Zprávy / News

Co je v ČNK nového X

Václav Cvrček, Pavel Vondřička, Martina Waclawičová
Ústav Českého národního korpusu, Filozofická fakulta UK Praha
{vaclav.cvrcek; pavel.vondricka; martina.waclawicova}@ff.cuni.cz

V rámci jubilejního roku 2019, v němž projekt ČNK oslavil 25. výročí svého založení, byla jako hlavní inovace portálu www.korpus.cz představena aplikace Slovo v kostce (viz KGA 20/2019). To nebyl nicméně jediný počin, kterým jsme oslavili toto půlkulaté jubileum. Vedle pravidelné aktualizace vlajkových korpusů – zejm. synchronního psaného korpusu SYN (verze 8) a paralelního korpusu InterCorp (verze 12) – bylo zveřejněno několik korpusů nových, konkrétně: korpus polooficiální internetové komunikace NET, korpus monologů ORATOR a autorské korpusy Karla Čapka a Jana Čepa (autor R. Změlík). Dokumentace ke všem těmto korpusům je dostupná v rámci wiki stránek projektu ČNK (wiki.korpus.cz).

Ještě větší rozšíření ovšem zaznamenal segment uživatelských nástrojů, kterému se budeme věnovat v tomto příspěvku. Ještě v roce 2019 přibyly do rodiny korpusových aplikací statistická kalkulačka Calc a prohlížeč frekvenčních seznamů Lists, v roce 2020 pak následovaly Mapka a KorpusDB. Vzhledem k tomu, že se jedná o specializovanější nástroje, úzce zaměřené na jeden typ výzkumného úkolu, vyhradili jsme jim v nástrojové liště portálu korpus.cz specifické místo, které je označeno symbolem aplikačního menu (tzv. „čokoláda“).

První jmenovaná aplikace s názvem **Calc** je dostupná na adrese www.korpus.cz/calc. Jedná se o nástroj pro statistické vyhodnocování úloh, s nimiž se uživatel korpusu při své práci běžně setkává. Podle typu úlohy je kalkulačka rozdělena do několika modulů, mezi nimiž lze přepínat pomocí menu na levé straně. Jednou z nejčastějších otázek, kterou při korpusovém výzkumu uživatelé řeší, je to, zda mají dva jevy, které v korpusu vyhledali, statisticky signifikantně odlišnou frekvenci. Calc pomůže s vyhodnocením tohoto problému, ať už se jedná o úlohu, v níž porovnáváme dvě slova v jednom korpusu (např. varianty jednoho jevu), jedno slovo ve dvou korpusech (např. porovnání různých historických období), či dvě slova ve dvou korpusech (např. ekvivalenty při kontrastivním výzkumu). Důraz je přitom kladen na jednoduchost a interaktivitu zadávání a srozumitelnou prezentaci výsledků, které jsou doplněny o vysvětlující text v češtině a angličtině.

Další moduly jsou určeny pro řešení specifičtějších úloh. Modul nazvaný „1 jev – víc vzorků“ pomáhá určit spolehlivost manuální analýzy konkordance na základě několika náhodných vzorků. Modul s inverzním názvem „Víc jevů – 1 vzorek“ slouží

naopak k rozhodnutí o doloženosti určitých jevů v rámci konkordance (např. analýza jednotlivých významů). Zbývající dva moduly vyžadují pro vyhodnocení vstupních údajů informaci o jazyce, v němž je analýza prováděna – jeden počítá lexikální bohatost textu na základě míry zTTR, která využívá referenční hodnoty (specifické pro každý jazyk a registr), a není proto závislá na délce textu, druhý se pokouší odhadnout, jaká je adekvátní korespondence n-gramů v různých jazycích (rozsah víceslovných jednotek se v jazycích mění podle jejich typologické charakteristiky, např. izolační jazyky mají tendenci vytvářet rozsáhlejší víceslovné jednotky než jazyky s flexí). Vedle češtiny jsou poslední dva moduly schopné vyhodnocovat výsledky i pro dalších 7 jazyků: angličtinu, němčinu, francouzštinu, polštinu, španělštinu, nizozemštinu a portugalštinu.

Posledním nástrojem zveřejněným v roce 2019 je **Lists** (www.korpus.cz/lists), který reaguje na potřebu některých uživatelů ČNK procházet frekvenční seznamy (např. při vybírání vhodných slovních stimulů v rámci psychologického či neurologického výzkumu). Nabízí snadné prohlížení a srovnávání frekvenčních seznamů z vybraných korpusů ČNK (SYN 2000, SYN 2005, SYN 2010, SYN 2015 a ORAL v1) a dále jejich interaktivní filtrování, které může být založeno jak na informaci o frekvenci, tak i na podobě konkrétních slovních tvarů či lemmat. Vedle absolutní frekvence jednotek nabízí aplikace také relativní frekvenci (ipm), průměrnou redukovanou četnost (ARF) a průměrnou redukovanou četnost na milion slov (ARFn). Tabulka pro lemmata je navíc doplněna o informaci o slovních druzích, kterou lze rovněž využít pro filtrování.

Data ze zdrojových korpusů poskytují informaci o čtyřech hlavních registrech (beletrie, oborová literatura, publicistika jsou obsaženy v korpusech řady SYN a konverzace pochází z korpusu ORAL). Toto třídění aplikace Lists využívá k tomu, že na základě registrů lze informace o frekvencích jednotek dále filtrovat a vzájemně porovnávat.

Aplikace **KorpusDB** (zveřejněná v roce 2020) je dostupná na adrese db.korpus.cz a obsahuje aktuálně databázi všech rozpoznaných slovních tvarů všech lemmat vyskytujících se v korpusech ČNK a jejich frekvenčních distribucí. Kromě souhrnu všech korpusů současného psaného jazyka (celý korpus SYN v8) pokrývají frekvenční statistiky i korpusy mluveného jazyka (ORAL v1, ORTOFON v1) a korpusy diachronní (DIAKORP v6, dosud nezveřejněný korpus textů 19. století). Všechna data byla pro účel vytvoření databáze označkována stejným způsobem.

Kromě vyhledávání konkrétních slovních tvarů nabízí rozhraní náhled celých paradigmat jednotlivých lemmat, kde lze pomocí filtrů snadno nalézt gramatické kategorie, v nichž dochází k reálné konkurenci různých variant v jazykových datech. V přehledu se zobrazuje rychlé srovnání relativní frekvence jednotlivých variant v pěti základních skupinách textů: beletrii, oborové literatuře a publicistice v rámci současného psaného jazyka, a dále v diachronních a mluvených korpusech. Zobrazit lze však i podrobnější frekvenční charakteristiky podle dalších textových atributů korpusů ČNK či podle sociolingvistických kategorií mluvčích v mluvených korpusech.

V rámci současného psaného jazyka tak lze rychle posoudit např. i případné vývojové tendence v čase. V úvahu je třeba brát nedostatky v datech a chyby v disambiguaci homonymních tvarů – typickým problémem je částečná normalizace v transkripci mluvených dat i extrémní obtížnost správného rozlišení homonymie pádů v mluveném jazyce.

V červenci 2020 byla dále zveřejněna webová aplikace **Mapka** (www.korpus.cz/mapka/), která je doplňkem ke korpusu DIALEKT (viz KGA 17/2018) a zároveň přináší vhled do nářeční diferencovanosti českého jazyka. Korpus DIALEKT reprezentuje v Mapce několik druhů sítí lokalit, ve kterých probíhalo nahrávání nářečních projevů. Je možné si zobrazit lokality zařazené do současné verze korpusu DIALEKT, přičemž lze vyčlenit zvlášť ty, které se vztahují ke staré časové vrstvě sběru (od 50. do 80. let 20. stol.) a k nové časové vrstvě sběru (od 90. let 20. stol. do současnosti). Aktuální stav sběru nářečního materiálu lze sledovat na dalších sítích lokalit. Jde jednak o místa nahrávání, jednak o místa, odkud pocházejí zaznamenaní mluvčí (vzhledem ke způsobu výběru mluvčích se tyto sítě téměř shodují). Další možností je zobrazení sítě výzkumných lokalit Českého jazykového atlasu. Do budoucna se počítá i se začleněním dat vycházejících z dalších korpusů mluvené češtiny, ORALu a ORTOFONu.

Mapka poskytuje i základní informace o nářečním členění českého jazykového území. Lze si zobrazit buď základní nářeční oblasti, nebo podrobnější dělení na nářeční podskupiny, úseky a typy. Vedle toho je možné si na mapě zvolit administrativní členění ČR na okresy nebo kraje. V mapě jsou vloženy přehledy charakteristických nářečních rysů příslušící hlavním regionům (Čechy, Morava a Slezsko) a všem základním nářečním oblastem. Popis je vždy zaměřen na rysy hláskoslovné a tvaroslovné. Příklady nářečních jevů byly vybírány především z aktuální verze korpusu DIALEKT, doplňkově i z přepisů, které budou zveřejněny v další verzi korpusu. Každou nářeční oblast navíc reprezentují dvě ukázky autentických projevů nářečních mluvčích ve formě zvukového záznamu a dialektologického a ortografického přepisu. Ukázky pocházejí z korpusu DIALEKT a byly zvoleny tak, aby demonstrovaly typické nářeční rysy daných nářečních oblastí. Pro každou oblast bylo vybráno po jedné ukázce ze staré i nové časové vrstvy sběru. Postupně budou přidávány ukázky nářečních projevů i k dalším nářečním podskupinám, úsekům a typům. Pro potřeby vlastního výzkumu je do Mapky začleněna možnost vyhledávání obcí nebo částí obcí v rámci ČR a jejich zařazení do systému nářečního územního členění. Vyhledané body lze následně v mapě zakreslit a vytvořit tak vlastní mapu, kterou je možné stáhnout a vytisknout.