Slovensko-čínsky paralelný korpus

Ľuboš Gajdoš

Slovak-Chinese Parallel Corpus

The aim of the paper is to describe a preliminary work on building a parallel Slovak-Chinese corpus, a motivation for building it, its parameters, goals and future perspective. The lack of appropriate materials (e.g. dictionaries, a grammar which reflects differences of these two typologically and genetically distinct languages) has been the primal impetus for this project. Being based on texts used in translation courses, in the first stage (3–5 years), the corpus is meant to be a didactic tool for students in the Department of East Asian Studies, Faculty of Philosophy, Comenius University. In the second stage, the corpus should be opened to corpus-based linguistic research (e.g. stratification of Chinese language etc.) and its practical application (e.g. lexicography). Currently (spring 2013) the frame of the corpus is already created, the parametrization of the corpus is still underway, as well as the questions of morphological (part of speech tagging), style-genre annotations etc.

Aj vďaka dostupnosti informačných technológií prináša korpusová jazykoveda za ostatné približne dve dekády so sebou veľké možnosti nielen v oblasti lingvistického výskumu jazyka a jeho praktickej aplikácie (napr. v lexikografii), ale korpus sa veľmi často stáva aj didaktickou, v niektorých prípadoch dokonca veľmi potrebnou, pomôckou pri osvojovaní si cudzieho jazyka.[1] Hoci skúmanie

Príspevok vznikol v rámci riešenia grantu VEGA 1/0253/13 2013–2014.

1 Podrobnejšie pozri napr. Karin Aijmer ed., *Corpora and Language Teaching* [Korpusy a výučba jazyka] (Amsterdam: John Benjamins Publishing Company, 2009), alebo *Corpus Research from*

metódami korpusovej lingvistiky (čínsky *yǔliàokù yǔyánxué* 語料庫語言學) v Číne začína neskôr ako v iných krajinách,[2] je možné v súčasnosti pozorovať zvýšený záujem nielen lingvistov o tento zaujímavý spôsob skúmania jazyka. V Číne (na Taiwane) i v zahraničí existuje k dnešnému dňu niekoľko korpusov čínskeho jazyka voľne dostupných cez webové rozhranie.[3] Uveďme niektoré:

- *Běijīng dàxué Hànyǔ yǔyánxué yánjiū zhōngxīn* 北京大學漢語語言學研究中心 [Centrum čínskej lingvistiky Pekinskej univerzity], <ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=xiandai> (cit. 14. mája 2013). Rozsah korpusu je 477 miliónov čínskych znakov (vrátene klasickej čínštiny), podiel textov v súčasnom jazyku je 300 miliónov tokenov. V korpuse je možné vyhľadávať len konkordancie (kľúčové slová), nie je dostupná slovnodruhová anotácia.
- *Centre for Translation Studies*, University of Leeds. Dostupné na internete <corpus.leeds.ac.uk/query-zh.html> (cit. 20. mája 2013). Korpus má rozsah 280 miliónov tokenov, je automaticky morfologicky anotovaný (slovnodruhová anotácia), je v ňom možné vyhľadávať konkordancie a kolokácie.

Ako sme už načrtli, korpus je možné využiť aj ako didaktickú pomôcku, navyše, berúc do úvahy *status quo* v oblasti výučby čínskeho jazyka so zameraním na potreby slovenských hovoriacich pri osvojovaní si čínskeho jazyka, ukazuje sa budovanie slovensko-čínskeho paralelného korpusu ako jedno z potenciálnych riešení. Absencia relevantnej odbornej literatúry reflektujúcej uvedené špecifické požiadavky totiž stavia často pedagóga do pozície »hľadača« *ad hoc* riešení.[4]

Phrase to Discourse [Korpusový výskum od frazeologizmu k výpovedi], ed. Eileen Fitzpatrick (Amsterdam: Rodopi, 2007).

2 Porovnaj Liu Kaiying 劉開瑛, *Zhōngwén wénběn zìdòng fēncí hé biāozhǔn* 中文文本自動分詞和標準 [Automatická segmentácia na slová a jej kritériá] (Beijing : Shangwu yinshuguan, 2000), 171.

3 Okrem korpusov sú dostupné napríklad aj slovníky založené na korpuse. Pozri: *Deutscher Wortschatz,* Universität Leipzig. Dostupné online <corpora.informatik.uni-leipzig.de/?dict=zh> (cit. 3. mája 2013).

4 Tu máme predovšetkým na mysli slovensko-čínsky slovník, gramatiku čínskeho jazyka, ktorá by uspokojivo zodpovedala na otázky pri kontrastívnom pohľade na tieto dva typologicky (geneticky) odlišné jazyky. Na druhej strane je potrebné dodať, že (paralelný) jazykový korpus nie je univerzálnym riešením spomenutých problémov, ale v mnohých ohľadoch je ho možné využiť aspoň ako zdroj jazykových dát.

Počiatky

Na samom počiatku myšlienky vybudovania slovensko-čínskeho paralelného korpusu bol autor tohto článku, ktorému sa podarilo pre spoluprácu získať prof. Janu Benickú.

V súvislosti s prípravnými prácami pri budovaní korpusu by sme chceli vyjadriť naše poďakovanie prof. Pavlovi Žigovi za cenné rady a pripomienky, ktorými nás nasmeroval na Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied, oddelenie Slovenského národného korpusu. Rovnako tak je potrebné na tomto mieste poďakovať RNDr. Radovanovi Garabíkovi a PhDr. Márii Šimkovej, ktorí nám pomohli v počiatočnej orientácii v problematike budovania korpusu. RNDr. Garabík sa okrem toho podujal realizovať budovanie paralelného korpusu po technickej stránke. Jeho skúsenosti s budovaním iných paralelných korpusov sú neoceniteľným vkladom do celkovej práce.

Ciele

Ako už bolo spomenuté, primárnym podnetom budovania korpusu bolo riešenie otázok didaktického charakteru, ktoré úzko súvisia s vyučovaním čínskeho jazyka na našom akademickom pracovisku. V prvej fáze je preto rozhodujúce práve toto hľadisko. Cieľom je teda vytvoriť paralelný korpus, v ktorom si používateľ—študent môže vyhľadať kľúčové slovo (konkordancie) v kontexte jedného jazyka (napr. v čínštine) s tým, že navyše bude mať možnosť vidieť preklad celého kontextu kľúčového slova v druhom jazyku (v tomto prípade v slovenčine). To znamená, že kontext kľúčového slova v čínskom jazyku slúži ako pomôcka pri jeho osvojovaní si, okrem toho, slovenský preklad pomáha používateľovi pri hľadaní slovenského ekvivalentu, keďže v súčasnosti (jar 2013) nie je na knižnom trhu slovensko-čínsky slovník. Vyhľadávať bude možné tak v čínskej, ako aj v slovenskej strane korpusu. Ukážka rozhrania a výsledok vyhľadávania kľúčového slova:[5]

5 Texty použité v skúšobnej verzii paralelného korpusu neprešli jazykovou úpravou.

Parametre korpusu

Paralelný korpus v prvej fáze (3–5 rokov) je mienený ako špecializovaný, rozsahom malý (do 500 tisíc tokenov), obsahujúci písané texty. Korpus bude dostupný cez webové rozhranie len pre študentov, príp. aktívnych prispievateľov a to len v rámci Filozofickej fakulty UK. Čínska strana korpusu bude morfologicky (slovnodruhovo) a bibliograficky (štýlovo-žánrovo) anotovaná, doplnená o prepis do transkripčnej abecedy Hanyu pinyin (s vyznačením tónov).

Realizácia

V súčasnosti (jar 2013) je možné povedať, že kostra systému paralelného korpusu je už vybudovaná a v štádiu riešenia sú otázky parametrizácie korpusu. Hlavne sú to otázky spojené so slovnodruhovou a bibliografickou anotáciou, prepisom čínskych slov do transkripčnej abecedy, tokenizáciou čínskeho textu a pod. Pri slovnodruhovej anotácii sú použité automatizačné nástroje s tým, že chyby pri značkovaní budú ručne opravované.

V prvej fáze bude korpus pravidelne dopĺňaný o texty v čínskom jazyku a ich preklad do slovenského jazyka, ktoré sú obsahom výučby niektorých kurzov na

našom pracovisku. Spomeňme aspoň niektoré z nich: Základy jazyka tlače a masmédií, Čítanie novinových textov, Stredoveká čínština atď. Už z tohto krátkeho prehľadu je zrejmé, že korpus nebude obsahovať len texty súčasnej čínštiny, ale budú v ňom aj texty stredovekej, príp. klasickej čínštiny, primárne sa však bude jednať o synchrónny korpus, pričom diachrónna časť bude len doplnkom. Zo súčasnej čínštiny to budú texty prevažne v publicistickom štýle, príp. umelecké texty. To teda znamená, že prioritou nie je textami pokryť všetky domény jazyka, ale len už spomenuté.

Texty budú pre korpusové použitie (napr. vyčistenie textu, paralelné zarovnanie viet atď.) upravované v rámci pripravovaných grantových schém za aktívnej účasti študentov magisterského stupňa.

V druhej fáze by sa mal paralelný korpus rozširovať aj o texty z iných štýlov súčasného čínskeho jazyka, najmä o hovorový štýl a neskôr aj o hovorený jazyk. Túto neskromnú ambíciu, pravda, nie je možné realizovať v rámci výučby (alebo len v obmedzenej miere), ale do budúcnosti bude potrebné pripraviť projekty, ktoré budú zamerané na výskum čínskeho jazyka, napr. v oblasti stratifikácie čínskeho jazyka. Okrem toho je plánované rozširovať korpus aj o vydané preklady knižných publikácií v čínskom alebo slovenskom jazyku, príp. o publikácie, ktoré sú preložené z tretieho jazyka.

Perspektíva

Z uvedeného je zrejmé, že ťažiskom budovania paralelného korpusu je jeho praktické využitie vo vyučovacom procese a veríme, že sa nám prostredníctvom pripravovaného paralelného čínsko-slovenského korpusu podarí poskytnúť študentom na našom pracovisku silný didaktický nástroj.

Na druhej strane sa však tiež domnievame, že by bolo škodou v strednodobom horizonte (pri dostatočnom rozsahu korpusu) nevyužiť aj iné príležitosti, ktoré korpus (korpusová jazykoveda) ponúka. Tu máme predovšetkým na mysli perspektívu lingvistického skúmanie jazyka, napr. v oblasti syntaxe alebo stratifikácie čínskeho jazyka, ale rovnako tak dúfame, že vyplníme biele miesta na knižnom trhu (napríklad formou čínsko-slovenského slovníka) a dáme tak aj bežnému záujemcovi o poznanie tohto, akiste, zaujímavého jazyka možnosť hlbšie preniknúť do jeho tajov.

Univerzita Komenského v Bratislave, Katedra východoázijských štúdií