustý: *mlha, obočí, vlasy, srst, déšť, polévka*
hustotní: *hranice, profil, kontrast*

5.2.3. SyD

Anotace
Cvičení tohoto typu je vhodné zařadit do výuky, pokud chceme žákům ozřejmit použití dubletních tvarů, které se liší stylovou platností a/nebo příslušností k jiné vrstvě českého jazyka (nespisovné X spisovné). Tento list je věnován tvarům *děkuju* X *děkuji*, stejný postup lze využít např. pro tvary *péct/píct*, *moci/moct*, *píši/píšu*, *mohu/můžu*, *bychom/bysme* atd. Pracovní list vyžaduje samostatnou práci s korpusovou aplikací *SyD* a obsahuje tři úkoly, které vedou k tomu, aby si žák význam adjektiv osvojil vlastními silami, a to samostatným objevením pravidel a následně jejich aplikací.

Postup práce
Před obdržením pracovního listu by měli žáci vidět obě morfologické varianty a měli by si jich sami všimnout v textu. V tomto případě se učitel může na začátku zeptat, jak můžeme v češtině děkovat. Je pravděpodobné, že žáci budou znát obě varianty *děkuju* a *děkuji* a možná i další (viz řešení).

Učitel se poté zeptá žáků, kdy by kterou variantu použili. Může je vyzvat, aby si představili, že mluví s kamarádem/ředitelem/učitelem nebo že píší oficiální nebo neoficiální e-mail. Žáci možná budou sami vědět, že varianta *děkuji* je formálnější a častěji se používá v psaných textech. I přesto pro ně může být zajímavé podívat se do korpusu na rozložení v psaném a mluveném jazyku. Předtím než se do korpusu podívají, můžou hádat, jak výsledek dopadne – tj. kdy se častěji používá která varianta a jaký je jejich poměr. Poté žáci samostatně nebo ve dvojicích vyřeší úkol 2 a 3. Předem musejí znát postup práce s aplikací. Pokud žáky výsledky zaujmou, mohou se dále podívat na rozložení variant *děkuji* a *děkuju* v psaných textech. Zjistí tak, že varianta *děkuji* se nejčastěji používá v odborných textech, *děkuju* naopak v beletrii. Učitel se následně žáků může zeptat, která další slovesa mají takové dublety, a žáci se mohou podívat, zda se tyto dublety chovají stejně. Zajímavé je např. sloveso *pracovat*, které se v psaném jazyce v první osobě singuláru častěji objevuje ve formě *pracuji*, v mluveném jazyce se v korpusu vyskytuje jen podoba *pracuju*. Nakonec by žáci měli sami obě varianty vhodně použít. Učitel si předem může připravit různé situace, ve kterých mají žáci někomu ústně nebo písemně poděkovat a použít jednu nebo druhou dubletu. O vhodnosti použití lze samozřejmě diskutovat. Žáci by před vlastní produkcí měli znát nejčastější kolokace, a to především kolokaci *děkovat někomu za něco*. Kolokace najdou po kliknutí na stejnojmennou ikonu, která se v aplikaci zobrazí spolu s výsledky. Zde si můžou kliknutím na jednotlivá slova zobrazit příkladové věty. Kolokace samozřejmě mohou hledat také přes rozhraní KonText.

Pracovní list

DĚKUJU TI ZA DÁREK. DĚKUJI VÁM ZA POMOC.

1) Jak můžeme v češtině děkovat?

2) Co je typické pro psaní a co pro mluvení? Podívejte se do korpusu (https://syd.korpus.cz/).

Pro psaní je typické:

Pro mluvení je typické:

3) Doplňte do tabulky procenta z grafů.

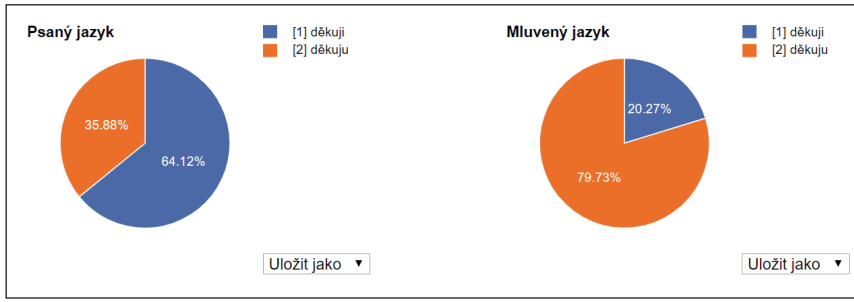
varianta	psaní	mluvení
děkuji		
děkuju		

4) Zahrajte se spolužákem minidialog. Spolužákovi musíte v dialogu za něco poděkovat.

5) Napište krátký e-mail učiteli nebo učitelce. V e-mailu musíte za něco poděkovat.

Postup práce s korpusem

Do jednoho okénka aplikace *SyD* zadejte tvar *děkuji*, do druhého *děkuju*. Nechte zaškrtnuté políčko a=A (nezáleží na tom, jestli je první písmeno malé nebo velké), ale nezaškrtávejte políčko lemma (chcete hledat jen tyto konkrétní tvary slovesa). Klikněte na tlačítko Porovnat varianty. Zobrazí se vám tento výsledek:



Obrázek 49

Řešení

1. děkuji, děkuju (děkuju/i moc, díky, dík, díkec, ď)
2. Pro psaní je typické *děkuji*. Pro mluvení je typické *děkuju*.

3.

varianta	**psaní**	**mluvení**
děkuji	64,12 %	20,27 %
děkuju	35,88 %	79,73 %
varianta	**psaní**	**mluvení**
děkuji	64,12 %	20,27 %
děkuju	35,88 %	79,73 %

5.2.4. KWords

Anotace
Cvičení tohoto typu pomáhají žákům při psaní textů určitého žánru nebo zadání. Pomocí srovnání příslušných textů (zde charakteristik) a vyváženého referenčního korpusu, který obsahuje texty všech žánrů, mohou žáci sami zjistit, která slova jsou pro dané texty typická. Tato slova poté mohou vědomě použít při psaní vlastního textu. Lze vytvořit také offline variantu, kdy si učitel klíčová slova najde předem a žákům je pouze prezentuje. Třetí úkol z tohoto pracovního listu by tak nevyžadoval práci s korpusem ve výuce, ostatní úkoly mohou žáci plnit stejným způsobem.

Postup práce
Pracovní list by měl být zařazen do výukového bloku, během kterého si mají žáci osvojit dovednosti potřebné k psaní textu určitého žánru. Tento pracovní list je věnován charakteristice, lze jej ovšem obměnit pro potřeby jakéhokoliv jiného žánru. Než žáci začnou s pracovním listem pracovat, měli by se seznámit s tím, jak charakteristika vypadá. Na začátku výukového bloku může učitel pomocí brainstormingu zjistit, co si žáci pod pojmem charakteristika vybaví (úkol 1). Sami se mohou pokusit formulovat, co by měla dobrá charakteristika obsahovat a jak by měla být strukturována. Své domněnky by si poté měli ověřit na vhodně zvolené příkladové charakteristice, kterou by si měli přečíst a společně s učitelem okomentovat z hlediska použité slovní zásoby, struktury i obsahu (úkol 2).[44] Práce s korpusem může být zařazena na toto místo pro zpestření a jako alternativní způsob nacházení klíčových slov typických pro daný žánr. Učitel by se žáků nejprve měl zeptat, která klíčová slova nacházejí v příkladovém textu. Žáci by si je měli vyznačit a poté společně s učitelem diskutovat o tom, zda to jsou opravdu slova typická pro daný žánr. Dále je možné využít vlastní práci s korpusem. Učitel si předem připraví soubor textů daného žánru, které následně s žáky zanalyzuje pomocí aplikace *KWords*. Textů musí být několik, aby obsahovaly dostatek materiálu pro analýzu. Ideální je rozsah několika set až několika tisíc slov. Příliš dlouhé texty analyzuje program moc dlouho.[45] Při vkládání je nutné odstranit informaci, odkud byly texty zkopírovány, jinak by se názvy stránek pravděpodobně objevily mezi klíčovými slovy. Žáci mohou samozřejmě najít texty sami během hodiny, v tom případě je ovšem třeba počí-

44 Okomentované příklady textů různých žánrů, které mají být žáci schopni napsat během maturity z češtiny, naleznete například na stránce http://www.statnimaturita-cestina.cz.

45 V tomto pracovním listu byly použity texty ze stránek http://www.statnimaturita-cestina.cz/charakteristika a http://www.cesky-jazyk.cz/slohovky/charakteristiky/#axzz5IEYwRMn2.

tat s vyšší časovou náročností. Výhodou je naopak to, že se žáci naučí hledat zdroje inspirace a také budou během práce aktivnější. Alternativně může učitel shromáždit soubor prací žáků s češtinou jako mateřštinou. Na žáky s OMJ může působit motivačně, že analyzují texty svých spolužáků.

Analýzu mohou žáci provádět samostatně, jednodušší je ovšem práce v plénu. Výsledkem analýzy jsou texty s červeně vyznačenými klíčovými slovy. Tato slova lze také zobrazit ve formě frekvenčního seznamu (viz postup práce s korpusem). Žáci si mohou na výsledcích ověřit, zda se nalezená klíčová slova shodují s jejich představou o slovní zásobě typické pro daný žánr. V klíčových slovech se nejspíš objeví slova ze skupiny vzhled, vlastnosti, zájmy, činnosti a jména. Pravděpodobně žáci naleznou také časová určení. Tato slova by měli shromažďovat do skupin a říct, proč se v charakteristice objevují.

Aplikace umožňuje také zobrazit, jak se jednotlivá klíčová slova propojují (viz postup práce s korpusem). Na analýzu a interpretaci výsledků by měla navázat vlastní produkce. Žáci by měli napsat vlastní charakteristiku a použít např. deset klíčových slov, která nalezli.

Pracovní list

1. Co je typické pro žánr charakteristiky? O čem se v charakteristikách nejčastěji píše?

2. Přečtěte si text charakteristiky. Najděte v ní slova a obraty, které jsou pro charakteristiku typické.

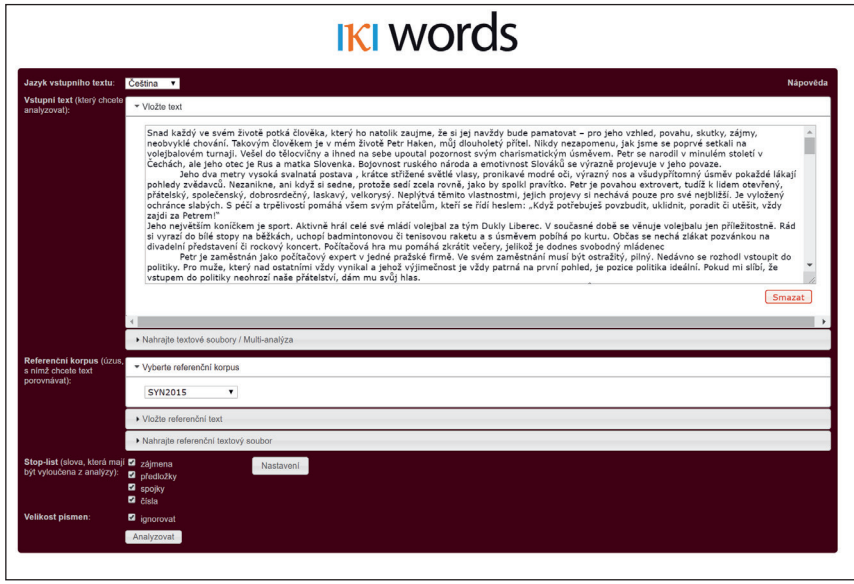
3. Porovnejte v aplikaci KWords texty charakteristik s ostatními texty. Najděte klíčová slova typická pro charakteristiku a seskupte je do skupin podle významu. Vymyslete pro každou skupinu název a doplňte další příklady.

např. *vzhled: vlasy, nos, světlé*

4. Napište vlastní charakteristiku a použijte v ní alespoň deset klíčových slov.

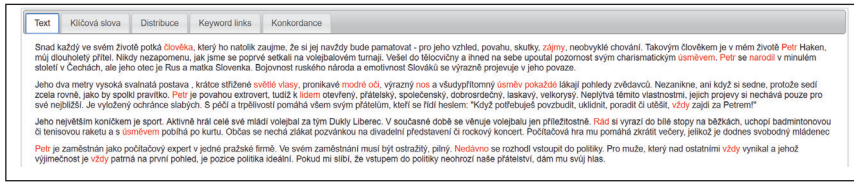
Postup práce s korpusem

1. Vstupní text, který chcete analyzovat (v tomto případě texty charakteristik v rozsahu alespoň 500 slov), zkopírujte do příslušného políčka aplikace *KWords*, které se rozbalí, když myší najedete na ikonu Vložte text. Jako referenční korpus ponechte *SYN2015*. Pomocí stop-listu můžete z analýzy vyloučit všechna nabízená slova (zájmena, předložky, spojky, čísla). Ponechte zaškrtnuté ignorovat velikost písma. Nakonec klikněte na *Analyzovat*.



Obrázek 50

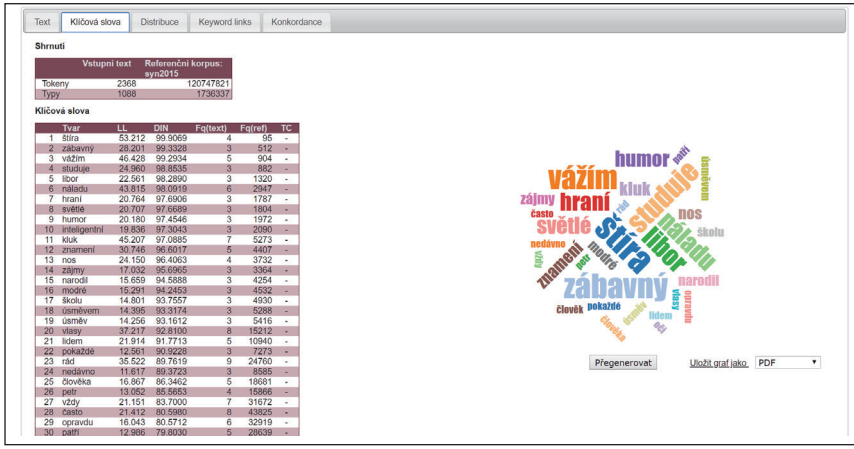
2. Zobrazí se vám analyzovaný text s červeně vyznačenými klíčovými slovy, tedy se slovy, která se v analyzovaném textu vyskytují relativně častěji než v textech z referenčního korpusu.[46]



Obrázek 51

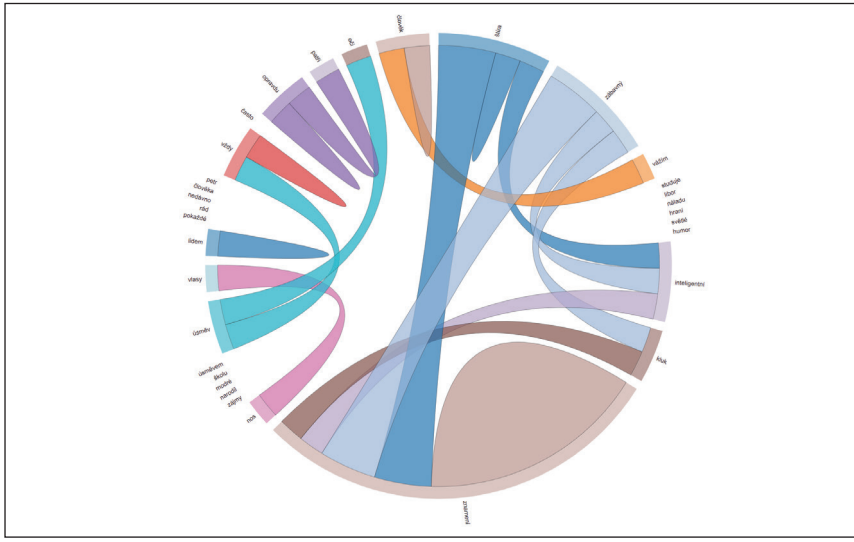
46 Příklad je text ze stránky http://www.statnimaturita-cestina.cz/charakteristika.

3. Pokud chcete zobrazit klíčová slova jako frekvenční seznam, klikněte na záložku *Klíčová slova* v horní liště. Zobrazí se vám jednak seznam, jednak barevný graf.



Obrázek 52

4. Pokud kliknete na záložku *Keyword links*, zobrazí se vám graf, který znázorňuje, jak jsou jednotlivá klíčová slova propojena, tedy v jakém kontextu se vyskytují.



Obrázek 53

Řešení

Výsledky úkolů 2 a 3 závisí na zvolených vzorových textech. V textech zvolených pro výše uvedenou analýzu byla nalezena tato klíčová slova:

štíra, vážím, zábavný, studuje, Libor, náladu, hraní, světlé, humor, inteligentní, kluk, znamení, nos, zájmy, narodil, modré, školu, úsměvem, úsměv, vlasy, lidem, pokaždé, rád, nedávno, Tomáš, člověka, Petr, vždy, často, opravdu, patří, člověk, oči

V našem případě je na prvním místě slovo *štíra*, které odkazuje na znamení. Z toho lze usuzovat, že se v charakteristikách často mluví o znamení daného člověka. Na prvním místě je proto, že názvy znamení nejsou tak typické pro texty ostatních žánrů. Dalším slovem je *vážím*, jelikož zadání vybraných charakteristik znělo *Charakteristika člověka, kterého si vážím*. Ostatní slova lze shromáždit do skupin, např. *vlastnosti* (*zábavný*, *inteligentní, humor*), *vzhled* (*světlé vlasy*, *nos*) nebo *časová určení* (*pokaždé*, *nedávno*, *často*), která ukazují na to, že součástí charakteristiky bývá i vyprávění nějakého zážitku, který nás s daným člověkem pojí. Skupiny lze samozřejmě rozšiřovat o další příklady.

Použitá literatura

Cvrček, V. a kol. (2010). Mluvnice současné češtiny. Praha: Karolinum.

Čermák, F. – Schmiedtová, V. (2004): Český národní korpus – základní charakteristika a širší souvislosti. In: Národní knihovna, knihovnická revue 15, 2004, č. 3., str. 152–168.

Čermák, F. (2011). Korpusy včera, dnes a zítra. In: Korpusová lingvistika Praha 2011 2 Výzkum a výstavba korpusů (eds. F. Čermák a kol.). Praha: Nakladatelství Lidové noviny.

Český národní korpus – základní charakteristika a širší souvislosti. *Národní knihovna, knihovnická revue*, 15, 2004, č. 3., str. 152–168

Ellis, R. – Barkhuizen, G. (2005). Analysing Learner Language. Oxford: Oxford University Press.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In: Corpora and Language Teaching (ed. K. Aijmer). Amsterdam and Philadelphia: Benjamins, 2009, str. 13–32.

Han, N.-R. et al. (2010). Using an Error-Anotating Learner Corpus to Develop an ESL/EFL Error Correction System. Dostupné online 13. 12. 2018: https://www.cs.rochester.edu/~tetreaul/han-lrec10-final.pdf

McEnery, T. – Hardie, A. (2012). Corpus Linguistics. Method, Theory and Practice. Cambridge: Cambridge Universitŷ Press.

Sinclair, J. (Eagles, 1996). Preliminary recommendations on Corpus Typology. Dostupné z http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Slovník spisovného jazyka českého [online]. Dostupný z http://ssjc.ujc.cas.cz/

Šebesta, K. – Škodová, S. (2012). Čeština – cílový jazyk a korpusy. Liberec: Technická univerzita v Liberci.

Šebesta, K. (2010). Korpusy češtiny a osvojování jazyka. In: Studie z aplikované lingvistiky / Studies in Applied Linguistics, 2, str. 11–33.

Štindlová, B. (2011). Žákovský korpus češtiny a evaluace jeho chybové anotace. Praha: FF UK.

Šulc, M. (1999). Korpusová lingvistika. První vstup. Praha: Karolinum.

Turton, N.D. and J.B. Heaton (1996). Longman Dictionary of Common Errors. Harwich: Longman.

http://akces.ff.cuni.cz/

http://wiki.korpus.cz/doku.php/cnk:syn2015

https://wiki.korpus.cz/doku.php/cnk:uvod

http://www.webcorp.org.uk/live/
https://kontext.korpus.cz/first_form
https://lindat.mff.cuni.cz/cs/
Learner corpora around the world, dostupné 13. 12. 2018: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html
https://merlin-platform.eu/
http://www.cesky-jazyk.cz/slohovky/charakteristiky/#axzz5IEYwRMn2

Profily autorů

PhDr. Tomáš Gráf, Ph.D. (tomas.graf@ff.cuni.cz) vystudoval anglistiku na FF UK, kde od roku 2010 působí jako odborný asistent v Ústavu anglického jazyka a didaktiky. Ve svém výzkumu se zabývá žákovským jazykem především z pohledu přesnosti a plynulosti. Vyučuje jazykovou akvizici a předměty týkající se didaktiky anglického jazyka a přípravy budoucích učitelů angličtiny.

Mgr. Věra Hejhalová, Ph.D. (vera.hejhalova@ff.cuni.cz) je česká germanistka působící v Ústavu germánských studií FF UK. Vystudovala Pedagogickou fakultu Jihočeské univerzity v Českých Budějovicích (Učitelství pro střední školy, aprobace německý jazyk – latina). Na Filozofické fakultě Univerzity Karlovy se profilovala v doktorském programu germánské jazyky a literatury. Vyučuje korpusovou lingvistiku, didaktiku němčiny jako cizího jazyka, gramatiku a frazeologii němčiny. Centry jejího badatelského zájmu jsou didaktika němčiny jako cizího jazyka, korpusová lingvistika a frazeologie němčiny.

Mgr. Barbora Kukrechtová (barbora.kukrechtova@ff.cuni.cz) je absolventkou magisterského studijního oboru učitelství češtiny jako cizího jazyka v Ústavu českého jazyka a teorie komunikace FF UK (DP „Mlčeti stříbro, mluviti zlato. Rozvíjení mluvních dovedností v současných učebnicích češtiny pro cizince“). Pracuje jako lektorka češtiny pro cizince v Ústavu bohemistických studií se zaměřením na ústní a písemné vyjadřování. Zároveň se jako doktorandka ÚČJTK zabývá využitím jazykových korpusů ve výuce češtiny pro cizince, a to především s ohledem na kolokace sloves pohybu. Vyučuje také v magisterském programu Učitelství češtiny jako cizího jazyka.

Prof. PhDr. Karel Šebesta, CSc. (karel.sebesta@ff.cuni.cz) působí v Ústavu českého jazyka a teorie komunikace FF UK, pedagogicky i badatelsky se zaměřuje na aplikovanou lingvistiku a didaktiku (českého) jazyka jako jazyka prvního i druhého/cizího. Je mj. předsedou oborové rady doktorského studijního oboru Didaktika konkrétního jazyka, iniciátorem a vedoucím dlouhodobého projektu AKCES – Akviziční korpusy českého jazyka, prvního komplexu speciálních korpusů sloužících studiu osvojování a dalšího vývoje českého jazyka i pedagogickým účelům, autorem řady publikací a učebnic v oboru.

Mgr. Kateřina Šormová, Ph.D. (katerina.sormova@ff.cuni.cz) je absolventkou doktorského studia na FF UK, obor český jazyk. V současné době působí jako odborná asistentka v Ústavu českého jazyka a teorie komunikace FF UK, vyučuje zejména v programech navazujícího magisterského studia Učitelství českého jazyka a literatury a Učitelství češtiny jako cizího jazyka, zaměřuje se na osvojování češtiny jako prvního a druhého/cizího jazyka, na čtenářskou gramotnost a na jazykové testování a hodnocení včetně sebehodnocení. Spolupodílela se na vybudování databanky jazykových projevů romských mluvčích češtiny ROMi a korpusu CZESL.

Rejstřík

AKCES 14, 15
akviziční korpusy 8
akviziční korpusy češtiny 12, 13, 14, 16, 17, 33, 42, 43, 47
angličtina pro pracovní účely 19
angličtina pro akademické účely 19
anotace 17, 22, 23, 26, 29, 31, 34, 37, 44, 45, 46
anotační systémy 17
atrice 12
autentičnost 34, 40

Bilingual Corpus of Chinese English Learners 37
British Academic Spoken English 20
British Academic Written English 20
British National Corpus 23, 27, 38

Cambridge Learner Corpus 18, 24, 25, 36, 37
Compar 15
Corpus of Learner English 20
corpus query language 50
CQL 50, 61
CZESL 15, 37, 39, 42, 43, 44, 45, 46, 47

část slova 50
české žákovské korpusy 37
Český národní korpus 35, 47
čítanky 25

DIAKORP 14
disambiguace 44, 45

emendace 22, 26, 37, 45, 46
English for Academic Purposes 19, 20
English for Occupational Purposes 19
English Grammar Profile 18, 25
English Vocabulary Profile 18, 25, 28
explicitní segmentace textu 43

falešní přátelé 26
faux amis 26
fráze 28, 50, 52
frekvence 24, 26, 27, 54, 57, 59, 67, 75

CHILDES 17
Chungdahm Corpus 36, 37

i.p.m. 54, 58
Interactive Spoken Language Education 41
InterCorp 14, 47, 70, 71, 72

jazyk dětí a mládeže 15
jazyk dětí předškolního věku 15
jazyk krajanských komunit 15
jazyk nerodilých mluvčích češtiny 15
jazyk sociokulturně znevýhodněných skupin 15
jazyk ve vzdělávacím kontextu 15
Just the Word 27

kolokace 27, 28, 29, 35, 58, 59, 60, 64, 84, 85, 89, 91, 93
konkordance 54, 55, 70, 71, 84, 85, 91
KonText 47, 48, 60, 75, 79, 89, 91, 93
kookurence 58
korpus 11, 12, 13, 14
korpusové pozice 43, 44
korpusy abstraktů 32
korpusy atriční 13
korpusy diachronní 14, 47, 61
korpusy obecné 11, 13, 14, 34, 35, 37, 39
korpusy osvojování mateřštiny 17
korpusy speciální 11, 12, 13, 14, 16, 20, 34, 40
korpusy synchronní 14, 35, 37, 38, 45, 47, 61
korpusy terapeutické 13, 15
korpusy zaměřené na jazyk krajanů 12
korpusy žákovské 12, 15, 16, 20, 22, 25, 35, 37, 39, 40, 42, 43
krajanské korpusy 13
KWords 60, 67, 68, 69, 70, 96, 98, 99

learner corpora 16, 17, 34
lemma 44, 45, 46, 50, 52, 57, 59, 61, 66, 70, 83, 84, 85, 91, 95
lemmatizace 37, 45
LINDAT/CLARIN 15
lingvistická anotace 34

Longman Grammar of Spoken and Written English 30

MERLIN 15
metatextové informace 34, 43
mezijazyk 16, 21, 23, 39
Michigan Corpus of Academic Spoken English 20
Michigan Corpus of Upper-level Student Papers 20
Morfio 60, 65, 66, 67, 89, 90
morfologická analýza 44, 45
morfologické značkování 22, 45

nedokončené segmenty 32
nestandardní jazyková data 13, 36, 45

opakované segmenty 32

paralelní korpus 14, 47
pauzy 30, 32
plynulost 27, 30, 31, 32, 33
poziční atributy 44
pragmatika 31
Pražský závislostní korpus 14
projevy vernakulární 42
promíchat 55

regulární znaky 50, 51
ROMi_1.0 15, 42

SERR 17, 18, 39, 44
seznamy slovní zásoby 18, 25
SCHOLA 2010 15
slovní tvar 57, 61, 66, 69, 70
specializované korpusy mluveného jazyka 14
specializované korpusy psaného jazyka 14
správná varianta 22
strukturní atributy 34, 42, 43, 44
SyD 60, 61, 62, 63, 64, 93, 94, 95
synchronní korpusy mluveného jazyka 14
synchronní korpusy psaného jazyka 14

tempo mluvy 32
tokenizace 44
tokeny 43, 44
Treq 60, 70, 71, 72
typ dotazu 49, 52, 54, 56, 75, 83, 85

Varieties of English for Academic Purposes 20
velikost 29, 34, 35, 36, 37, 38, 39, 40, 54, 55
videokorpusy 21
vyváženost 34, 35, 37
vzorek 11, 55

Web jako korpus 35
WebCorp 35
wiki 49, 50, 59, 64, 67, 69, 71

žáci s odlišným mateřským jazykem 23, 26, 61
žáci s OMJ 47, 65, 67, 70, 97

Resumé/Summary

Corpora in language teaching

The present monograph is written in the belief that corpus linguistics offers a wide array of tools which are useful and directly useable by L1 and L2 language teachers. It invites them to explore language using basic exploratory corpus linguistics methods and to use these as tools both for learning about language and for designing materials and activities which will guide students to learn about language and increase their linguistic competence. The specific target group for this publication is those teachers whose classes include students with mother tongues different from the class majority (Czech) and thus often with a lower-than-native command of that language. It is especially these students who need careful language work, as the improvement of their linguistic skills in the language of school instruction has a direct bearing on their success at school.

Acknowledging the fact that corpus linguistics is a field with which Czech teachers may still be largely unfamiliar, the book, first of all, provides a concise introduction to corpus linguistics. Starting, in Chapter 1, by describing and defining what language corpora are and outlining what types of corpora exist and how they have contributed to language teaching, it then explores the area of corpus research and its relevance to language pedagogy. In Chapter 2 it discusses acquisitional corpora and the link between learner corpus research and such fields as second language acquisition, language testing and assessment and language for specific and academic purposes. It further describes how corpus linguistics can contribute to learner-language analysis especially with regard to grammar, vocabulary, phraseology, discourse, pragmatics, fluency and accuracy and discusses its impact on language teaching. Chapter 3 furthers our understanding of the field by focussing on such key parameters of language corpora as their size, the authenticity of the recorded language they contain, their structural and positional attributes, and methods and implications of corpus annotation.

The latter half of the publication aims to introduce teachers to concrete techniques of exploiting language corpora. It does so by introducing the teachers to the Czech National Corpus and the tools which are available through its related web pages, and progresses methodically from the basics of formulating queries and setting various attributes and filters to wor-

king with the concordance listing and accompanying web apps, which are directly exploitable in work with students with different L1s, such as SyD, Morfio, KWords, and Treq.

One of the book's most interesting and useful aspects is the inclusion of sample practical worksheets. The authors have designed these to illustrate how corpus tools can be used for the development of teaching materials whilst bearing in mind the characteristic needs of the target group of students, i.e. the need of many of them to improve their knowledge of Czech in order to be able to succeed in the Czech educational system. The exercises include a variety of didactic approaches, and each worksheet is accompanied by a key which includes both the instructions for carrying out the task using one of the corpus tools and a possible solution. The book thus also offers highly practical concrete solutions for teachers who need to prepare language exercises to aid students' progress in Czech as a second language.

Kateřina Šormová, Karel Šebesta a kol.
Korpusy v jazykovém vyučování

Vydala Univerzita Karlova, Filozofická fakulta,
nám. Jana Palacha 2, Praha 1

Obrázky Marek Šorm
Typografická osnova František Štorm
Sazba z písma Skolar Dušan Neumahr
Vyrobila Togga, spol. s r. o., Praha
Vydání první, Praha 2019

Tomáš Gráf

Věra Hejhalová

Barbora Kukrechtová

Karel Šebesta

Kateřina Šormová

FILOZOFICKÁ FAKULTA
Univerzita Karlova

ISBN 978-80-7308-897-2