Francúzsko-slovenský paralelný korpus

Dorota Vasilišinová – Radovan Garabík

Jazykovedný ústav Ľ. Štúra SAV
813 64 Bratislava, Slovakia
korpus@korpus.juls.savba.sk, http://korpus.juls.savba.sk

XVI. kolokvium mladých jazykovedcov (Častá-Papiernička 8. – 10. 11. 2006)

Abstrait. Le corpus parallèle FRASK est un corpus composé de textes littéraires français accompagnés de leur traduction en slovaque. Les textes parallèles ont été alignés et annotés aux niveaux structurel (paragraphes) et linguistique (étiquetage de la partie du discours et lemmatisation). Cet article présente le procès de la formation du corpus parallèle français-slovaque et les problèmes qui ont survenus au cours de ce travail.

Úvod

Paralelné korpusy sú špeciálnou oblasťou korpusovej lingvistiky (Rosen, 2005). Paralelný korpus tvorí zväčša originálny text v cudzom jazyku a jeho preklad do iného jazyka alebo jazykov (v našom prípade do slovenčiny) v elektronickej podobe. Počítačovo spracované jazykové dáta je možné ďalej využiť napr. na:

- porovnanie lexikálnej, morfologickej a syntaktickej štruktúry jazykov
- skúmanie problémov pri preklade textov
- vytvorenie dvoj- a viacjazyčných slovníkov
- výučbu cudzích jazykov
- trénovanie systémov automatického prekladu
- aplikáciu štatistických metód a pod.

S projektom budovania paralelných korpusov sa v oddelení Slovenského národného korpusu (SNK) Jazykovedného ústavu Ľ. Štúra SAV začalo v roku 2005. V súčasnosti je k dispozícii rusko-slovenský a francúzsko-slovenský paralelný korpus, plánuje sa vytvorenie česko-slovenského paralelného korpusu.

Formát a spracovanie textov

Texty vchádzajúce do korpusu sú spracovávané v niekoľkých fázach, pričom sa vždy aplikuje

jeden konkrétny spôsob konverzie a spracovania predchádzajúcej úrovne. Takýto modulárny prístup umožňuje ľahko zmeniť spracovanie textov v tej-ktorej fáze, napríklad pri náhrade jedného nástroja iným, lepším. Najprv sa texty skonvertujú z pôvodného vstupného formátu (ktorým môže byť napríklad HTML, MS Word alebo iný formát bežne používaný v DTP) na spoločný textový formát v UTF-8 kódovaní. Odseky sú označené prázdnym riadkom, inak sa v texte nenachádzajú žiadne iné formátovacie alebo štruktúrne značky, čo umožňuje jednoduché ručné porovnanie dvoch paralelných textov a ich úpravu – je potrebné ručne zarovnať začiatok a koniec textu, pretože práve tu sa najviac vyskytujú prípadné rozdiely (napríklad doslov prekladateľa, predhovor, iná štruktúra úvodného nadpisu a pod.). V ďalšej fáze sa text prevedie na formát TEI XML, ktorý používajú ako svoj vstupný formát nástroje na segmentáciu, morfologickú analýzu a značkovanie. Po morfologickej analýze je vytvorený ďalší špeciálny formát, v ktorom je každá veta na samostatnom riadku (namiesto slov sú vo vetách už iba lemy) a odseky sú oddelené špeciálnym znakom ¶, čo umožňuje jednoduché ďalšie spracovanie vo fáze automatického zarovnávania. Po zarovnaní sú výsledky zarovnania doplnené do TEI XML súborov ako odkazy na protistojaci súbor v tvare napr. <s link="31+32+33">, čo znamená, že tejto vete zodpovedajú vety s poradovým číslom 31, 32, 33 v druhom súbore. Text je potom skonvertovaný do tvaru vhodného pre korpusový manažér.

FRASK – francúzsko-slovenský paralelný korpus

Francúzsko-slovenský paralelný korpus FRASK (http://korpus.juls.savba.sk/frask/) je vytvorený z prozaických textov francúzskych autorov a ich prekladov do slovenčiny, ale postupne predpokladáme jeho doplnenie aj inými typmi textov (odbornými, publicistickými). Paralelný korpus FRASK má v súčasnosti vo francúzskej časti 315 599 tokenov, 13 004 viet a v slovenskej časti 194 478 tokenov, 12 286 viet. Tento nepomer medzi počtom tokenov a viet vo francúzskych a slovenských textoch je spôsobený pravdepodobne rozdielnou syntaktickou a morfologickou stavbou oboch jazykov (napr. vo francúzštine používanie členov pri substantívach, konjugácia slovies atď.).

Lingvistická analýza textu

Text je segmentovaný na vety použitím jednoduchého heuristického algoritmu podľa prítomnosti interpunkčných znamienok na konci vety. Text je ďalej lematizovaný a morfologicky analyzovaný; v slovenskej časti používame morfologický analyzátor popísaný v

Levenshtein Edit Operations as a Base for a Morphology Analyzer (Garabík, 2005) v kombinácii s TNT dezambiguátorom (Brants). Systém gramatických značiek je popísaný v manuáli pre morfologickú anotáciu SNK (Garabík – Gianitsová – Horák – Šimková, 2004).
V čase písania tohto článku nie je ešte aplikovaná lematizácia vo francúzskej časti textov, predpokladáme však použitie francúzskeho nástroja Flemm v3.1 (Analyseur Flexionnel du français pour des corpus étiquetés), ktorý vypracovalo laboratórium ATILF (Analyse et Traitement Informatique de la Langue Française) v Nancy vo Francúzsku.
Každému textu je priradená aj štýlovo-žánrová a bibliografická anotácia, ktorá obsahuje informácie o type textu, žánri a doméne, teda odbornej oblasti, ktorej sa daný text týka, ďalej bibliografické údaje a iné relevantné vlastnosti textu.

Zarovnávanie

Na zarovnávanie textov bol použitý program hunalign (zdroj http://mokk.bme.hu/resources/hunalign). Zarovnávanie prebieha na úrovni viet pomocou paragrafov, dĺžky viet, vstupného slovníka a automatizovaného slovníka. Pri procese zarovnávania si program vytvorí vlastný slovník na základe paralelných textov. Tento slovník sa na zefektívnenie procesu zarovnávania ručne skontroloval a upravil, pri slovesných tvaroch sa vo francúzštine použil infinitív, vlastné mená a ich ekvivalenty (ako Maniflore – Ejkvietková, Cornemuse – Gajdík) sa odstránili, hoci sú pre prekladateľa zaujímavé, ale pri ďalšej expanzii paralelného korpusu by mohli spôsobovať problémy zanášaním „šumu“ do slovníka. Problematickými sa ukázali aj slová elidované, ako je napr. *aujord'hui*, keď sa vo vstupnom slovníku dané slovo vyskytlo v podobe *dnes – aujourd* a *dnes – hui* (apostrofy sa v texte automaticky vydelili medzerami pri tokenizácii). Slovník sa z tohto dôvodu musel očistiť aj od slov, v ktorých je apostrof.

Úspešnosť zarovnania

Na zistenie úspešnosti zarovnania sme štatisticky porovnali výskyt slovenského slova ALEBO a jeho paralelné francúzske ekvivalenty v korpuse pri použití dvoch typov slovníka. Po zarovnaní textu za pomoci automatizovaného slovníka bolo pri 144 výskytoch hľadaného slova:

- 86 úplne totožných viet (60 %)
- 29 viet (20 %) významovo správne preložených, pri ktorých sa však nezhodovala dĺžka vety (skrátené, pospájané), jednej slovenskej vete boli priradené aj 2 – 3 francúzske alebo naopak

– 29 nezhodných viet (20 %).

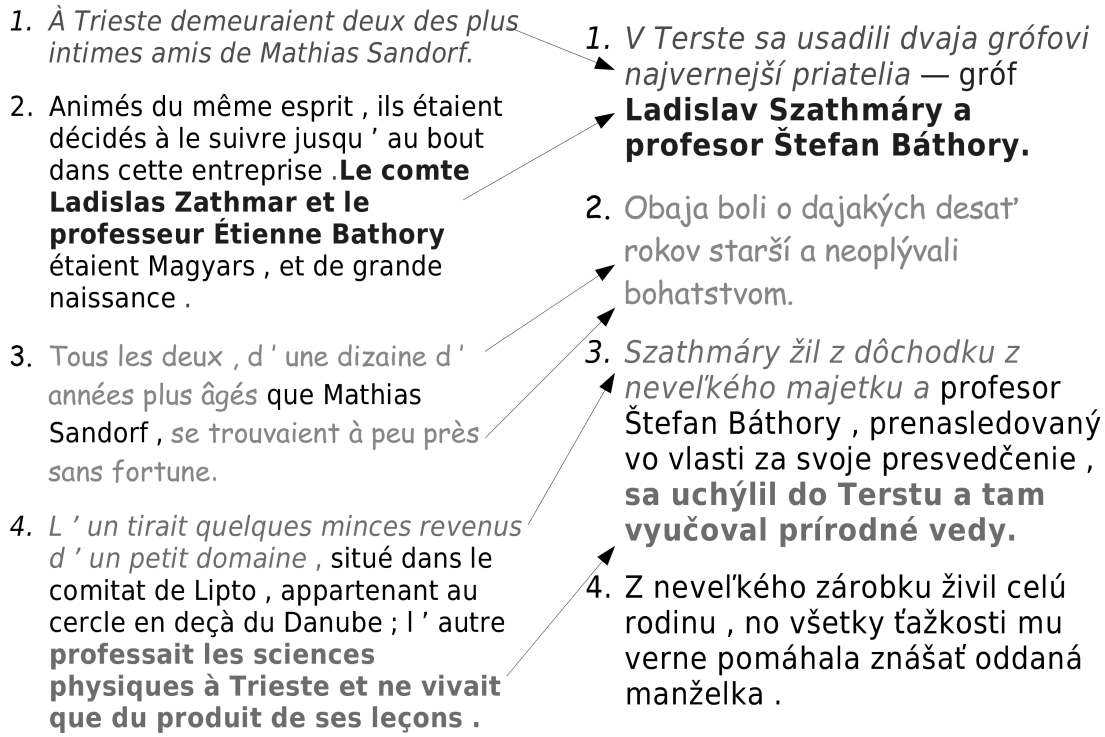
Následne bol automatizovaný slovník ručne opravený a texty v paralelnom korpuse sa znova zarovnali. Výsledky sú nasledujúce:

– 91 úplne totožných viet (63 %)
– 30 nie úplne zhodných viet (21 %)
– 23 nezhodných viet (16 %).

Ručnou opravou slovníka sa podarilo odstrániť 20 % chýb, teda úspešnosť zarovnania sa zlepšila. Po pridaní ďalších textov do paralelného korpusu a vygenerovaní nového slovníka ho však bude potrebné znova ručne opraviť.

Problematické momenty pri spracovaní a používaní korpusu

Najproblematickejším bodom pri zarovnávaní viet v paralelnom korpuse sa ukázali nepresnosti alebo chyby prekladateľa textu. Veľmi často sa stáva, že prekladateľ buď zo štylistických, alebo iných dôvodov skráti alebo predĺži dĺžku viet, vynechá niektoré slová alebo časti textu. Ako príklad uvedieme nasledujúce vety z románu od Julesa Verna – Nový gróf Monte Christo:



Ako naznačujú čísla viet a šípky, v tomto odseku došlo vzhľadom na zásahy prekladateľa k niektorým posunom v slovenskej časti textu. Prvá francúzska veta bola v slovenskom preklade doplnená menami, ktoré figurovali až v tretej vete. Druhá francúzska veta bola v preklade vynechaná úplne. Slovenský preklad v druhej vete pokračuje v poradí už štvrtou francúzskou vetou, v ktorej však boli niektoré slová vynechané a piata veta francúzskeho textu sa v slovenskom preklade nachádza na treťom mieste, pričom je doplnená o časť slovenského textu, ktorý však v origináli nenájdeme. Výsledkom tejto štylizácie je päť francúzskych viet ekvivalentných trom slovenským, pričom rozdiel nie je len v dĺžke a počte viet, ale aj v neekvivalentných častiach textu. V takomto prípade je priam nemožné dosiahnuť správne zarovnanie textu.

Ďalšou prekážkou sa ukázali niektoré znaky a interpunkcia. Francúzska graféma *œ* sa vo francúzsky písaných textoch vyskytuje v oboch grafických podobách, teda ako *oe* aj *œ*, v tomto paralelnom korpuse je napr. v diele Anatola Franca zapísaná ako *oe*, v texte od Julesa Verna ako *œ*. Z toho dôvodu sme sa rozhodli ponechať v textoch oba používané znaky a používateľov korpusu upozorniť na to, že pri vyhľadávaní by mali rátať s oboma alternatívami.

Nekoherentný je tiež spôsob zápisu úvodzoviek, ktoré sa líšia v textoch francúzskych a slovenských, na zjednotenie bolo potrebné niektoré znaky buď úplne odstrániť, alebo aspoň prepísať do inej formy tak, aby boli zhodné.

Vyhľadávanie v korpuse

Na vyhľadávanie v korpuse sa využíva korpusový manažér Manatee, vyhľadáva sa pomocou vlastného frontendu cez WWW rozhranie napísané v systéme Karrigell. Toto www rozhranie je prístupné v slovenčine, angličtine, francúzštine a krymskotatárčine. Vyhľadávať v korpuse je možné pomocou slova, lemy, regulárnych výrazov a gramatických značiek. Pre francúzske dokumenty je vyhľadávanie momentálne obmedzené na slovo a regulárny výraz.

Záver

Popísaný paralelný korpus môže slúžiť ako cenná pomôcka pre prekladateľov, pri výučbe jazyka (francúzštiny alebo slovenčiny), ako zdroj trénovacích dát pre systémy automatického prekladu, prípadne ako pomôcka pri tvorbe dvojjazyčného slovníka. Po rozšírení korpusu o odborné, publicistické a iné texty predpokladáme jeho využitie aj v širšej oblasti francúzsko-slovenských jazykových vzťahov.

Literatúra

BRANTS, T.: TNT – Statistical Part-of-Speech Tagging. http://www.coli.uni-saarland.de/~thorsten/tnt/

GARABÍK, Radovan: Levenshtein Edit Operations as a Base for a Morphology Analyzer. In: Computer Treatment of Slavic and East European Languages. Ed. R. Garabík. Bratislava: Veda 2005, s. 50-58.

GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária (2004): Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. Interný materiál. http://korpus.juls.savba.sk/publications

ROSEN, Alexandr: In Search of the Best Method for Sentence Alignment in Parallel Texts. In: Computer Treatment of Slavic and East European Languages. Ed. R. Garabík. Bratislava: Veda 2005, s. 174-185.