TISKOVÁ ZPRÁVA

Cílem výzkumného záměru MŠMT Český národní korpus a korpusy dalších jazyků (MSM0021620823), řešeného v letech 2005–2011, bylo extenzívní kontinuální mapování češtiny a vytváření jazykových korpusů, které v současné době představují materiálově nejrozsáhlejší strukturovanou bázi pro základní i aplikovaný výzkum celku českého jazyka v jeho různých podobách a v jeho starším i současném vývoji. Jako projekt svou podstatou kontinuální navazoval výzkumný záměr plynule na souvislou řadu projektů, na nichž řešitelé pracovali v letech 1996–2004 a k nimž patřil i bezprostředně předcházející stejnojmenný výzkumný záměr MSM 112100002. Výzkumný záměr byl realizován dvěma řešitelskými pracovišti na Filozofické fakultě Univerzity Karlovy – Ústavem Českého národního korpusu (ÚČNK) a Ústavem srovnávací jazykovědy (ÚSJ), na jeho řešení se však podílel i Ústav teoretické a komputační lingvistiky a více než dvacet lingvistických kateder a ústavů FF UK. V rámci řešení VZ byla dále rozvíjena i tradiční spolupráce s řadou pracovišť mimofakultních (Ústav formální a aplikované lingvistiky MFF UK) a mimouniverzitních (Ústav pro jazyk český a Ústav pro českou literaturu AV ČR, Fakulta informatiky MU, ČVUT), včetně pracovišť zahraničních (smluvní spolupráce s více než 20 institucemi).

K hlavním výsledkům řešení výzkumného záměru v ÚČNK patří především vytvoření a internetové zveřejnění reprezentativních, žánrově vyvážených a gramaticky označkovaných synchronních psaných korpusů SYN2005, SYN2010 a doplňkových korpusů publicistických textů SYN2006PUB, SYN2009PUB. Celkově byly uživatelům zpřístupněny korpusy o rozsahu zhruba 1,2 miliardy slovních tvarů určené k využití při tvorbě jazykových příruček, zejména slovníků, a k mnohostrannému jazykovému výzkumu současného psaného jazyka. Vedle toho byla v rámci výzkumného záměru významně rozšířena i řada specializovaných korpusů, jmenovitě korpusů paralelních (v jejich rámci byly systematicky paralelně uspořádány jak překlady českých textů do 22 jazyků, tak české překlady z 22 jazyků, o celkovém rozsahu přibližně 92 milionů slovních tvarů), dále korpusů mluvených (autentický mluvený jazyk v nahrávkách a přepisech, celkově přibližně 2 miliony slovních tvarů) a diachronního korpusu (české texty od konce 13. století zhruba do poloviny 20. století, celkový rozsah přes 4 500 000 slovních tvarů). Na řešitelském pracovišti ÚSJ patří k nejvýznamnějším výsledkům práce na výzkumném záměru mnohonásobné rozšíření diachronního korpusu arabštiny z necelých 20 milionů na 384 milionů slovních tvarů, zpracování korpusu jaghnóbštiny v rozsahu přes 100 000 slovních tvarů a vytvoření komplexního korpusu ugaritštiny (propojení korpusu s kritickými edicemi a slovníkem).

Souběžně s tvorbou a rozšiřováním korpusů byla během řešení výzkumného záměru dále propracovávána teorie gramatického i strukturního značkování korpusů, a to zejména ve směru morfologické desambiguace, syntaktického značkování a identifikace frazémů. Hlavním výsledkem této činnosti řešitelů bylo jednak výrazné zkvalitnění morfologické analýzy v korpusech současných psaných českých textů, jednak vytvoření koncepce lemmatizace a gramatického značkování v mluvených korpusech a v korpusech starších českých i jinojazyčných textů. V souvislosti s tím byly na jednotlivé korpusy aplikovány nové nástroje (bylo spuštěno webové rozhraní Sketch Engine) a další specializované nástroje byly vytvořeny (především InterText, software vyvinutý v ÚČNK pro správu, zpracování a zarovnání různojazyčných paralelních textů ve formátu XML).

Z dosavadních zkušeností s uplatněním výsledků projektu je zřejmé, že existence reprezentativních korpusů přinesla zásadní obrat do výzkumu v celém oboru lingvistiky a že řada výzkumných činností a úkolů je dnes bez nich už nepředstavitelná. O tom svědčí jak řada knižních i jiných publikací, jako jsou Statistiky češtiny (Praha, NLN 2009), jejichž vznik by byl bez korpusu prakticky nemožný, tak dlouhodobě vzrůstající počet registrovaných uživatelů Českého národního korpusu, který koncem roku 2011 překročil hranici 2 700.