# Využití webového korpusu Araneum při tvorbě nového výkladového slovníku češtiny

**Zdeňka Opavská, Vít Michalec, Pavla Kochová, Magdalena Kroupová, Michaela Lišková, Barbora Procházková, Barbora Štěpánková**

*Ústav pro jazyk český AV ČR, v. v. i.*
*E-mail: {opavska, michalec, kochova, kroupova, liskova, bprochazkova, stepankova}@ujc.cas.cz*

**Abstrakt**

Při tvorbě nového Akademického slovníku současné češtiny (ASSČ) se využívá kombinovaná materiálová základna. Primárním materiálovým zdrojem je synchronní korpus psaných textů SYN ÚČNK. V případě, že *korpus SYN poskytuje pro lexikografické zpracování lexikální jednotky nedostatečné, nepřesvědčivé nebo nejednoznačné údaje,* jsou pro dohledávání jazykových dat a jejich ověřování využívány další materiálové zdroje, a to korpusové i nekorpusové. O webové korpusy Aranea, jež byly vytvořeny pro češtinu, byla materiálová základna ASSČ rozšířena po jejich zpřístupnění v rámci ÚČNK. Díky největšímu z nich, korpusu Araneum Bohemicum Maximum, a to zejména v jeho dosud největší publikované verzi *Araneum Bohemicum III Maximum (dostupné mimo ÚČNK), získali zpracovatelé ASSČ velmi objemný zdroj dat, která je možno efektivně analyzovat prostřednictvím korpusových nástrojů. Zároveň se jedná o významný zdroj textů patřících do sféry běžné komunikace, a tedy o zdroj v jistém smyslu komplementární ke korpusu SYN (s převahou publicistických textů). Na základě dosavadní lexikografické práce lze konstatovat, že se korpus Araneum stává důležitým ověřovacím zdrojem pro řešení problémů týkajících se jak slovníkové makrostruktury, tak mikrostruktury. Jeho přínosnost pro tvorbu ASSČ ilustrujeme konkrétními příklady zaměřenými na problematiku výběru lexémů/lexií a na uvádění gramatické informace.*

**Klíčová slova:** korpus SYN; korpus Araneum; Akademický slovník současné češtiny; webové korpusy; výkladový slovník

## 1 O Akademickém slovníku současné češtiny (ASSČ)

Akademický slovník současné češtiny (ASSČ) připravovaný od roku 2012 v oddělení současné lexikologie a lexikografie ÚJČ AV ČR je novým všeobecným výkladovým slovníkem, s plánovaným rozsahem 120–150 tisíc hesel. Jeho cílem je popsat slovní zásobu současné češtiny, užívanou ve veřejné oficiální a poloficiální komunikaci i v komunikaci běžné (tj. neveřejné, neoficiální). Současná čeština je vymezena jazykem 3 žijících generací, orientačně rokem 1945. Podrobně ke koncepci slovníku viz Kochová, Opavská (2016).

## 2 Materiálová základna ASSČ

Na rozdíl od předcházejících akademických všeobecných výkladových slovníků [1] se materiálovou základnou ASSČ nestala cílená excerpce, ale elektronické texty, zejména korpusy ÚČNK. Primárním materiálovým zdrojem je synchronní korpus psaných textů SYN ÚČNK, v současné době ve verzi 6 (SYN6). Zdroji p*ro dohledávání a ověřování jazykových dat v případě, že korpus SYN nepostačuje (tj. poskytuje nedostatečné, nepřesvědčivé nebo nejednoznačné údaje),* jsou korpusy řady ORAL ÚČNK, mediální archiv společnosti Newton Media, a. s., a internetové texty[2]. Doplňkově jsou využívány databáze a sbírky oddělení současné lexikologie a lexikografie ÚJČ AV ČR. O webové korpusy Aranea (Araneum Bohemicum Minus, Araneum Bohemicum Maius, Araneum Bohemicum Maximum) byla materiálová základna ASSČ rozšířena po jejich zveřejnění v rámci ÚČNK.[3] Využíváním korpusu Araneum Bohemicum Maximum, včetně dosud nej*větší verze Araneum Bohemicum III Maximum (ARIII)* [4] , se zefektivnila lexikografická práce ve srovnání s prací s nestrukturovanými, resp. nedostatečně strukturovanými nekorpusovými elektronickými zdroji – k datům se lexikografové dostávají snáze, rychleji. Webová data je možné podrobovat analogickým systémovým analýzám jako u korpusů psaných textů SYN a získávat porovnatelné výsledky ze dvou materiálově různých zdrojů.

## 3 Zásady zařazování lexikálních jednotek / lexií do ASSČ

Při tvorbě hesláře ASSČ využíváme trojici kritérií: frekvenční kritérium, kritérium rozšíření v úzu a systémové kritérium. Žádné z nich není primární, uplatňují se vždy v kombinaci.

Heslář ASSČ vzniká a) vyhodnocením hrubého hesláře, který byl vytvořen ze subkorpusu skládajícího se z reprezentativních korpusů SYN2000, SYN2005, SYN2010 (lexémy s frekvencí 5 a více) a b) jeho metodickým doplněním o další lexikální jednotky na základě slovotvorných a sémantických vztahů. Oba uvedené postupy se provádějí na korpusu SYN, v případě jeho nedostatečnosti na dalších zdrojích (v současné době nabývá na významu korpus Araneum). Obecným, základním frekvenčním kritériem pro zařazení lexikální jednotky do hesláře je minimálně 5 výskytů v korpusu SYN. Frekvenční kritérium je doprovázeno analýzou zdrojů (nezařazují se např. jednotky s výskytem pouze v jednom zdroji nebo v limitovaném počtu zdrojů).

Limitovaný výběr se provádí u některých lexikálněsémantických skupin. Přísnějším kritériím podléhají zvláště terminologické lexémy, a to vzhledem k jejich velkému počtu a povaze. Nezařazují se terminologické jednotky a lexie s nižší frekvencí a/nebo vázané čistě na odborné zdroje.

---

[1] Příruční slovník jazyka českého, Slovník spisovného jazyka českého, a Slovník spisovné češtiny pro školu a veřejnost.

[2] Ke složení webu z hlediska typů textů a k jeho využití jako slovníkového zdroje viz Grouws et al. (2013: 1366n.). O nevýhodách využívání webu pro lingvistickou práci a o webových korpusech viz Čermák (2017: 83).

[3] Benko (2014a, b, c). – Od počátku byl využíván největší z nich, tj. Araneum Bohemicum Maximum. – Pro češtinu existuje rovněž webový korpus czTenTen (viz Suchomel 2012). Je dostupný v rámci aplikace Sketch Engine; ke korpusům řady TenTen viz Jakubíček et. al. (2013). Korpus czTenTen není při tvorbě ASSČ využíván.

[4] Je dostupná m*imo ÚČNK, a tedy i mimo prostředí korpusového manažeru KonText.*

V omezené míře se dale zařazují a) profesionalismy a slangismy, které nejsou vázány čistě na profesní a zájmovou komunikaci; b) dialektismy a regionalismy, které splňují požadavek širší územní platnosti a které více pronikají do beletristických a publicistických textů; c) neologismy, které se staly pevnou součástí slovní zásoby, tedy nejsou omezeny pouze na krátkou dobu užívání nebo na určitý dobově ohraničený společenský kontext.

Co se týče slovotvorných hnízd nebo otevřených slovotvorných řad, jedná se buď o rozšiřování hesláře, nebo o jeho redukci. Slovotvorná hnízda jsou systémově doplňována o lexikalizované a uzualizované deriváty, tj. jejich zařazování není dáno pouze frekvencí (jedná se např. o adverbia, přechýlené názvy). Redukce se týká členů otevřených slovotvorných řad tvořených zejména radixoidními kompozity a předponovými deriváty. Zde se uplatňuje vyšší frekvenční hranice[5].

## 4 ASSČ a přístup corpus-based vs. corpus-driven

ASSČ je slovník, jehož přístup ke korpusovým datům v podstatě odpovídá přístupu corpus-based. O tom zejména svědčí koncepce tohoto slovníku, jež vychází z dosavadních teoretických poznatků jednotlivých lingvistických disciplín, nevyužívání pouze korpusových dat, přihlížení ke zpracování lexikálních jednotek ve výkladových slovnících jak českých, tak zahraničních. Přístup corpus-driven se více uplatňuje při výstavbě hesláře (viz výše), při výběru typických kolokací do exemplifikace, do jisté míry při identifikaci lexikograficky dosud nezachycených významů. Z těchto důvodů můžeme hovořit o kombinaci obou přístupů[6]. Je však třeba zdůraznit, že data získaná z korpusu jsou vždy vyhodnocována na základě stanovené teoretické koncepce slovníku. Korpusová data mohou tuto koncepci korigovat, ale nevytvářejí ji.

## 5 Využití webového korpusu Araneum při tvorbě ASSČ

V úvodu jsme konstatovali, že materiálová základna ASSČ je tvořena kombinací korpusových a nekorpusových zdrojů, přičemž korpus SYN je primárním materiálovým zdrojem a trojice reprezentativních korpusů SYN2000, SYN2005 a SYN2010 tvoří základ hesláře. S ostatními zdroji (tj. jinými, než jsou korpusy řady SYN) se pracuje v okamžiku, kdy korpus SYN pro lexikografická rozhodnutí poskytuje nedostatečná data. Taková byla situace v roce 2012 (v době počátku lexikografického projektu) a taková je i situace současná, a to i přes kontinuální nárůst rozsahu korpusu SYN[7]. Od zpřístupnění webového korpusu Araneum ustupuje využití nekorpusových zdrojů do pozadí. Oproti nim totiž korpus Araneum poskytuje data ve strukturované podobě a při jejich analýze lze použít nástroje pro lingvistickou analýzu, podobně jako je tomu při práci s korpusem SYN. K nevýhodám využívání webového korpusu náleží nemožnost podrobněji analyzovat typy zdrojů a fakt, že jde o směs korigovaných i

[5] Např. pro výrazy začínající komponentem *dvoj-/dvou-* byla stanovena hranice přibližně 100 dokladů v korpusu SYN. Při aplikaci frekvenčního kritéria je třeba zvláště přihlížet k opakování lexikální jednotky v jednom nebo několika málo zdrojích.

[6] Srov. Kosem (2016: 78–79). K vymezení přístupu corpus-based a corpus-driven v korpusové lingvistice viz Tognini-Bonelli (2001), v české lingvistice srov. např. Nový encyklopedický slovník češtiny (Karlík, P, Nekula M. & Pleskalová J., 2016; autorem hesel corpus-based výzkum, corpus-driven výzkum je V. Cvrček).

[7] V současné době je k dispozici verze 6 z 18. 12. 2017 s rozsahem 4 834 739 998 pozic.

nekorigovaných textů, což komplikuje vyhodnocení dat zvláště v tak velkém měřítku. Na druhou stranu nekorigované texty poskytují obraz soudobého jazykového úzu, a to zejména ve sféře běžné každodenní komunikace. Tím se webový korpus Araneum stává vůči korpusu SYN, v němž převažují publicistické texty, komplementárním zdrojem, neboť obsahuje velké množství jazykových projevů neformálních, nepřipravených. Proto je vhodným ověřovacím zdrojem pro lexikální jednotky vázané na běžně mluvené projevy, např. citoslovce, některá synsémantika, frazémy. Jako ověřovací zdroj také slouží v případě lexikálních nebo morfologických variant, některých valencí nebo lexikálních jednotek či lexií, u nichž se v ASSČ uplatňuje limitovaný výběr (termíny, neologismy apod.). Z hlediska lexikografického tak můžeme konstatovat, že využití webového korpusu Araneum má dopad jak na výstavbu hesláře, tak na výstavbu heslové stati, ovlivňuje zejména řešení problému zařazení/nezařazení určité lexikální jednotky / lexie do slovníku, dále pomáhá při zjišťování a ověřování vybraných gramatických informací a při hledání vhodných exemplifikačních dokladů. V příspěvku se zaměříme na dva z uvedených aspektů, totiž na otázku zařazování lexémů/lexií do slovníku a na problematiku podávání gramatické charakteristiky u příznakových jednotek.

## 5.1 Zařazování lexikálních jednotek / lexií do ASSČ

Korpus Araneum se využívá jako ověřovací zdroj při rozhodování o zařazení/nezařazení lexikální jednotky / lexie[8] do hesláře, resp. do slovníku[9]. V příspěvku tento problém ilustrujeme na problematice přechýlených názvů a zdrobnělin (5.1.1), významů uvedených v předcházejících výkladových slovnících a/nebo strukturních významů (5.1.2), nových významů (5.1.3), termínů (5.1.4), lexikálněsémantické skupiny názvů potravin a pokrmů (5.1.5), neologismů (5.1.6), frazémů (5.1.7), citoslovcí a synsémantik (5.1.8).

Pro vyhledávání výskytů lexikální jednotky v korpusu SYN[10] a Araneum obecně platí, že je třeba počítat s možností chybné lemmatizace (zvláště u periferních jednotek), dále je třeba automaticky nebo ručně provádět filtraci (lexikograficky) nerelevantních konkordancí. Tomu je také třeba přizpůsobit techniku vyhledávání (srov. poznámky v jednotlivých bodech).

### 5.1.1 Přechýlené názvy a zdrobněliny

Přechýlené názvy a zdrobněliny jsou do slovníku zařazovány i při poměrně nízké frekvenci, ale s dodržením podmínky lexikalizace a uzualizace. Pokud korpus Araneum potvrzuje velmi nízkou frekvenci v korpusu SYN, je tím také potvrzena nedostatečná uzualizace daného lexému. Do slovníku proto nejsou zařazena např. substantiva[11] *agorafobička* (SYN6 1×, ARIII 5×), *ariánka* (SYN6 0×, ARIII 2×), *aritmetička* (SYN6 2×, ARIII 0×), *barikádnice* (SYN6 2×, ARIII 0×), *biomedička* (SYN6 7×, ARIII 5×[12]), *dekadentka* (SYN6 2×, ARIII 2×).

---

[8] V příspěvku někdy místo termínu lexie používáme výraz význam.
[9] Frekvenční kritérium se při rozhodování o zařazení lexikální jednotky kombinuje s jinými kritérii (viz výše).
[10] V rámci dále předložených analýz je jako „korpus SYN“ označován korpus SYN ve verzi 6, jako „korpus Araneum“ korpus Araneum Bohemicum III Maximum.
[11] U všech frekvenčních údajů jsou odfiltrovány pro slovníkové zpracování nerelevantní doklady (zejména propria, resp. užití výrazu s velkým písmenem).
[12] Pouze ze 2 zdrojů, doklady nejsou relevantní.

Pokud jde o deminutiva, v případech, jako je např. *beranička* (SYN6 10×, ARIII 14×[13]), *bioobchůdek* (SYN6 37×, ARIII 61×[14]), *blafáček* (SYN6 28×, ARIII 22×[15]), *bubáček* (SYN6 2×, ARIII 4×), *bumbáníčko* (SYN6 5×, ARIII 54×[16]), potvrzuje korpus Araneum analýzu provedenou v korpusu SYN, tj. nezařazení těchto jednotek do slovníku.

Naopak zjištěný frekvenční rozdíl u substantiva *anonymka* (SYN6 2×, ARIII 198×) ukazuje, že původní rozhodnutí nezařazovat tuto jednotku do slovníku z důvodů nízké frekvence v korpusu SYN a k jeho vázanosti na internetové diskuse bude třeba případně přehodnotit na základě dalšího zkoumání.

### 5.1.2 Významy uvedené v předcházejících výkladových slovnících. Strukturní významy

Jako vhodný nástroj může sloužit korpus Araneum i pro ověřování existence a užívání daného významu v současném úzu v případě, že je zachycen v předcházejících výkladových slovnících a/nebo se jedná o význam, který lze ze strukturního (systémového) hlediska očekávat, avšak korpus SYN ho nezachycuje nebo ho zachycuje v minimální míře.

U slovesa *docpat* potvrzuje korpus Araneum lokální význam ‚nesnadně dopravit míč, puk do branky' (SYN6 32×, ARIII 15×) i význam temporální ‚dokončit naplňování' (SYN6 2×, ARIII 2×[17]). Ve slovníku je zachycen i význam ‚doplnit na určitou míru' (SYN6 11×, ARIII 45×). Vzhledem k nedostatečné uzualizaci není zařazen význam ‚s úsilím dosáhnout nějakého cíle' (SYN6 11×, ARIII 14×). Araneum dále dokládá 2 významy, které SYN nezachycuje – ‚dodatečně dodat' a ‚dát někomu jídlo, a tím ho zcela zasytit'. Tyto významy jsou vzhledem k absenci v korpusu SYN a neustálenosti, resp. vázanosti na jeden zdroj nezařazeny.

U reflexiva *docpat se* se díky ověření v korpusu Araneum do slovníku zařazuje význam ‚dodatečnou konzumací jídla se zcela zasytit' (SYN6 2×, ARIII 41×). Další 2 významy ‚doplnit na určitou míru' (SYN6 2×, ARIII 0×) a ‚dokončit konzumaci jídla' (SYN6 0×, ARIII 3×) jsou vzhledem k nízké frekvenci a neustálenosti, resp. okazionálnosti nezařazeny.

U slovesa *dočurat/dočůrat* převažuje v obou korpusech lokální význam ‚domočit do určité vzdálenosti' (SYN6 62×, ARIII 128×) nad temporálním ‚přestat močit' (SYN6 2×, ARIII 52×[18]). Ačkoli je temporální význam vzhledem k lokálnímu okrajovější, je ze systémového hlediska zařazen.

### 5.1.3 Nové významy

Při vyhledávání nových významů[19] lze korpus Araneum využívat přímo jako identifikační zdroj nebo jako zdroj potvrzující, resp. vyvracející data korpusu SYN. V korpusovém

[13] Po odfiltrování názvů samice králičí rasy.
[14] U členů otevřených slovotvorných řad je stanovena vyšší frekvenční hranice.
[15] Omezeno na specifické prostředí – mluvu sportovních komentátorů.
[16] Omezeno na malý počet zdrojů, opakující se kontexty.
[17] Ačkoli chybí dostatek dokladů, tento typ významu ze systémového hlediska zařazujeme.
[18] Jde z velké části o doklady z textů pornografického charakteru.
[19] Významy v korpusovém materiálu obecně identifikujeme buď pomocí náhodného vzorku (300 výskytů), nebo na základě zjišťování slovního okolí (s využitím nástroje WordSketch nebo nástroje kolokace a frekvenční distribuce).

materiálu se podařilo identifikovat nový význam substantiva *baldachýn* ‚posuvné zařízení k zastínění terasy, zimní zahrady, skleníku ap.‘. Frekvence kolokátů [20] v korpusu SYN potvrzujících tento význam se pohybují v jednotkách, v korpusu Araneum v desítkách [21]: *pergola* SYN6 1×, ARIII 76×; *markýza* SYN6 3×, ARIII 35×; *clona* SYN6 0×, ARIII 25×; *exteriérový* SYN6 0×, ARIII 14×; *venkovní* SYN6 0×, ARIII 14×; *sluneční* SYN6 0×, ARIII 29×. Podobná situace je u lexikalizovaného významu deminutiva *brzdička* ‚pomůcka k zafixování polohy šňůrky, gumy ap.‘: *šňůrka* SYN6 4×, ARIII 45×; *stažení* SYN6 0×, ARIII 38×; *stáhnout* SYN6 1×, ARIII 22×; *stahovací* SYN6 1×, ARIII 20×; *stahování* SYN6 0×, ARIII 12×; *guma* SYN6 0×, ARIII 28×; *gumička* SYN6 1×, ARIII 28×; *zip* SYN6 0×, ARIII 18×; *nohavice* SYN6 0×, ARIII 15×; *šňůra* SYN6 3×, ARIII 11×; *bunda* SYN6 0×, ARIII 11×; *kapsa* SYN6 1×, ARIII 11×.

Opačný případ, kdy je identifikován nový význam slova, ale pro svou okrajovost není do slovníku zařazen, můžeme demonstrovat na substantivu *bodovačka*. Slovní okolí potvrzující profesní význam ‚bodová svářečka‘ vypadá v obou korpusech následovně: *plech* SYN6 2×, ARIII 1×; *svářečka* SYN6 1×, ARIII 2×; *svářet*, *svařovat* SYN6 0×, ARIII 0×; *pneumatický* SYN6 2×, ARIII 2×; *stojanový* SYN6 0×, ARIII 1×; *elektronický* SYN6 0×, ARIII 1×.

### 5.1.4 Termíny

U terminologických hesel je nutné kromě frekvence aplikovat další kritéria k zařazení (viz bod 3). Primárně jde o kontrolu, zdali se lexikální jednotka nevyskytuje pouze nebo převážně v odborných zdrojích (odborných textech), a tedy že se šíří mimo (úzce) odbornou komunikaci. Ani relativně vysoký výskyt není důvodem k zařazení úzce odborného termínu. Vysoká frekvence v korpusu Araneum v porovnání s korpusem SYN může však být impulsem k opětovnému zvážení zařazení jednotky do hesláře. Nevýhodou korpusu Araneum však je, že vzhledem k absenci funkce třídění podle typů textů není možné provést rychlou analýzu (srov. 5.1.5). Analýza dat z korpusu Araneum potvrzuje nezařazení např. těchto výrazů: *atomizátor* (SYN6 194×, z toho publicistika 0×, malý počet zdrojů; ARIII 40×, malý počet zdrojů, a to pouze odborných), *bioskop* (SYN6 11×, ARIII 12×), *bočník* (SYN6 63×, z toho publicistika 6×; ARIII 412×, zprav. odborné zdroje z elektrotechniky), *delaminace* (SYN6 40×, z toho publicistika 4×; ARIII 156×, zprav. odborné zdroje), *berylióza* (SYN6 6×, ARIII 27×, v obou korpusech odborné medicínské zdroje), *bromová voda* (SYN6 23×, ARIII 25×, v obou korpusech v odborných zdrojích). Podrobnější vyhodnocení a případné přehodnocení nezařazení bude potřeba naopak provést např. u výrazů: *aramid* (SYN6 39×, z toho publicistika 20×; ARIII 196×, různé zdroje zaměřené nejen na techniku a chemii, ale např. i na textilnictví, oděvnictví a sport), *bootovat* (SYN6 64×, ARIII 1117×, frekvence se velmi liší, v korpusu Araneum není slovo vázáno pouze na odborné zdroje).

### 5.1.5 Lexikálněsémantická skupina: názvy potravin a pokrmů

Jednou ze specifických lexikálněsémantických skupin, které podléhají limitovanému výběru, jsou názvy potravin a pokrmů. Názvy exotických, cizích jídel a potravin zařazujeme do slovníku, pokud splňují předpoklad jisté rozšířenosti užití v našem jazykovém a kulturním

[20] V rozsahu 5 pozic vlevo a 5 pozic vpravo od KWIC.
[21] Při analýze je třeba zohlednit, že se některé zdroje opakují. Manuálně byly po zobrazení slovního okolí vyřazeny kontexty, v nichž je substantivum *baldachýn* v jiném než zkoumaném významu.

prostředí a jejich další fungování v jazyce pokládáme za perspektivní. Do uvedené lexikální skupiny náleží i názvy mexických národních pokrmů, jejichž základ tvoří placky plněné různorodými směsmi. Jde o lexémy *tortilla(s)*, *burrito(s)*, *fajita(s)*, *quesadilla(s)*, *enchilada(s)*, *taquito(s)*, *tostada(s)* a *chimichanga(s)*. Vzhledem k tomu, že se jedná o skupinu exotických pokrmů, lze tu za optimální považovat výskyt dané lexikální jednotky (po vytřídění neprůkazných nebo nevhodných dokladů s nízkou výpovědní hodnotou [22]) minimálně v 50–100 dokladech jasně ilustrujících její užití v běžných kontextech (např. *dám si / objednám si / koupím / k večeři si uvařím / připravím* apod.). Z porovnání doloženosti zkoumaných výrazů v relevantních kontextech v korpusu SYN a Araneum (viz tabulka 1) vyplývá, že do slovníku by bylo možné – vedle lexému *tortilla(s)*[23] – zařadit i názvy *burrito(s)*, *quesadilla(s)*, *fajita(s)*. Na hranici zařazení je výraz *enchilada(s)*. Vzhledem k nerovnoměrnému a v čase klesajícímu výskytu se předpokládá jeho nezařazení. Nejméně doložená (a do hesláře nezařaditelná) jsou substantiva *taquito(s)*, *tostada(s)*, *chimichanga(s)*.

| Lexém | SYN6 | ARIII |
|---|---|---|
| **tortilla/tortillas** | 2718 (2603) / 75 (70) | 3222 (3119) / 89 (85) |
| **burrito/burritos** | 238 (88) / 183 | 409 (376) / 259 |
| **quesadilla/quesadillas** | 68 / 73 | 112 / 151 |
| **fajita/ fajitas** | 29 (10) / 133 (124) | 40 (31) / 221 (116) |
| **{enchilada/enchiladas}** | 21 (13) / 86 (83) | 24 (19) / 133 (132) |
| taquito/taquitos | 2 / 4 | 0 / 6 |
| tostada/tostadas | 4 (3) / 19 | 13 (8) / 21 |
| chimichanga/chimichangas | 7 / 11 (10) | 11 (9) / 6 |

Tabulka 1.

Vysvělivky k tabulce 1: lexém tučně: velmi pravděpodobné zařazení do slovníku; lexém tučně ve složených závorkách: nejisté zařazení do slovníku; lexém netučně: nezařazení do slovníku. Čísla v závorkách uvádějí počet dokladů po odfiltrování nerelevantních dokladů a výskytů v jiných významech než pokrm (omáčka, koření).

Předběžné výsledky získané z frekvenční analýzy výskytů daného lexému je třeba potvrdit či korigovat frekvenční analýzou výskytů lexému podle typu textů, případně podle výskytů v jednotlivých letech. Tu je možno automaticky provést v korpusu SYN, nikoli však ve webovém korpusu Araneum [24]. Závěry proto vycházejí z údajů v korpusu SYN. Lexémy

[22] Faktorem zpochybňujícím oprávněnost zařazení dané lexikální jednotky do hesláře je její častý nebo převažující výskyt ve výčtech, v kuchařských receptech, obchodních nabídkách, dále výskyt s výrazem „tzv.“ nebo v uvozovkách, který svědčí o její cizosti a neusazenosti v lexikálním systému přijímajícího jazyka.

[23] Substantivum *tortilla* je zachyceno v Novém akademickém slovníku cizích slov, ostatní výrazy nikoli.

[24] Pro otestování možnosti pracovat s typy textů i v korpusu Araneum byly materiálové zdroje experimentálně rozděleny do dvou skupin. Do kategorie publicistických textů byly zařazeny odborné blogy, které se zaměřují na určitou tematickou oblast (cestování, životní styl) a bývají součástí nějakého zpravodajského serveru (idnes.cz), dále publicistické a studentské webmagazíny, lifestylové a hobby magazíny, stránky soukromých rádií a televizí věnované vaření. K

splňující frekvenční kritérium pro zařazení mají většinou dvojnásobný (resp. několikanásobný) výskyt v publicistice oproti oborové literatuře, okrajově se vyskytují v beletrii. Převažující výskyt v publicistice potvrzuje šíření jednotky mimo úzce odbornou oblast, a tedy zařazení jednotky do slovníku.

### 5.1.6 Neologismy

Pokud jde o neologismy, ilustrujeme využití korpusu Araneum na dvou příkladech. První z nich se týká otázky zařazení jednotlivých členů neologického slovotvorného hnízda a zároveň reprezentuje problematiku zařazování nových a/nebo sociálně příznakových lexikálních jednotek (slovotvorné hnízdo *boulder*), druhý se týká otázky zařazení neologického derivátu v počátcích jeho užívání (adjektivum *hotspotový*).

Výrazy ze slovotvorného hnízda *boulder*[25] jsou doloženy jak v korpusu SYN, tak v korpusu Araneum. Do ASSČ byly zařazeny lexémy: *boulder*[26] jako substantivum i jako nesklonné adjektivum (SYN6 325×, ARIII 4739×), *bouldering* (SYN6 1010×, ARIII 2798×), *boulderingový* (SYN6 233×, ARIII 404×) a *boulderový* (SYN6 189×, ARIII 704×). Zařazeny nebyly jednotky, které jsou nové, resp. sociálně vázané a které teprve zvolna opouštějí periferii slovní zásoby: *boulderovka* (SYN6 19×, ARIII 220×), *boulderista* (SYN6 25×, ARIII 256×), *boulderistka* (SYN6 2×, ARIII 5×), *boulderovací* (SYN6 5×, ARIII 39×), *boulderovat* (SYN6 14×, ARIII 143×), *boulderování* (SYN6 41×, ARIII 128×) a výraz složený: *bouldermatka* (SYN6 32×, ARIII 474×).[27]

Při adaptačním procesu výpůjček se mění ve formální podobě některých výrazů skupina *-der-* na *-dr-* (kolísá i výslovnost). V textech obsažených v korpusu SYN se dodržuje původní podoba slov rigidněji než v textech obsažených v korpusu Araneum, srov. *bouldr* (SYN6 1×, ARIII 443×), *bouldrový* (SYN6 1×, ARIII 50×), *bouldrovací* (SYN6 10×, ARIII 62×). K uvedené formální změně *-der-* na *-dr-* ale podle korpusového materiálu vůbec nedochází u výrazů *bouldering* a *boulderingový* – hláska *e* je v české výslovnosti patrně důsledně realizována. Vzhledem k menší míře korektorských zásahů ve zdrojích webového korpusu Araneum lze data v tomto korpusu chápat jako určitý indikátor adaptačních změn (adaptované podoby mohou v jazyce koexistovat s neadaptovanými). Adaptované podoby *bouldr*, *bouldrový* (zatím) nejsou v ASSČ podány jako varianty hesel v záhlaví slovníkových odstavců, a to kvůli jejich nízké doloženosti v korpusu SYN.

---

„ostatním“ analyzovaným zdrojům byly přiřazeny firemní weby, osobní kuchařské a cestovatelské weby a blogy, weby o zdravé výživě, internetové kuchařky, průvodce restauračními podniky aj. Toto rozdělení bylo ověřeno na výrazu *tortilla*. Ukázalo se při tom, že ve skupině „ostatní texty“ jsou i některé zdroje, které by bylo možné považovat za publicistické. V analýze typů textů v korpusu Araneum by bylo nutné dále pokračovat. Již nyní je třeba konstatovat, že takto ručně prováděná analýza typů textů v korpusu Araneum je velmi pracná a časově náročná.

[25] Frekvenční údaje u výrazů *boulder* a *bouldering* jsou uvedeny bez výskytů s velkým písmenem.

[26] *Boulder* je v rukopisu ASSČ definován jako ‚nízká umělá lezecká stěna, na kterou se leze bez zajištění lanem‘. Výraz zaznamenávají rovněž Nový akademický slovník cizích slov a Slovník současné češtiny.

[27] Stranou ponecháváme málo frekventované výrazy počínající dalšími písmeny abecedy: *oboulderování*, *oboulderovat*, *proboulderovat*, *přeboulderovat*, *streetbouldering*, *streetboulderový*, *uboulderovat*, *vyboulderovat* a *zaboulderovat* (tyto výrazy převážně slangové povahy jsou doloženy pouze v korpusu Araneum, nikoli v korpusu SYN).

Adjektivum *hotspotový*[28] není dosud příliš frekventované (SYN6 7× ve 2 různých zdrojích, ARIII 20×[29] v 18 různých zdrojích). Na 1. kolokační pozici vpravo od KWIC se v korpusu Araneum vyskytují výrazy *síť*, *služby*, *systém*, *řešení*; v korpusu SYN jsou oproti korpusu Araneum navíc výrazy *brána*, *funkce*, *režim*. Webový korpus tedy v případě adjektiva *hotspotový* nabízí větší množství příkladů z více zdrojů, ale s menší variabilitou výrazů na 1. pozici vpravo od KWIC. Doloženost adjektiva je pro zařazení do slovníku dostatečná, omezený počet – navíc pouze specializovaných – zdrojů v korpusu SYN však svědčí v neprospěch zařazení hesla. O definitivním zařazení jednotky do slovníku bude rozhodnuto při závěrečné redakci hesel.

### 5.1.7 Frazémy

Vzhledem k tomu, že korpus Araneum obsahuje množství textů, které jsou příznačné pro běžnou komunikaci, je tento korpus vhodným ověřovacím zdrojem i pro zjištění, zda do slovníku určitý frazém zařadit, nebo nezařadit. Tak např. korpus Araneum potvrzuje, že přirovnání obsahující spojení *stará bába* (*naříkat*, *být zvědavý*, *být pověrčivý jako stará bába*) náleží k periferním jednotkám a do slovníku nebudou zařazena. Podobně se v něm ukazuje, že neologický frazém z konce 90. let *bába Dymáková* je stále okrajový (srov. SYN6 58× v 16 zdrojích, vázáno především na roky 1998, 1999, sporadické výskyty i v dalších letech, ARIII 25× v 17 zdrojích[30]), a nebude proto do slovníku zařazen. V případě parémie *komu není shůry dáno, v apatyce nekoupí* korpus Araneum potvrzuje zařazení do slovníku (srov. SYN 58× v 28 zdrojích, ARIII 130× v 88 zdrojích). Frekvence frazému *rektální alpinismus* v korpusu Araneum nejen vede k zařazení této lexikální jednotky do slovníku, ale také naznačuje, že by bylo potřebné zvážit zachycení i její varianty *řitní alpinismus* (srov. *rektální alpinismus*: SYN6 43× ve 20 zdrojích, ARIII 184× v 119 zdrojích, *řitní alpinismus* SYN6 10× v 6 zdrojích, ARIII 61× v 52 zdrojích). U frazému *(ne)dělat zagorku/Zagorku*, u něhož je frekvence v SYN6 49× v 21 zdrojích, korpus Araneum poměrně přesvědčivě ukazuje na jeho rozšíření v úzu, a je tedy důvodem pro jeho zařazení do slovníku (ARIII 107× v 96 zdrojích).

### 5.1.8 Citoslovce a synsémantika

Specifickou roli hraje webový korpus Araneum u těch výrazů, které jsou typické pro mluvený jazyk, vyjadřují emoce či postoje mluvčích, často jsou expresivními variantami výrazů používaných v neutrálním jazykovém projevu. Jedná se např. o varianty citoslovcí *fujky*, *čauky*, *ahojky/ahojda*. V korpusu SYN se tyto výrazy objevují také, zpravidla jsou však méně frekventované (*fujky* SYN6 5×, ARIII 271×; *čauky* SYN6 82×, ARIII 5274×; *ahojky* SYN6 207×, ARIII 146 710×; *ahojda* SYN6 14×, ARIII 5064×). V korpusu SYN jde zprav. o zprostředkované užití, např. z beletrie, reprodukovaných rozhovorů; v korpusu Araneum jde častěji o přímé použití v původní komunikační funkci. Lze tedy hovořit o doložení

[28] Adjektivum *hotspotový* je derivátem výrazu anglického původu *hotspot* 1. ‚veřejný přístupový bod pro bezdrátové připojení k Internetu', 2. ‚záchytné centrum pro uprchlíky'.
[29] Výraz je chybně lemmatizován jak ve webovém korpusu Araneum, tak v korpusu SYN. Tomu bylo potřeba přizpůsobit vyhledávací strategii.
[30] Pro plné potvrzení uvedeného by ještě bylo potřeba mít k dispozici dataci jednotlivých konkordancí. To však korpus Araneum neumožňuje.

neformálního psaného jazyka.[31] Vlastní užití výrazů se však příliš neliší.

Korpus Araneum může sloužit i pro ověření posunu v užívání některých synsémantik. Na základě frekvence byly do hesláře ASSČ zařazeny spojky, které byly v předcházejících výkladových slovnících označeny jako zastaralé, např. *anobrž*, *alébrž*. Podle dokladů v korpusu SYN se jimi uživatelé často snaží ozvláštnit text. V některých případech však uživatelé tyto spojky používají v odlišném významu, než který je uváděn ve slovnících. Webový korpus Araneum u těchto spojek potvrzuje jak tendenci ke zvýšení užívání těchto spojek (*anobrž* SYN6 99×, ARIII 281×; *alébrž* SYN6 73×, ARIII 622×), tak určité rozvolnění povědomí o jejich významu, srov. příklad, který dokládá použití spojky *anobrž* v jiném než původním odporovacím významu (ale, nýbrž), a to ve významu příčinném (protože): *že jsem vlastně ten nejšťastnější člověk, anobrž nemám í-kvé tykve a nechytám včerejší den do síťky na motýly*. Ve slovníku jsou proto zachyceny oba významy.

## 5.2 Gramatická charakteristika hesel v ASSČ

Pokud jde o zjišťování gramatických informací, při soustavné lexikografické práci se ukazuje, že korpus Araneum, zejména díky specifickému charakteru jazykového materiálu, umožňuje ověřit existenci některých gramatických forem v reálném úzu. Jedná se zejména o:

a) stupňování kvalifikačních adjektiv, resp. adjektiv, která nabyla kvalifikační význam, a stupňování paralelně derivovaných příslovcí. Tak např. u adjektiv *bezporuchový*, *bezpracný*, *blbuvzdorný* a adverbií *bezporuchově*, *bezpracně*, *blbuvzdorně* není stupňování podchyceno v žádném z dostupných slovníků, včetně speciálních.[32] V komunikaci se uvedená adjektiva a adverbia užívají v kvalifikačním významu, o čemž svědčí také doklady stupňovaných tvarů. V korpusu SYN jsou sice jednotkové nebo zcela žádné výskyty, v korpusu Araneum však frekvence (někdy významně) narůstá: *bezporuchovější* 4×, *bezporuchověji* 2×; *bezpracnější* 4×, *nej/bezpracněji* 4×; *nej/blbuvzdornější* 68×, *blbuvzdorněji* 5×.[33] Nejčastěji platí, že adjektivní stupňování je dostatečně doloženo jak v korpusu SYN, tak Araneum (*nej/festovnější*: SYN6 34×, ARIII 61×), avšak potencialita stupňování derivovaného příslovce je potvrzena až v korpuse Araneum (*festovněji* ARIII 6×);

b) užívání řidších morfologických forem u kolokvialismů, resp. slangismů. Expresivní kolokviální výraz *bengo* je relativně významně doložen jak v korpusu SYN (cca 180×), tak v korpusu Araneum, kde je však frekvence několikanásobně vyšší (cca 630×).[34] Několikanásobná frekvence v korpusu Araneum oproti korpusu SYN se ukazuje u tvaru 1. pl.,

---

[31] O rysech psaného a mluveného jazyka srov. např. Nový encyklopedický slovník češtiny (Karlík, P, Nekula M. & Pleskalová J. (2016); autorkou hesla Projevy mluvené a psané je J. Hoffmannová).
[32] Tento stav může mít různé důvody: a) tento údaj nebyl soustavně podchycován, např. v Slovníku spisovného jazyka českého byl uváděn jen u nepravidelně tvořených forem, b) zkoumaná adjektiva nebyla do hesláře dané příručky zahrnuta, např. adjektivum *blbuvzdorný* není registrováno v Slovníku spisovného jazyka českého, Slovníku spisovné češtiny pro školu a veřejnost a v Internetové jazykové příručce), c) při zpracovávání příslušné příručky nebyl dostatek jazykového materiálu, který by užívání stupňovaných forem prokazoval (srov. *bezporuchový*, *bezpracný* v Internetové jazykové příručce).
[33] Tvary se někdy dále ověřují přímo na internetu, kde jsou pak výskyty zprav. ještě značně vyšší.
[34] Frekvenční údaje jsou po odfiltrování výskytů s počátečním velkým písmenem, výskyty psané verzálkami byly naopak ponechány (s výjimkou odfiltrování kolokací *AC BENGA*, *BENGA ČAVE*), dále byly vyřazeny (ojedinělé) citace romštiny, resp. vězeňského slangu (*Ara bengo!*).

resp. 4. pl. *benga* (SYN6 146×, ARIII 451×).[35] V případě všech dalších morfologických tvarů a jejich forem se ukazuje větší rozmanitost v materiálu korpusu Araneum, a to jak kvantitativně (vyšší frekvence forem shodně doložených v obou korpusech), tak kvalitativně (v korpusu Araneum jsou doloženy formy neregistrované v korpusu SYN). Srov. 1., 4. sg. *bengo*[36] (SYN6 18×, ARIII 72×), další formy pro 1. pl. *bengové/bengy* (SYN6 0×, ARIII 3×/5×), 2. pl. *bengů/beng* (SYN6 2×/0×, ARIII 11×/2×)[37]. V korpusu Araneum je doložen také další, systémově náležitý tvar 4. pl. *bengy*, ale v problematických kontextech. Morfologická informace v hesle *bengo* je tedy na základě ověření v jazykovém materiálu tato: 2. j. -ga, 3., 6. j. -govi, -gu, 4. j. -ga, -go, 1. mn. -ga, -gové, 2. mn. -gů, -g, 3. mn. -gům, 4. mn. -ga, 6. mn. -gách, 7. mn. -gy m. živ. i s.;

c) užívání morfologických forem příznačných pro běžně mluvený jazyk. Silně expresivní až zhrubělé výrazy *blít* a *fuckovat/fakovat* patří k slovesům 3. třídy, u nichž je v 1. os. sg. a 3. os. pl. možno užít dva stylově rozlišené soubory koncovek: neutrální *-i* a *-í* vedle příznakových *-u* a *-ou*. V případě příznakových výrazů se však jeví jako velmi problematické (byť formálně možné) uvádět ve slovníkové morfologické informaci tvary v naznačeném pořadí (tj. nejdříve neutrální, avšak ke knižnosti směřující koncovky *-i*, resp. *-í*, pak teprve *-u*, *-ou*). Přestože toto pořadí by se jevilo jako náležité podle jazykového materiálu v korpusu SYN, korpus Araneum potvrzuje jazykově přirozené přiřazování koncovek příznačných pro mluvený jazyk k příznakovým slovním základům, srov. *bliji* SYN6 0×, ARIII 11×, *bliju* SYN6 26×, ARIII 158×, *blijí* SYN6 12×, ARIII 63×, *blijou* SYN6 9×, ARIII 88×; *fuckuji/fakuji* SYN6 6×/0×, ARIII 1×/3×, *fuckuju/fakuju* SYN6 0×/11×, ARIII 14×/17×, *fuckují/fakují* SYN6 6×/7×, ARIII 5×/7×, *fuckujou/fakujou* SYN6 1×/1×, ARIII 19×/21×.[38] Frekvenční údaje z korpusu Araneum svědčí o tom, že by bylo v morfologické informaci u těchto sloves náležité uvádět tvary na *-u* a *-ou na prvním místě*.

## 6 Závěrem

V příspěvku jsme demonstrovali, že díky nástrojům webového korpusu Araneum je oproti vyhledávání na internetu možné rychleji a snáze získat odpovědi na otázky v případě, že korpus SYN poskytuje pro lexikografická rozhodnutí málo dat nebo neposkytuje data žádná. Zároveň je však potřeba zdůraznit, že webový korpus Araneum není vhodné pro ASSČ využívat jako základní materiálový zdroj (jak pro tvorbu hesláře, tak pro zpracování jednotlivých slovníkových hesel), ale pouze jako podpůrný, doplňkový zdroj. Je totiž nutné počítat s jeho nevyvážeností, pokud jde o typy textů, a také nemožností zjistit jejich dataci[39], včetně časové distribuce.

---

[35] V obou korpusech je velmi nedokonalé tagování formy *benga* z hlediska tvarové homonymie, např. v korpusu SYN je 43 výskytů označeno jako genitiv sg., ale jen 3× jde o správný tag. V korpusu Araneum není (z mnohem většího množství) ani jeden výskyt označen jako genitiv sg., manuálně jsou v něm však tyto tvary (s minimální frekvencí) identifikovatelné.

[36] Zároveň výskyty slova v těchto tvarech v příhodném kontextu umožňují doložit rodové kolísání mezi životným maskulinem a neutrem.

[37] Forma *beng* je v korpusu Araneum doložena mimo lemmatizované tvary.

[38] Frekvence jsou uvedeny včetně negativních forem, tj. *nebliji*, *nebliju* ap., a včetně psaní verzálkami.

[39] Je možné zjistit jen datum stažení.

K výhodám korpusu Araneum náleží bezpochyby jeho objemnost. Pro texty, které obsahuje, je příznačná nespisovnost, expresivita, osobitá aktivita mluvčího a jeho vztah k danému prostředí; důležitým komunikačním cílem je vedle předávání informací rovněž navazování a rozvíjení kontaktu s dalšími komunikanty[40]. Korpus Araneum je tak vůči korpusu SYN (s převahou publicistických textů) komplementárním zdrojem v tom smyslu, že obsahuje pestré žánry elektronické komunikace (např. internetové články, interview, diskuse, různorodé marketingové texty v e-shopech ad.) a rovněž řadu textů náležejících do sféry běžné (každodenní) komunikace, tj. projevů neformálních, nepřipravených. Informace o typech textů si však musí badatel (zdlouhavě, výběrově) zjišťovat sám. Do budoucna by proto bylo vítané inkorporovat nástroj pro automatizované třídění typů textů i do korpusu Araneum. Přes naznačené nevýhody je třeba konstatovat, že korpus Araneum se díky korpusovým nástrojům, s jejichž pomocí je možné analyzovat obrovské množství dat, stal důležitým a osvědčeným materiálovým zdrojem pro lexikografické analýzy při tvorbě všeobecného výkladového slovníku.

## 7 Literatura

Benešová, L., Křen, M. & Waclawičová, M. *ORAL2013: reprezentativní korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha 2013. Dostupné z: http://www.korpus.cz [30/06/2018].

Benko, V. *Araneum Bohemicum Maius, verze 15.04*. Ústav Českého národního korpusu FF UK, Praha 2015. Dostupné z: http://www.korpus.cz [30/06/2018].

Benko, V. *Araneum Bohemicum Maximum, verze 15.04*. Ústav Českého národního korpusu FF UK, Praha 2015. Dostupné z: http://www.korpus.cz [30/06/2018].

Benko, V. *Araneum Bohemicum III Maximum (Czech, 17.04)*. UNESCO Katedra plurilingválnej a multikultúrnej komunikácie. Univerzita Komenského v Bratislave. Dostupné z: http://unesco.uniba.sk [30/06/2018].

Benko, V. *Araneum Bohemicum Minus, verze 15.04*. Ústav Českého národního korpusu FF UK, Praha 2015. Dostupné z: http://www.korpus.cz [30/06/2018].

Benko, V. (2014a). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček, K. Pala (eds.) *Text, speech, and dialogue: 17th international conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014: proceedings*. Cham: Springer, s. 257–264.

Benko, V. (2014b). Compatible Sketch Grammars for Comparable Corpora. In A. Abel, Ch. Vettori, N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User In Focus, 15–19 July 2014*. Bolzano/Bozen: Eurac Research, s. 417–430.

Benko, V. (2014c). Je webový korpus „horší“? In *Korpusová lingvistika Praha 2014. 20 let mapování češtiny*. Abstrakty. IV. pražská konference korpusové lingvistiky pořádaná u příležitosti 20. výročí založení ČNK. Univerzita Karlova, Praha, s. 21–23.

Čermák, F. (2017). *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum.

Čermák, F., Blatná, R., Hlaváčová, J., Klímová, J., Kocek, J., Kopřivová, M., Křen, M., Petkevič, V., Schmiedtová, V. & Šulc, M. *SYN2000: žánrově vyvážený korpus psané češtiny*.

[40] Srov. Jílková (2016: 108).

Ústav Českého národního korpusu FF UK, Praha 2000. Dostupné z: http://www.korpus.cz [30/06/2018].

Čermák, F., Doležalová-Spoustová, D., Hlaváčová, J., Hnátková, M., Jelínek, T., Kocek, J., Kopřivová, M., Křen, M., Novotná, R., Petkevič, V., Schmiedtová, V., Skoumalová, H., Šulc, M. & Velíšek, Z. *SYN2005: žánrově vyvážený korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha 2005. Dostupné z: http://www.korpus.cz [30/06/2018].

[Grouws et al. (2013)] Grouws, R. H., Heid, U., Schweickard, W. & Wiegand, H. E. (2013). Dictionaries: an international encyclopedia of lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin: De Gruyter Mouton.

Hoffmannová, J., Homoláč, J., Chvalovská, E., Jílková, L., Kaderka, P., Mareš, P., Mrázková, K. (2016). *Stylistika mluvené a psané češtiny*. Praha: Academia.

*Internetová jazyková příručka. Praha: Ústav pro jazyk český AV ČR. Dostupné z: http://prirucka.ujc.cas.cz [30/06/2018].*

[Jakubíček et al. (2013)] Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*. Lancaster, s. 125–127.

Jílková, L. (2016). Elektronická komunikace. In Hoffmannová, J., Homoláč J., Chvalovská, E., Jílková L., Kaderka, P. & Mareš, P. *Stylistika mluvené a psané češtiny*. Praha: Academia, s. 105–143.

Karlík, P., Nekula, M. & Pleskalová, J. (2016). *Nový encyklopedický slovník češtiny on-line*. Dostupné z: https://www.czechency.org/index.html [30/06 2018].

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

Kopřivová, M., Waclawičová, M.: *ORAL2006: korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha 2006. Dostupné z: http://www.korpus.cz [30/06/2018].

Kosem, I. (2016). Interrogating a Corpus. In Durkin, P. (eds.) *The Oxford handbook of lexicography*. New York: Oxford University Press, s. 76–93.

Křen, M., Bartoň, T., Cvrček, V., Hnátková, M., Jelínek, T., Kocek, J., Novotná, R., Petkevič, V., Procházka, P., Schmiedtová, V. & Skoumalová, H. *SYN2010: žánrově vyvážený korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha 2010. Dostupné z: http://www.korpus.cz [30/06/2018].

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P. & Zasina, A.: *Korpus SYN, verze 6 z 18. 12. 2017*. Ústav Českého národního korpusu FF UK, Praha 2017. Dostupné z: http://www.korpus.cz [30/06/2018].

*Mediasearch – mediální archiv*. Praha: NEWTON Media, a. s. Dostupné z: http://mediasearch.newtonit.cz [30/06/2018].

Nový akademický slovník cizích slov (2005). Praha: Academia.

*Příruční slovník jazyka českého* (1935–1957). Praha: Státní nakladatelství, Školní nakladatelství, Státní pedagogické nakladatelství.

*Slovník současné češtiny* (2017). Internetový slovník současné češtiny, verze 2.0. Praha: Lingea s.r.o. Dostupné z: https://www.nechybujte.cz/slovnik-soucasne-cestiny

[30/06/2018].

*Slovník spisovné češtiny pro školu a veřejnost* (2003). Praha: Academia. (1. vyd. 1978; 2., opr. a dopl. vyd. 1994; 3., opr. vyd. 2003).

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

Suchomel, V. (2012). Recent Czech Web Corpora. In P. Rychlý, A. Horák (eds.) *Proceedings of the Sixth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2012*. Brno: Tribun EU, s. 77–83.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: Philadelphia: Benjamins.

Waclawičová, M., Kopřivová, M., Křen, M. & Válková, L. *ORAL2008: sociolingvisticky vyvážený korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha 2008. Dostupné z: http://www.korpus.cz> [30/06/2018].