

Tematická reprezentativnost korpusů

Reprezentativnost korpusů je idea obecně velmi obtížně podchytitelná a s určitými přestávkami již po dlouhou dobu diskutovaná (srovej zde stať J. Králíka, s. xx-xx).

První otázkou, kterou by bylo možno si položit, je otázka, do jaké míry korpus odpovídá (nebo v budoucnu bude odpovídat) jazyku - jeho produkci i recepci - z hlediska média přenosu, tedy poměru mezi jazykem mluveným a slyšeným versus psaným a čteným. Současné zkušenosti s vytvářením mluvených korpusů, s jejich zpracováním a především s jejich velkou časovou a finanční náročností ukazují, že texty mluvené a psané budou vzhledem k poměru, ve kterém se vyskytují v běžném životě (odhady hovoří o 90-95 % mluveného jazyka a pouhých 10, ale možná spíš jen 5 % jazyka psaného), pravděpodobně vždy nevyvážené a že korpusy, ve kterých by tento poměr odpovídal realitě života, budou alespoň v dohledné době několika desítek let příliš malé na to, aby poskytly dostatek materiálu pro základní výzkum na všech jazykových rovinách. Jistě budou pomocníkem pro rozlišení frekvence užití v oblasti mluveného a psaného jazyka, jistě budou velmi cenným zdrojem pro specializovaná studia mluveného jazyka, např. morfologická či syntaktická, která jsou velmi potřebná, pro lexikografická studia však podobné korpusy nebudou ještě velmi dlouhou dobu dostatečně veliké.

Je tedy třeba počítat s tím, že nejméně několik následujících desetiletí budou všechny větší korpusy odrážet především jazyk psaný. (Nejde o nic nového, i náš starší typ materiálové základny pro lexikografii - lístkový katalog - zpracovával téměř výhradně češtinu psaných textů, tento fakt je jen třeba při různých studiích brát v úvahu.) Také projekt Český národní korpus (ČNK) obsahuje převážně texty (původem) psané.

Zcela samozřejmě se otázka reprezentativnosti hned ve svém dalším kroku rozpadá na problém objemu a obsahu. Jak velký musí být korpus, aby ho bylo možno vzhledem k cílovému užití nazvat užitečně reprezentativním? Jak zajistit, aby ve velkých všeobecných korpusech byla obsažena co nejširší tematická škála?

Velikost korpusů vycházela v jejich úplných začátcích spíše z lidských, technických a finančních možností prvních korpusových lingvistů a pohybovala se od několika set tisíc k přibližně dvěma milionům slovních výskytů. Byla na nich sice podniknuta zajímavá zkou-

mání morfologická i jiná, velmi brzy se však ukázalo, že pro základní výzkum, který by se neomezoval na nižší jazykové roviny, je třeba velikosti nepoměrně větší. V dnešní době je za (pravděpodobně dočasný) standard považována velikost 100 milionů slovních výskytů, kterou má snad nejcitovanější korpus 90. let, British National Corpus (BNC). Tuto velikost nabízí i první varianta reprezentativního korpusu současné psané češtiny SYN2000, zpřístupněná veřejnosti v roce 2000.

Na obsahové měřítko, tedy na zachycení široké škály témat jedni při budování korpusů rezignují (např. Bank of English, Frantext), zatímco druzí se snažili a snaží o jistou proporcionalitu a reprezentativnost.

Pracovníci jednomilionového Survey of English Usage (SEU), nejvýznamnějšího předchůdce dnešních elektronických korpusů, dělili psaný text do následujících kategorií:

- | |
|--|
| <p>A) text tištěný</p> <ul style="list-style-type: none">1 informativní<ul style="list-style-type: none">1.1 denní tisk, 1.2 vědecké texty, 1.3 administrativa, 1.4 právo2 výukový3 agitační4 imaginativní |
| <p>B) text netištěný</p> <ul style="list-style-type: none">1 korespondence<ul style="list-style-type: none">1.1 soukromá, 1.2 nesoukromá2 deníky3 souvislý text<ul style="list-style-type: none">3.1 imaginativní, 3.2 informativní. |
| <p>C) text psaný, určený k přednesu</p> <ul style="list-style-type: none">1 různé mluvené projevy2 hry3 zprávy4 slavnostní projevy5 příběhy |

Autoři prvního elektronického korpusu, dodnes vlivného jednomilionového Brown University Standard Corpus of Present-Day Edited American English (Brown Corpus), kategorizovali své texty následovně:

| |
|---|
| <p>informativní</p> <ul style="list-style-type: none">1 denní tisk—reportáže |
|---|

- 1.1 politické, 1.2 sportovní, 1.3 společenské, 1.4 krátké zprávy, 1.5 ekonomické, 1.6 kulturní
- 2 denní tisk—úvodníky
 - 2.1 institucionální, 2.2 osobní
- 3 denní tisk—recenze (knihy, divadlo, tanec, hudba)
- 4 náboženství
- 5 řemesla a koníčky
- 6 lidové tradice
- 7 krásná literatura, životopisy, paměti
- 8 odborné texty
 - 8.1 přírodní vědy, 8.2 medicína, 8.3 matematika, 8.4 sociologie a psychologie, 8.5 politika, právo a výchova, 8.6 společenské vědy, 8.7 technika a strojírenství
- 9 různé
 - 9.1 vláda, 9.2 nadace, 9.3 zprávy průmyslových podniků, 9.4 univerzitní materiály, 9.5 zprávy podnikových orgánů

imaginativní

- 10 romány a povídky
- 11 detektivní romány a příběhy s tajemstvím
- 12 sci-fi
- 13 dobrodružné romány a westerny
- 14 milostné romány a romance
- 15 humor

Jeden z nejcitovanějších korpusů poslední doby a v letech 1991 až 1995 nejprestižnější projekt anglické lingvistiky, British National Corpus (BNC), pracuje s velkými a tematicky širokými oblastmi. Základní dělení na literaturu imaginativní a informativní sice nemohl ponechat jako nejjemnější možnou distinkci, o mnoho více však ze svého projektu nezveřejnil. Literaturu informativní dělí na osm specifitějších „domén“ (*domain*): 1. umění - 2. víra a myšlení - 3. obchod a finance - 4. volný čas - 5. čistá věda a přírodní vědy - 6. aplikované vědy - 7. sociální vědy a ekonomie - 8. dění ve světě.

Jinou možností je vytvořit jemné síto mnoha vědních oborů. Tato varianta pohledu na tematické oblasti byla zvolena v Ústavu Českého národního korpusu (ÚČNK), byť si jeho pracovníci byli vědomi obtíží a problémů z takového eventuálně zvoleného přístupu vyplývajících. I dnes se však zdá, že pracovat jen s pěti nebo deseti kategoriemi by znamenalo zůstat v půli cesty, a dnešní obtíže s kategorizací (u děl multitematických či vysloveně mezioborových) budou v budoucnu alespoň částečně vyváženy možností pracovat s vědními oblastmi odděleně (např. geografie, meteorologie, stavebnictví).

V ÚČNK je pro zajištění vyváženosti korpusu SYN2000 (stomilionový reprezentativní korpus synchronní psané češtiny) používána relativně velmi jemná tematická škála. Vznikala na podkladě materiálu, který se zabýval různými charakteristikami a jejich kombinovatelností

(F. Čermák, interní materiál ÚČNK), dále Deweyova desetinného třídění pro knihovny, poučila se na zahraničních korpusových projektech a vzala v úvahu také praktické zkušenosti těch pracovníků ÚČNK, kteří se zmiňovanou škálou pracují při značkování textů.

V současnosti je seznam velkých oblastí a jejich dalšího vnitřního dělení už relativně ustálený.

Abych tematické kategorie ČNK představil v přehledné podobě, rozhodl jsem se uveřejnit je v tabulce, která bude umožňovat srovnání s Deweyovým desetinným tříděním pro knihovny a s tematickými oblastmi, ve kterých Oxford University Press (Oxford UP) využívá pro svůj lexikografický program excerpe externích odborníků. Tyto konfrontační soubory témat můžeme považovat jednak za soubory obecným užíváním vyzkoušené (zejména v případě Deweyova třídění), jednak za zkušenými lexikografy zvolené a stále prověřované kategorie (zejména v případě Oxford UP), které navíc zajímavě poukazují na historicky podmíněné oblasti čtenářského zájmu (srov. nepoměr mezi množstvím jemných kategorií např. v oblasti filosofie, medicína, domácí hospodářství, sport či minority).

Zkratky používané při vnější lingvistické anotaci (viz sloupec 1) jsou založeny na mezinárodních názvech oblastí, oborů.

| Značka v ČNK | ČNK | Dewey | Oxford UP |
|---------------------|--|---|-------------------------------|
| <i>typ textu:</i> | | | |
| ver | báseň | | |
| son | píseň | | písňové texty |
| scr | dramatický text, scénář | 792 divadlo | drama, divadelní písňový text |
| nov | román či jiný celek | | fikce, historický román |
| col | soubor povídek, jednotlivá povídka | 080 sborníky | |
| fac | lit. faktu | | |
| pub | publicistika (noviny a neoborné časopisy) | | |
| adm | administrativa | 350 veřejná správa, 351 ústřední vláda, 352 místní správa, 651 řízení administrativy, 652 psaní, 653 těsnopis | zaměstnání |
| sci | vědeckonaučná lit. | | |
| pop | populárně-naučná lit. | | |
| txb | učebnice | | |
| enc | abecedně, systematicky a jinak uspořádaná díla | 030 všeobecné encyklopedie | |
| mis | rozmanité | 000 všeobecná díla | |
| <i>žánr:</i> | | | |
| IMAGINATIVNÍ | | | |
| crm | detektivní, špionážní romány | | zločin |
| scf | vědecko-fantastická lit., fantasy | | sci-fi, fantasy |
| jun | lit. pro děti a mládež, báje, pověsti, legendy, bajky, pohádky | | beletrie pro děti, mytologie |
| FAC | lit. faktu | | |
| tra | cestopisy | | |
| mem | (auto)biografie, vzpomínky, deníky | 920 životopisy (kromě 929 genealogie a heraldika) | autobiografie, biografie |

| | | | |
|---------------------|--|---|--|
| chr | kroniky, letopisy, ročenky, deníky | | deníky |
| let | dopisy | | |
| INFORMATIVNÍ | | | |
| ARS | UMĚNÍ | 700 umění, 709 dějiny umění | umění, kritická teorie |
| mus | hudba | 780 hudba | hudba (blues, etnická hudba, hudební nástroje, jazz, nahrávání, opera, písňové texty, populární hudba, primitivní hudba, rock, vážná hudba), /umění/ múzická umění |
| cin | film | | film |
| tvf | televize | | |
| arc | architektura | 710 tvorba krajiny, urbanistika, 720 architektura | architektura |
| art | výtvarné umění, užité umění | 708 galerie, muzea, sbírky, 730 sochařství, 740 kreslení a dekorativní umění, 750 malířství, 760 tiskové techniky, 770 fotografie, 913 starožitnosti | /umění/ malířství, fotografie, klenotnictví, starožitnosti, kaligrafie, keramika, design, tisk (litografie) |
| the | divadlo, balet | | tanec (balet), /umění/ múzická umění |
| lit | literární věda | 800 literatura, 810 americká literatura, 820 anglická literatura, 830 německá literatura, 840 francouzská literatura, 850 italská literatura, 860 španělská literatura, 870 latinská literatura, 880 řecká literatura, 890 ostatní literatury | literatura, literární kritika, literární teorie |
| HUM | SOCIÁLNÍ VĚDY | | |
| his | dějiny, archeologie, odborné biografie | 509 dějiny a dílčí pojednání (o čisté vědě), 571 prehistorická archeologie, 900 dějiny, 930 dějiny starověku, 929 genealogie a heraldika | dějiny, sociální dějiny, genealogie, heraldika, dějiny starověku, archeologie, mince |
| psy | psychologie | 130 psychologické obory, 150 obecná psychologie | psychoanalýza |
| edu | pedagogika a osvěta | 370 výchova, 507 studium a výuka (čisté vědy), 707 studium a výuka (umění) | výchova a vzdělání, muzea, děti, rodičovství |
| soc | sociologie, komunikace, média | 070 novinářství, 300 společenské (sociální) vědy, 360 sociální péče, 366 sdružení, 367 společenské kluby, 368 pojištění | sociologie, soc. otázky, žurnalistika, zpravodajství |
| phi | filosofie, etika | 100 filosofie, 110 metafyzika, 120 metafyzické teorie, 140 filosofická témata, 170 etika, | filosofie (fenomenologie) |

| | | | |
|------------|---|---|--|
| | | 180 starověká a středověká filosofie, 190 moderní filosofie, 501 filosofie a teorie (čisté vědy), 577 filosofie biologie, 601 filosofie a teorie (techniky), 701 filosofická hlediska (umění) | |
| inf | informace a knihovnictví | 370 výchova | výchova a vzdělání, muzea |
| pol | politologie | 320 politologie, 353 vláda Spojených států | politika, vláda |
| lin | lingvistika | 400 jazyk, 410 srovnávací jazykověda, 420 angličtina, 430 němčina, 440 francouzština, 450 italština, 460 španělština, 470 latina, 480 klasická řečtina, 490 ostatní jazyky, 508 sbírky, dialektologie | lingvistika (dialekt, gramatika, fonetika), hebrejština, angličtina ve světě, indická angličtina, jazyk, lexikografie, angličtina černochů, klasická filologie |
| eth | etnografie | 390 národopis a folklor | původní Američané |
| LAW | PRÁVO A BEZPEČNOST | | |
| jur | právo, kriminalistika | 340 právo | právo (zločin (mafie), kriminalistika, policie, vězení), dědictví, |
| mil | vojenství | 355 válečnictví, 356 pěší vojska, 357 jezdeckvo, 358 ostatní vojska a služby, 359 námořní síly | branná moc (vojenské letectvo, pozemní vojsko, válečné námořnictvo), zbraně |
| sec | bezpečnost | | špionáž |
| NAT | PŘÍRODNÍ VĚDY | 500 čistá věda | věda |
| arg | zemědělství, lesnictví, chov, pěstování | 630 zemědělství, 637 mlékárenský průmysl, 664 potravinářská technologie | zemědělství, farmaření, rybářský průmysl, lesní hospodářství, včelařství, koně |
| med | medicína | 610 lékařské vědy, 613 hygiena, 615 terapie a farmakologie | medicína (anatomie, bakteriologie, farmakologie, fyziologie, homeopatie, chirurgie, imunologie, neurologie, oftalmologie, psychiatrie, psychologie, veterinární medicína, zubní lékařství), zdraví (alternativní medicína (akupunktura, aromaterapie)) |
| zoo | zoologie | 562 paleozoologie bezobratlých, 563 jednoduché formy, 564 měkkýši, 565 ostatní fosilní bezobratlí, 566 paleozoologie obratlovců, 567 anamnia, 568 sauropsida, 569 savci, 590 zoologické vědy | zoologie (entomologie, ornitologie), vývoj zvířecích druhů |
| bot | botanika | 561 paleobotanika, 580 botanic- | botanika |

| | | | |
|------------|-----------------------------|--|---|
| | | ké vědy | |
| bio | biologie | 574 biologie, 575 organický vývoj, 576 mikrobiologie, 578 mikroskopy a mikroskopie, 579 sbírky a konzervování | genetika, biologie (cytologie, histologie, mikrobiologie), paleontologie, biochemie a biotechnologie |
| ant | antropologie | 572 antropologie, 573 fyzická antropologie | antropologie |
| che | chemie | 540 chemie, 660 chemická technologie | biochemie, chemie |
| mat | matematika | 310 statistika, 510 matematika | matematika (statistika) |
| log | logika | 160 logika | |
| ggr | geografie | 910 zeměpis, cesty | fyzikální zeměpis, geografie (kartografie), topografie |
| ast | astronomie | 520 astronomie a příbuzné vědy | astronomie |
| phy | fyzika | 530 fyzika | fyzika (fyzika částic, nucleární fyzika), optika (mikroskopie), akustika, mechanika |
| met | meteorologie | | meteorologie |
| geo | geologie, hydrologie | 550 vědy o Zemi, 552 petrologie | petrografie, oceanografie, mineralogie, geologie, speleologie |
| env | ekologie, životní prostředí | 710 tvorba krajiny, urbanistika | ekologie, ochrana přírody, životní prostředí |
| TEC | TECHNIKA | 600 technika (kromě 601 filosofie a teorie), 620 inženýrství | astronautika |
| tra | doprava, spoje | 383 pošty, 384 telekomunikace, 385 železniční doprava, 386 vnitrozemská vodní doprava, 387 námořní a letecká doprava, 388 dálniční a městská doprava | telekomunikace, vysílání (rozhlas, televize), letectví (vojenské letectvo), námořnictvo (válečné námořnictvo), doprava (železnice), automobilismus (motorkářství), lodě, komunikace = doprava a spoje, kosmonautika |
| ene | energetika | | energie (ropný průmysl) |
| ind | průmysl, technika | 654-655 polygrafie a vydavatelství, 658 řízení průmyslu, 670 výroba, 678 gumové a plastové materiály, 680 ostatní výroby | technologie, biotechnologie, metalurgie, elektronika, audio (nahrávání), látky (barvení látek), tisk (litografie), řemesla, nakladatelská práce, průmysl, zpracovatelský průmysl, důlní průmysl, balení a obaly, strojírenství, těžba dřeva, dřevo, tesařské řemeslo, instalatérské práce |
| bui | stavebnictví | 690 pozemní stavitelství | budova |
| com | informatika a počítače | | věda o počítačích (umělá inteligence, elektronické vědy) |
| sta | normalizace a metrologie | 389 metrologie a normalizace | taxonomie, měření času |

| | | | |
|------------|--|---|--|
| ECN | EKONOMIE A ŘÍZENÍ | | |
| eco | ekonomie, obchod, bankovníctví | 330 ekonomika, 334 družstevnictví, 337 celní politika, 381 vnitřní obchod, 382 mezinárodní obchod, 656-657 účetnictví, 659 ostatní problémy obchodu | finance (bankovníctví), obchod (maloobchodní podnikání), zaměstnání, vlastnictví, marketing, účetnictví, obchod, mince, hospodářství |
| man | management, řízení | | reklama (maloobchodní podnikání) |
| mer | zbožiznalství a spotřebitel | | spotřebitel |
| BEL | VÍRA | | |
| rel | náboženství | 200 náboženství, 210 přírodní teologie, 220 bible, 230 dogmatická teologie, 240 devocionální a praktická teologie, 250 pastorační teologie, 260 křesťanská církev, 270 dějiny křesťanské církve, 280 křesťanské církve a sekty, 290 ostatní náboženství | náboženství (budhismus, křesťanství (anglikánství, římskokatolická církev), hinduismus, islám, judaismus, pravoslavná církev), new age |
| sup | nadpřirozeno, okultní vědy, magie | | nadpřirozeno, okultní vědy, magie, astrologie |
| LIF | ŽIVOTNÍ STYL | | obecné zájmy (mužské), ženy |
| hou | domácí hospodářství, stravování, odívání, byt, ruční práce | 640 domácí hospodářství | jídlo (řeznictví), domov (zařízení domu, osvětlení), /řemesla/ = pletení, šití, háčkování, barevné sklo, výroba prošíváných dek, krejčovství, tetování, textil, víno, móda, kadeřnictví, domácnost, kuchařství, dámské krejčovství |
| spo | sport | 796 atletika a hry ve volné přírodě nebo na hřišti, 797 vodní sporty, 798 jezdecké sporty, 799 sportovní rybářství, myslivectví, střelba | sport (americký fotbal, atletika, automobilové závody, badminton, baseball, basketbal, bodybuilding, bojová umění, bowling, box, bruslení, býčí zápasy, curling, cyklistika, fotbal, golf, gymnastika, holubářství, jachting, jezdectví, kanoistika, karate, kriket, kulečnick, lacrosse, lyžování, parašutismus, plachetnice, plavání, pozemní hokej, rodeo, rugby, rybaření, skateboarding, skoky do vody, snowboarding, sportovní potápění, střelba, sumo, surfing, šerm, šipky, tenis, veslování, vodní sporty, volejbal, vzpírání, windsurfing, wrestling, závěsné létání, zimní hokej), jóga, sokolnictví, horolezectví, lukostřelba |

| sct | společenský život | 060 všeobecné společnosti | Svatby |
|-----|--|--|---|
| amu | zábava, hry, volný čas, cestování | 791 veřejná zábava, 793 hry a zábavy v místnosti, 794 hry šikovnosti a dovednosti, 795 hry nahodilé | závody (sázení), hry (karetní hry), domácí mazlíčci (psi, kočky), skauting, známky, studentské zájmy, zájmy teenagerů, cestování, zahradničení, hobby (musí mít vždy další specifikaci), volný čas, knihy, bridž, karavanning, šachy, komiksy, kutilství, zábava, sex |
| min | skupiny se specifickými zájmy - důchodci, etnické zájmy, postižení, homosexuálové, drogy | | mláďí, důchod, feminismus, gayové a lesbičky, invalidita, drogy, etnické zájmy, Hispanoameričané, Afroameričané, Asioameričané |
| reg | region | 354 ostatní země, 940 Evropa, 950 Asie, 960 Afrika, 970 Severní Amerika, 980 Jižní Amerika, 990 ostrovy v Pacifiku | venkov, regionální zájmy (Amisch, Haiti), Rusko, rozvojové země, Antily, Evropa, Indie, Japonsko, Židovství, Střední východ, Newfoundland, Afrika, Aljaška, Asie, Austrálie, Kanada, Čína |

Na závěr je třeba říci, že téma/námět textu (*topic*) patří dodnes k poněkud kontroverzním bodům korpusových projektů. Přes veškerou snahu je totiž obtížné zaručit, že všechny texty (dnes i v budoucnu) budou snadno zařaditelné k některé z dříve vzniklých kategorií. Je také poněkud nesnadné definovat (pro neodborníky, jakými lingvisté v ostatních oborech jsou) jejich hranice a poskytnout tak lidem, kteří pro účely zařazení do korpusu texty kategorizují, explicitní vodítko a kritérium jiné, nežli je jejich osobní náhled na věc.

Přes tyto nevýhody však upřednostňuje většina projektů a také ČNK alespoň základní pomocnou síť kategorií, protože - vzato do důsledků - seznam určitelných kategorií může být téměř nekonečný a předmětem zkoumání se za určitých okolností může stát i velmi malá a přesně vymezená podoblast. Rezignovat kvůli podobným obtížím na jakékoli - byť pro někoho příliš obecné, pro jiného třeba nepřesné - dělení, by bylo rozhodně škoda.

Literatura:

Aston, G. - Burnard, L.: *The BNC Handbook*. 1. vyd. Edinburgh University Press, Cambridge 1998.

Čermák, F.: Czech National Corpus: A case in many contexts. *International Journal of Corpus*

Linguistics, 2, 1997, s. 181-197.

Čermák, F. - Králík, J. - Kučera, K.: Receptce současné češtiny a reprezentativnost korpusu.

SaS, 58, 1997, s. 117-124.

Kennedy, G.: *An Introduction to Corpus Linguistics*. Addison Wesley Longman Limited, Harlow 1998.

Králík, J.: Vyvážení zdrojů reprezentativního korpusu synchronní češtiny SYN2000. *SaS*, 62, 2001, s. xx-xx.

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

Ústav Českého národního korpusu FF UK

nám. Jana Palacha 2, Praha 1