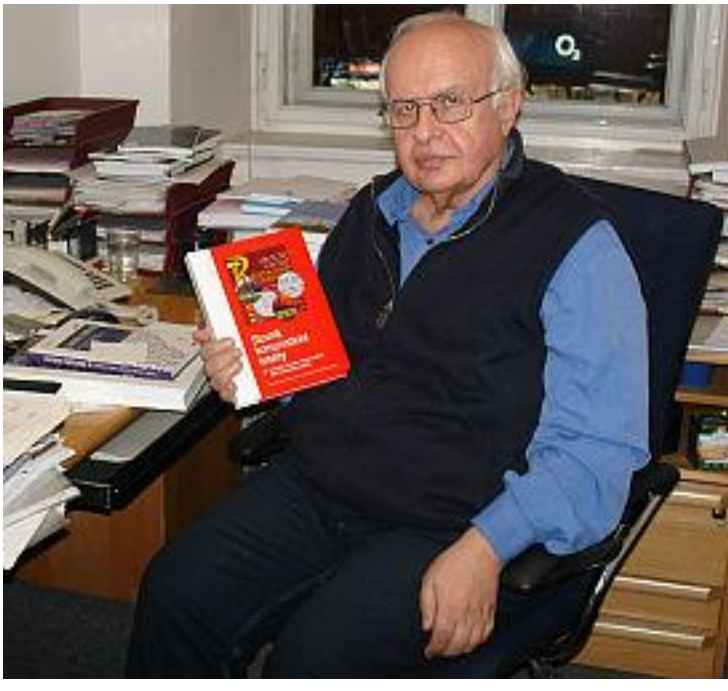
Slovník komunismu zpřístupněn veřejnosti

Slovník komunismu zpřístupněn veřejnosti

Na konci loňského roku vyšel Slovník komunistické totality, který je prvním pokusem o zmapování jazyka oficiální komunistické propagandy. Slovník je dílem odborníků z Ústavu Českého národního korpusu, ojedinělého pracoviště, které funguje při Filozofické fakultě UK. O nové publikaci, ale i o činnosti ústavu jsme si povídali s jeho ředitelem prof. PhDr. Františkem Čermákem, DrSc.

Co to vlastně je „ten“ korpus, kterému se váš ústav věnuje?
Korpus je soubor počítačově uložených textů (v případě mluveného jazyka – přepisů záznamu mluvy), který slouží k jazykovému výzkumu. Všechny naše korpusy jsou postupně zveřejňovány na internetu, aby mohly sloužit i veřejnosti. V současné době zahrnují přes 3 miliardy slov, díky čemuž je náš korpus tím největším v Evropě a jedním z největších i na světě. Ústav Českého národního korpusu je unikátním pracovištěm v ČR, které na svých úkolech spolupracuje s přibližně 15 dalšími vědeckými institucemi.



Z jakého popudu vznikl projekt Český národní korpus zaměřující se na budování počítačového korpusu především psané češtiny a co už obsahuje?
Český národní korpus vznikl původně z úsilí a potřeb lingvistů, dnes se o něj ale zajímají jak literáti, tak historici, sociologové a další odborníci. Důraz je kladen na současný jazyk, který máme podchycen poměrně vyčerpávajícím způsobem. Žádný jazyk totiž není možné popsat úplně, zvláště například v oblasti odborné terminologie.
Udělám malou odbočku. Kdysi se mohla česká lingvistika opírat o činnost Pražského lingvistického kroužku, dnes už se ale z této slávy čerpat nedá. Češtině stále chybí velký slovník současného jazyka. Myslím, že to je úkol, který nesplnila Akademie věd. Proto jsme si řekli, že připravíme korpusy, které by mohly postavit základy takovému dílu. Protože dnes už není možné takový slovník vytvořit bez rozsáhlého korpusu.
Prozatím jsme vydali Frekvenční slovník češtiny, který vycházel ze 100 milionů slov v korpusu. Tento slovník zaznamenává mimo jiné 50 000 nejběžnějších obecných slov, což usnadní práci například autorům překladových slovníků, kteří budou moci vhodněji vybrat ta nejčastější používaná slova. Velmi intenzivně také sbíráme mluvený jazyk po celých Čechách a na Moravě, v roce 2007 jsme například vydali Frekvenční slovník mluvené češtiny. Snažíme se jít ale i do minulosti, na webu jsme například zpřístupnili diachronní korpus v první verzi. Jeho rozšiřování je ovšem omezeno jak kapacitou našich odborníků, tak dostupností a rychlostí zpracování textů. Najdete tu vybrané texty z Komenského či Husa. Třetím naším důležitým počinem je projekt Intercorp, tedy projekt paralelních korpusů. Jeho cílem je vybudovat paralelní synchronní korpusy pro většinu jazyků studovaných na FF, vždy pro daný jazyk a češtinu.
Na konci roku vyšla rozsáhlá publikace Slovník komunistické totality. Proč jste se zaměřili právě na toto období a jak dlouho váš tým slovník připravoval?
Už dlouho jsem měl dojem, že právě takováto příručka uživatelům schází. Vlastní zpracování bylo jen špičkou ledovce, samo trvalo asi rok. Daleko déle trvalo dát dohromady potřebná data. Na základě doporučení od historiků jsme vybrali

čtyři kritická čtvrtletí Rudého práva. Ta jsme zadali ke skenování, což zabralo několik let. Dále jsme do výzkumu zařadili asi 100 nejvýznamnějších propagátorských příruček.

Co všechno ve slovníku uživatelé najdou?

Vybraná slova jsme seřadili podle frekvence a pomocí indexů srovnali se současným jazykem. To by ale bylo málo. Málokoho zajímají jen suchá čísla. Proto jsme dodali i dobové kolokace, tedy typické kombinace slov, které pro tu dobu byly příznačné. Knížka má i obsáhlou úvodní studii, jež nabízí přehled specifických rysů z oblasti pragmatiky a sémantiky, na závěr jsme zařadili některé typické texty a pamflety té doby. Uživatelé zároveň získávají úplný korpus Totalita dokonce na CD, mohou si ho zkopírovat do svého počítače a sami v něm vyhledávat cokoliv dalšího. Víc už se do takového díla nevešlo, protože by bylo jen obtížně prodejné.

Jak velký tým na publikaci pracoval?

Samotný tým čítal asi 10 lidí, v to nepočítám externisty, kteří se podíleli na skenování.

Považujete téma jazyka komunistické propagandy za uzavřenou kapitolu, nebo se mu budete v nějaké podobě věnovat i dál?

Bohužel pro další výzkum už nemáme data. Archiv Rudého práva nemá ani KSČM, my to za ně pořizovat nebudeme. Nevylučuji ale, že se k tématu někdy vrátíme. Nyní se více chceme soustředit na nezpracované 16. a 17. století a období baroka. Specializované menší studie z daného materiálu jsou ale pravděpodobné. do boxu

Co je to korpus

Korpus je soubor počítačově uložených textů (v případě mluveného jazyka – přepisů záznamu mluvy), který slouží k jazykovému výzkumu. K práci s tímto korpusem slouží speciální vyhledávací program. S jeho pomocí je možné vyhledávat slova a slovní spojení v kontextu a zjistit jejich frekvenci v korpuse i původní textový zdroj. Umožňuje i další zpracování nalezeného (např. abecední třídění apod.). U některých korpusů lze vyhledávat například i podle slovních druhů (a mnoha dalších gramatických kategorií).

Co je Český národní korpus

Český národní korpus (ČNK) je akademický projekt zaměřený na budování rozsáhlého počítačového korpusu především psané češtiny. Pracuje na něm Ústav Českého národního korpusu na Filozofické fakultě Univerzity Karlovy v Praze (ÚČNK). Od svého založení roku 1994 má ÚČNK na starosti budování ČNK, jeho rozvoj a rovněž činnosti související, zvláště v oblasti výuky a pěstování oboru korpusová lingvistika.

Čermák, F., Cvrček, V., Schmiedtová, V. (eds): Slovník komunistické totality. Nakladatelství Lidové noviny, Praha 2010. ISBN 978-80-7422-060-9

(Lucie Kettnerová)