Niektoré technické aspekty budovania a používania korpusov

Pavol Domin: Oddelenie Slovenského národného korpusu, JÚĽŠ SAV

Abstract

In the paper “Some technical aspects of building and using corpora” are discussed some technical questions in building the text corpora. The aim is to briefly illustrate what steps are necessary for corpus builders to provide corpus data for users. Solutions of some technical problems in building and providing text corpus chosen by Slovak National Corpus are presented.

V súvislosti s budovaním Slovenského národného korpusu (ďalej SNK) analyzujeme v tomto príspevku niektoré technické aspekty tvorby a používania korpusov. Cieľom je načrtnúť technickú náročnosť spracovania textových dát do formy požadovanej používateľmi korpusov a predstaviť najdôležitejšie technické kroky, ktoré korpusové pracovisko musí podniknúť, aby poskytlo korpusové dáta v primeranej forme najmä na použitie v lingvistickom výskume.

Pod textovým korpusom sa najčastejšie rozumie rozsiahly vnútorne štruktúrovaný a ucelený súbor jazykových dát, ktoré sú elektronicky uložené a spracovávané. Časť textov (niekedy značná) poskytnutých do korpusu nie je v elektronickej forme. Tieto texty v prvom kroku treba digitalizovať: texty sa zoskenujú a vhodným softvérom na optické rozpoznávanie znakov prevádzajú do elektronickej podoby vhodnej na ďalšie spracovanie.

Texty v elektronickej podobe sa ďalej archivujú v pôvodnom formáte. Každý text sa označí jedinečným identifikačným reťazcom a uvedú sa k nemu základné anotačné údaje (názov textu, dátum dodania textu, zdroj textu, licencia na prístup k textu). Technicky je v SNK táto časť *vonkajšej anotácie* realizovaná zápisom príslušných údajov do osobitných textových súborov. Aby bolo možné jednoducho a efektívne vyhľadávať údaje vo vonkajšej anotácii archívu textov, tieto údaje sa importujú do databázy. Používateľské prostredie na prezeranie vonkajšej anotácie textov SNK je realizované webovským rozhraním nad touto databázou. Vonkajšia anotácia archívu môže byť zvládnutá čiastočne automaticky a čiastočne ľudským anotátorom. Aj časť textov nachádzajúcich sa v súčasnosti v archíve SNK bola anotovaná ručne a časť bola pridaná automaticky rozdelením veľkej skupiny textov.

Štruktúry a formáty textov poskytnutých do SNK môžu byť veľmi rozmanité. Pre ďalšie spracovanie je nevyhnutné všetky texty z archívu previesť do jednotného formátu. Z technického hľadiska existujú dva principiálne spôsoby ako uskutočniť tento prevod. Prvý spôsob je, že na textové dáta archívu sa aplikuje program, ktorý je schopný konverziu spraviť automaticky. Pre veľa štandardných textových formátov sú takéto konverzné nástroje dostupné. Trochu horšia situácia je, ak takéto nástroje pre príslušný formát neexistujú. Vtedy treba buď vyvinúť vlastný konverzný nástroj (ak je tento formát známy), alebo konvertovať ručne. Ručná konverzia

vyžaduje otvorenie príslušného textu v textovom editore schopnom čítať daný typ textu a následný export tohto textu do jednoduchšieho (v ideálnom prípade konečného jednotného) formátu. Tento postup je časovo aj personálne veľmi náročný, technicky je zvládnuteľný len pre málo textov. Preferovaným spôsobom je preto používanie automatických konverzných nástrojov.

Najvýznamnejší z faktorov, ktorý textové korpusy stavia do špecifickej pozície oproti ostatným veľkým dátovým súborom, je fakt, že používateľa nezaujímajú len dáta priamo obsiahnuté v texte. Používateľ potrebuje z korpusových dát veľmi často selektovať informáciu, ktorá v pôvodných textoch nie je priamo obsiahnutá. Sem patrí napríklad hľadanie v textoch len z určitého zdroja alebo obdobia, prípadne hľadanie len v textoch určitého štýlu alebo žánru. Patrí sem aj hľadanie všetkých slov s určitým konkrétnym základným tvarom, gramatickou kategóriou a podobne. Z toho plynú jednak požiadavky na softvér, ktorý musí umožniť tieto informácie vyhľadávať, a jednak na samotné texty, do ktorých tieto dodatočné informácie musia byť nejakým spôsobom dodané.

Anotáciu korpusových textov rozdeľujeme na *vonkajšiu* a *vnútornú*. Do vonkajšej anotácie patria informácie o celých dokumentoch, ako je autor, poskytovateľ, čas vydania, štýl a podobne. Keďže na tejto úrovni sa do textu dodáva informácia relevantná pre celý text alebo dokument, je (v princípe) možné dodávať túto informáciu ručne človekom. V rámci SNK bolo rozhodnuté, že do vonkajšej anotácie nebudú pridávané informácie, ktorých získanie a zapísanie k textu by bolo vzhľadom na technické, personálne a časové možnosti pracoviska SNK príliš náročné a ktoré nie sú striktne vyžadované používateľmi korpusu. Sú to informácie typu autor a žáner novinového článku, paginácia a podobne.

Pracovník SNK by mal mať možnosť pristupovať k dátam vonkajšej anotácie čo najkomfortnejším spôsobom súčasne minimalizujúcim možnosť vnášania preklepov a chýb pri zápise do databázy. Na pracovisku SNK sa vyvíja používateľské prostredie spĺňajúce tieto kritériá.

Do vonkajšej anotácie patrí aj značkovanie rôznych grafických a štruktúrnych častí textu, ako sú nadpisy, odseky, poznámky pod čiarou, tabuľky, obrázky a podobne. Úlohou korpusového pracoviska je vývoj softvéru, ktorý dokáže tieto informácie v texte rozoznať a označiť. Táto anotácia je vo viacerých prípadoch netriviálna, napríklad značkovanie začiatkov a koncov viet môže v niektorých prípadoch vyžadovať istý stupeň interpretácie textu. Napriek tomu je táto časť vonkajšej anotácie veľmi uspokojivo zvládnuteľná automaticky.

Oveľa komplikovanejšia situácia je pri vnútornej anotácii. Pod vnútornou anotáciou rozumieme vnášanie lingvistickej informácie, ktorá vyžaduje značný stupeň interpretácie textu. Ide o značkovanie takých informácií, ako je lema (základný tvar slova), morfologické kategórie slov, dezambiguácia slov a podobne. Je zrejmé, že toto nie je možné robiť vo veľkých korpusoch

ručne. Ľudským anotátorom môže byť značkovaná nejaká malá časť korpusu, na čo treba tiež vyvíjať príjemné počítačové prostredie. Vnútorné značkovanie celého korpusu je možné iba strojovo. Ak si uvedomíme, akým spôsobom funguje analýza textu u človeka, môžeme povedať, že automatická vnútorná anotácia vyžaduje, aby počítač dokázal „myslieť". Prakticky sa v súčasnosti lingvistické myslenie počítača implementuje dvomi spôsobmi. Jedným je štatistický prístup, čiže počítač sa na človekom vopred správne anotovanom korpuse *naučí* pravdepodobnosti výskytu rôznych slovných tvarov v určitom kontexte a takto naučené štruktúry potom aplikuje na neoznačkovaný text. Výhodou tejto metódy je, že počítač od začiatku pracuje s reálnym textom a dokáže dobre zachytiť štruktúry, ktoré sú v texte štatisticky relevantné. Nevýhodou je potreba čo najväčších ručne anotovaných korpusov a menšia úspešnosť pre okrajové javy. Alternatívny spôsob je naopak založený na definovaní presných gramatických pravidiel a ich implementácii do počítača. Výhodou je viera autorov týchto modelov, že dokážu zachytiť jazykové javy podstatne úspešnejšie ako štatistické modely. Nevýhodou je závislosť od aktuálnej gramatickej teórie (ktorá nemusí korešpondovať s reálnymi textovými dátami). Potreba počítačovej implementácie gramatiky navyše môže viesť, najmä pre flektívne jazyky s voľným slovosledom, k ohromnému množstvu pravidiel.

V SNK sa na vnútornú anotáciu a dezambiguáciu spočiatku použijú existujúce štatistické nástroje vyvinuté pre český jazyk (ale už aplikované aj na iné jazyky), ktoré treba pripraviť na rozpoznanie slovenčiny. V prípade neuspokojivej úspešnosti týchto nástrojov bude nutné postupne vyvinúť vlastné nástroje, pravdepodobne založené na gramatických pravidlách.

Anotované dáta je potrebné sprístupniť používateľovi. Existuje viacero možností, ako to urobiť. Jednou z nich je, že korpusové pracovisko sprístupní anotované texty na internete, používateľ si môže ich časť prostredníctvom internetu stiahnuť a študovať. Tento spôsob je reálne možný len pre malé korpusy, na reálne využitie korpusov je tento spôsob prakticky nepoužiteľný. Hlavné dôvody sú tri. Prvý je, že sa cez internet prenáša obrovské množstvo dát, pričom používateľa drvivá väčšina z nich nezaujíma. Druhý a podstatnejší dôvod je, že ak aj takýmto spôsobom požívateľ korpusové dáta získa, nebude schopný ich študovať bežnými počítačovými prostriedkami. Ak chce používateľ získať nejakú relevantnú informáciu, potrebuje väčšinou študovať dáta v ich celistvosti a také obrovské množstvo dát, aké je obsiahnuté v korpusoch, jednoducho žiadny textový editor neotvorí. Tretí, veľmi dôležitý dôvod je ten, že autorské práva často nedovoľujú sprístupnenie celých textov.

Riešením tejto situácie je, aby sa cez internet prenášali len tie informácie, ktoré používateľa zaujímajú. Preto musí existovať špecializovaný program, ktorý spracuje požiadavku používateľa, vyberie z korpusu požadovanú informáciu a až táto je prenesená cez internet. Toto sa dá realizovať tromi spôsobmi:

1. Používateľ sa cez internet nahlási na server, spustí špecializovaný program s požiadavkou na získanie informácie z korpusu, server sa pozrie na médium, na ktorom je korpus uložený, vyberie požadované dáta a odovzdá ich priamo používateľovi.
2. Ďalší spôsob je, že používateľ nekomunikuje priamo so serverom, ale požiadavky zadáva klientovi nainštalovanému na svojom vlastnom lokálnom počítači, ktorý požiadavku spracuje a pošle cez internet serveru poskytovateľa korpusu. Server vráti klientovi požadovanú informáciu a klient ju zobrazí. Server vyhodnocujúci požiadavky klienta a vyberajúci požadované dáta sa (niekedy spolu s klientom) nazýva korpusový manažér. Výhodou tohto spôsobu je zefektívnenie práce, keďže časť operácií s dátami môže robiť samotný klient. Nevýhodou je, že používateľ potrebuje na prístup ku korpusovým dátam špecializovaný klient.
3. Ďalší spôsob prístupu k dátam je cez webovský prehliadač. Používateľ sa vo svojom obľúbenom webovskom prehliadači nahlási na stránku poskytovateľa korpusu a tu zadá svoju požiadavku. Jeho požiadavka je špecializovaným programom, tzv. webovským rozhraním poskytnutá samotnému serveru. Tento požiadavku vyhodnotí a výsledky poskytne rozhraniu, ktoré ich cez internet pošle webovskému prehliadaču používateľa. Uvedený spôsob nevyžaduje od používateľa špecializovaný klient a odbremeňuje server, keďže časť požiadaviek môže spracovávať webovské rozhranie. Nevýhodou je, že zadávanie dotazov na korpus a zobrazovanie výsledkov je obmedzené schopnosťami internetových prehliadačov.

Je veľmi žiaduce zabezpečiť možnosť práce s korpusom aj pre výskumníkov, ktorí majú obmedzený prístup na internet. Preto sa väčšinou časť korpusových dát distribuuje na prenosných médiách, teda prakticky na CD nosičoch. Z technického hľadiska sú možnosti prístupu ku korpusovým dátam podobné ako v prípade prístupu cez internet. Rozdiel je v obmedzení veľkosti poskytovaných dát. Niekedy navyše autorské práva zakazujú tento spôsob sprístupnenia korpusu. Na rozdiel od prístupu ku korpusovým dátam cez internet, kde časť softvéru sa nachádza na vzdialenom serveri, je v tomto prípade nutné, aby všetok softvér potrebný na prístup ku korpusovým dátam bol nainštalovaný na počítači používateľa. To sa môže riešiť buď tým, že sa potrebný softvér dodáva spolu s korpusovými dátami na CD, alebo používateľ musí získať (zakúpiť si) tento softvér z iného zdroja. Môže ísť o rovnaké programy ako v prípade prístupu ku korpusu cez internet (server-klient, server-webovské rozhranie) alebo o špecializovaný softvér na prácu s korpusom na CD.

V SNK sa pripravuje sprístupnenie dát pomocou architektúry klient-server s využitím existujúceho korpusového manažéra Manatee s klientom Bonito (autorsky vypracovaný na Fakulte informatiky Masarykovej univerzity v Brne). Dáta SNK budú sprístupnené aj cez

webovské rozhranie na webovskej stránke SNK. Pracovisko SNK takisto plánuje vyvíjať vlastný korpusový manažér. Rovnako sa počíta s distribúciou časti korpusu na CD nosičoch.

Na záver uvedieme niekoľko poznámok k požiadavkám na korpusový manažér. Korpusový manažér musí vedieť rýchlo pracovať s konkordanciami a pristupovať k informáciám zadaným do textu pri vonkajšej i vnútornej anotácii. Na hľadanie v korpuse musí existovať jednoduchý a zároveň silný jazyk. Ako štandard sa minimálne požaduje schopnosť vyhľadávania podľa regulárnych výrazov, čo je jazyk špecializovaný na hľadanie zložitejších štruktúr v texte. Korpusový manažér musí byť súčasne schopný počítať na vybranej časti korpusu rôzne štatistiky. Najlepšie je, ak korpusový manažér ponúka príjemné používateľské rozhranie s možnosťou interaktívneho zadávania dotazov a ponúka aj služby, ako je autorizovaný prístup k vybraným častiam korpusu a pod.

Reálnu nepoužiteľnosť štandardných nástrojov na vyhľadávanie v texte ilustruje nasledujúci príklad. Vyhľadávanie v ad hoc utvorenom testovacom súbore ekvivalentnom 200 miliónom neanotovaných slov na súčasnom serveri SNK (1 GHz CPU, 256 MB RAM, 40 GB IDE HDD) trvá o niečo viac ako 1 minútu. Pri reálnej práci s korpusom (hľadanie zložitých textových konštrukcií a ich selekcia podľa rôznych kritérií, hľadanie podľa anotácií a pod.) je tento čas niekoľkonásobne dlhší.

Na prvý pohľad by sa zdalo, že ideálnym nástrojom na implementáciu korpusového manažéra môže byť databázový systém. Ukazuje sa však, že takéto systémy sú schopné uspokojivo pracovať s korpusmi len do určitých veľkostí a len pri nízkych nárokoch na informačnú hodnotu vyhľadávania. Po prevýšení istej veľkosti a pri potrebe vyhľadávať kombinácie rôznych prvkov vnútornej a vonkajšej anotácie začínajú kolabovať. Súvisí to so špecifickou distribúciou dát v textových korpusoch (málo slov je veľmi frekventovaných a veľmi veľa slov je frekventovaných veľmi málo).

Ukazuje sa, že na taký špecifický problém, ako je práca s korpusovými dátami, nie je možné použiť žiadny štandardný nástroj určený na prístup k rozsiahlym dátam. Zatiaľ jedinou cestou je aplikácia softvéru optimalizovaného na prácu s textovým korpusom.