ČEŠKI NARODNI KORPUS

Besedilo je nastalo po enomesečnem študijskem dopustu, ki sem ga kot štipendist CEEPUS-a preživel v Pragi spomladi 1997. Poleg Filozofske fakultete, Inštituta za češki jezik pri Češki akademiji znanosti in Inštituta jezikovne in strokovne priprave tujih študentov Karlove univerze sem obiskal tudi Inštitut Češkega narodnega korpusa na Filozofski fakulteti.[1]

1 Korpus je enovita, notranje strukturirana in standardno označena zbirka besedil v elektronski obliki, zbranih po izhodiščnih kriterijih glede na namen samega korpusa (Meyer, Mackintosh 1996: 260). V splošnem referenčnem korpusu so besedila zbrana tako, da ta z vidika različnih besedil predstavlja reprezentativno besedilno zbirko.

Aktualnost korpusov predvsem v zadnjem desetletju, njihova široka uporabnost tako za jezikoslovne kot tudi nejezikoslovne študije, novi pristopi k analizam, ki omogočajo širše in globlje razumevanje jezika, so le nekateri od razlogov za označevanje zadnjega desetletja tega stoletja kot desetletja korpusnega jezikoslovja (Čermák 1995a: 119). – Z jezikoslovnega vidika je uporaba korpusa izjemno dragocena, saj v analize jezika vnaša večjo verodostojnost; velik obseg načrtno zbranega gradiva namreč omogoča izpostavitev v jeziku tipičnega in zmanjšuje možnost interpretiranja le obrobnega kot temeljnega. Sistematično delo s korpusom tako pomeni možnost natančnejšega spoznavanja celovitega delovanja jezika, kot je bilo možno kadarkoli prej (Čermák 1995a: 119, 121; Svartvijk 1992: 8).

Prevladujoči področji uporabe korpusov sta leksikologija in predvsem leksikografija, ki že po tradiciji izhajata iz sistematično zbranega minimalnobesedilnega gradiva (McEnery, Wilson 1996: 90); prav slovarji pa so tudi prvi široki javnosti dostopni rezultati uporabe korpusov. Vendar se danes s pojavitvijo večjega števila širšemu krogu dostopnih različnih tipov korpusov uporabnost širi na vsa jezikoslovna področja, tako na raziskave jezikovnosistemskih lastnosti jezikov, še bolj pa na področja, ki so tradicionalno gradivno usmerjena – besediloslovje, stilistika, sociolingvistika, psiholingvistika, ipd. (McEnery, Wilson 1996: 98–101, 111) ter na področje uporabnega jezikoslovja in prevodoslovja.

2 Aktualnost korpusa v jezikoslovju z zavestjo, da ta lahko veliko pripomore ne le h kultiviranju in razumevanju jezika, ampak tudi k opisovanju in razumevanju sodobne družbe in kulture, kot se odraža v jeziku, je spodbudila razmišljanje o korpusu češčine že v začetku devetdesetih let. Večletna prizadevanja so pripeljala do ustanovitve Inštituta Češkega narodnega korpusa v okviru Filozofske fakultete Karlove univerze v Pragi septembra 1994, v celoti pa je delo steklo oktobra 1996, ko se je na Inštitutu zaposlila stalna skupina raziskovalcev in se je tudi delo preselilo v na novo opremljene prostore (ICNC).

Pri projektu sodelujejo tri češke univerze, in sicer Karlova univerza v Pragi (Inštitut teoretične in računalniške lingvistike, Inštitut čeških študij in Oddelek za češki jezik Filozofske fakultete, Inštitut za formalno in uporabno jezikoslovje, Fakulteta za matematiko in fiziko), Češka tehnična univerza v Pragi (Oddelek za računalništvo in inženiring) in Masarykova univerza v Brnu (Oddelek za češki jezik Filozofske fakultete, Fakulteta za informacijsko tehnologijo) ter Češka akademija znanosti. Celotno delo v okviru Inštituta

[1] Zahvaljujem se prof. dr. Františku Čermáku, prof. dr. Věri Schmiedtovi in mag. Michalu Šulcu, ki so me prijazno seznanjali z delom svoje skupine. Pogovor z njimi je v marsičem podlaga za ta zapis.

Češkega narodnega korpusa na Filozofski fakulteti v Pragi koordinira prof. dr. František Čermák.

Inštitucionalna povezanost je omogočila tudi lažje pridobivanje finančnih sredstev, potrebnih za nemoteno delovanje Inštituta. Glavnino sredstev je Inštitut pridobil pri državni agenciji in agenciji Karlove univerze, ki financirata raziskovalne projekte, finančno pa ga podpitara tudi dve večji češki banki in založništvo Lidové noviny.

2.1 Projekt Češkega narodnega korpusa je izredno obsežen, saj predvideva kot celoto sinhrono in diahrono besedilno zbirko tako pisnih kot govornih besedil. Gre torej za večdelni splošni korpus češčine, ki pa bo kasneje omogočal tudi oblikovanje novih korpusov za posebne raziskovalne namene, npr. korpusa strokovnih besedil. Prav specializirani korpusi strokovnih besedil postajajo vse aktualnejši. Če je v začetku uporabe korpusov zaradi izredne dinamike razvoja strok terminologija in terminografija uporabljaja v veliki meri še tradicionalne pristope (Mayer, Mackintosh 1996: 285), sta z vzpostavljeno dinamiko gradnje korpusov in njihovega nenehnega nadgrajevanja postali to področji, ki prav zaradi možnosti hitrega sprotnega opazovanja jezikovnih sprememb vse bolj temeljita ne delu s korpusom.

Trenutno je delo na Inštitutu usmerjeno predvsem k oblikovanju sinhronega pisnega dela korpusa. Določitev sinhronega vidika je v izhodišču formalistična. Predvideva zajetje besedil, izdanih po letu 1989, ko se pojavi dovolj velika količina besedil v elektronski obliki, hkrati pa velja še omejitev starosti avtorja. Postavljena je v leto 1890, tj. rojstno leto K. Čapka, katerega jezik se še danes prepoznava kot sodobna češčina; njegova besedila so v razmerju do govorjenega jezika postavila tudi pisni standard govorjenega knjižnega jezika, sprejemljivega za celotni češki prostor. – Tako postavljena merila bi npr. izločila svetopisemska besedila in v celoti besedila klasične literature. Ker pa to seveda ni njihov namen, so postavljena dovolj prožno in se sproti oblikujejo glede na doseganje reprezentativnosti korpusa. Merila reprezentativnosti se v celoti oblikujejo na novo, zato je razumljivo, da nastajajo ob srečevanju s konkretnimi problemi v diskusiji v okviru korpusne skupine.

Za doseganje reprezentativnosti je oblikovana okvirna mreža parametrov glede na jezikovno zvrst, besedilno vrsto, žanrsko pripadnost, medij, v katerem se pojavljajo, ipd. – Razmerja med besedili posameznih zvrsti oziroma besedilnih vrst so določena predvsem glede na podatke o knjižnični izposoji in glede na rezultate anket branosti. Tako so v izhodišču določena količinska razmerja zajemanja umetnostnih (t. i. imaginativnih) in neumetnostnih (t. i. informativnih) besedil. Ankete branosti so v okviru umetnostnih kriterij količinskega zajemanja besedil posameznih literarnih zvrsti (poezije, proze, dramatike), za razmerja med publicističnimi (časopisnimi, revijalnimi) so upoštevani še podatki o nakladi, med strokovnimi pa tudi število periodičnih publikacij ter sama količina besedil posamezne stroke.

2.2 Problem starejših besedilnih zbirk je bil ob hitro se razvijajoči računalniški tehnologiji v veliki meri njihovo hitro zastarevanje. Prav zaradi zagotavljanja trajnosti pa tudi izmenljivosti elektronskih zapisov je bilo potrebno poskrbeti za njihovo standardizacijo. Tako je logično, da tudi češki korpus za računalniški zapis strukture besedil uporablja standard SGML (Standard Generalised Markup Language), ki kot ISO-standard 8879 dejansko zagotavlja trajnost in izmenljivost elektronskih zapisov. Za označevanje same strukture besedila pa so uporabljena s SGML skladna priporočila za označevanje besedil TEI (Text Encoding Initiative). Izbor označevalcev in nivoja označevanja TEI je

pomemben element pri postavitvi korpusa, saj prav tovrstno označevanje omogoča njegovo kasnejšo uporabnost.[2]

2.3 Sredi leta 1997 je bilo v češkem korpusu zbranih 40 milijonov besed,[3] v glavnem iz časopisov (Mladá fronta Dnes, Lidové noviny, Hospodářské noviny), do konca leta pa je predvideno zajetje še nadaljnjih 30 milijonov. – Tudi pri sprejemanju besedil velja pragmatično načelo sprejemanja vsega elektronsko dostopnega, šele v končni fazi se prav zaradi doseganja reprezentativnosti predvideva elektronsko branje in po potrebi tudi vtipkavanje besedil, predvsem manj obsežnih, s katerimi pa se vendarle izredno pogosto srečujemo, npr. besedila na vozovnicah, obrazcih, javni napisi, obvestila ipd. – Del korpusa je že javno dostopen na naslovu http://ucnk.ff.cuni.cz/cnc, za raziskovalne namene pa dostop ni omejen in si ga češki raziskovalci lahko zagodovijo prek telneta.

Navedenke

Čermák, František, 1995a: Jazikovy korpus: Prostředek a zdroj poznání. *Slovo a slovesnost* 56. 119–140.

– – 1995b: Komputační lexikografije. *Manual lexikografije*. Ur. František Čermák, Renata Blatná. Praha: Nakladatelství H&H. 50–71.

Erjavec, Tomaž, 1996/97: Računalniške zbirke besedil. *Jezik in slovstvo* 2/3. 81–95. [http://nl.ijs.si/ tomaz/Bib/SlKorpus/slKorpus-la2/slKorpus-la2.html]

ICNC – *The Institute of the Check National Corpus*. [http://ucnk.ff.cuni.cz/cnc]

McEnery, Tony, Wilson, Andrew, 1996: *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Meyer, Ingrid, Mackintosh, Kristen, 1996: The Corpus from a Terminographer's Viewpoint. *International Jurnal of Corpus linguistics* I/2. 257–285.

Svartvijk, Jan, 1992: Corpus linguistics comes of age. *Direction in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8. 8. 1991*. Ur. Jan Svartvijk. Berlin, New York: Mouton de Gruyter. 7–13.

Vojko Gorjanc
Filozofska fakulteta v Ljubljani

[2] Natančneje o vsem tem v članku T. Erjavca, kjer je predstavljen tudi kratek zgodovinski pregled razvoja korpusov, njihova tipologija in uporabnost (Erjavec 1996/97).

[3] Obseg korpusov je največkrat izražen v številu besed, zgolj izrazno definiranih, torej brez upoštevanja pomenskega kriterija.