Dvojjazyčné prekladové ekvivalenty z paralelného slovensko-francúzskeho korpusu

Radovan GARABÍK

Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied, Bratislava

Résumé : Nous présentons une base de données de type dictionnaire d'équivalents français-slovaque extraits d'un corpus parallèle slovaque-français aligné sur les phrases et les mots. Dans cet article nous abordons la structure de base du corpus, composé d'un sous-corpus de textes de fiction et d'un sous-corpus de textes administratifs et techniques et la méthode d'extraction et d'élagage des équivalents de traduction bilingues. La base de données est accessible sur le portail des dictionnaires de l'Institut Ľ. Štúr (Institute of Linguistics) et offre de riches possibilités d'interrogation et de visualisation des équivalents de traduction courts alignés sur les mots.

Mots-clés : base de données, corpus, traduction français-slovaque

1 Úvod

Slovensko-francúzsky[1] paralelný korpus sa skladá z dvoch hlavných častí, ktoré nazývame podkorpus beletrie a podkorpus voľne dostupných textov. Podkorpus beletrie neobsahuje iba beletriu, ale aj nejakú literatúru faktu. Podkorpus voľne dostupných textov je tvorený prevažne textami, ktoré sú voľne dostupné na Internete – ide väčšinou o texty a preklady vznikajúce v inštitúciách Európskej únie. Korpus je podrobnejšie popísaný v (Vasilišinová, Garabík 2007), hoci nová verzia, z ktorej vychádza tento článok, je už podstatne rozšírená a vylepšená.

V tabuľkách 1 a 2 je uvedený stručný prehľad veľkosti podkorpusov a v tabuľke 3 štruktúra podkorpusu voľne dostupných textov – hoci podkorpus beletrie patrí svojim rozsahom medzi korpusy menšie, očakávame, že kvalita a rozmanitosť textov tohto

[1] V názve ide o konvenciu – korpus je formálne obojsmerný, obsahujúci preklady z francúzštiny do slovenčiny, zo slovenčiny do francúzštiny a preklady z tretích jazykov do oboch, francúzštiny aj slovenčiny. V skutočnosti podkorpus beletrie obsahuje výlučne preklady z francúzštiny do slovenčiny.

typu výrazne prispeje ku kvalite výsledného paralelného korpusu. Podľa očakávania je rozsah francúzskych textov (v tokenoch) konzistentne väčší než slovenských, čo súvisí hlavne s prítomnosťou krátkych ortograficky oddelených funkčných slov, ako sú hlavne členy a osobné zámená vo francúzštine.

	slovenská časť	francúzska časť
tokeny	4 318 156	5 547 763
slová	3 577 746	4 756 714
vety	289 515	280 120
dokumenty	58	

Tabuľka 1: Veľkosť podkorpusu beletrie

	slovenská časť	francúzska časť
tokeny	212 273 451	226 903 647
slová	141 342 890	164 347 551
vety	14 024 793	13 232 351
dokumenty	200 092	

Tabuľka 2: Veľkosť podkorpusu voľne dostupných textov

zdroj	slovenská časť	francúzska časť
DCEP: Digitálny korpus Európskeho parlamentu	10 379 400	11 422 930
DGT-Acquis	14 400 673	16 578 959
Európska centrálna banka	414 270	548 741
Európske centrum pre prevenciu a kontrolu chorôb	38 449	48 094
Úradný vestník Európskej únie	187 040 659	198 304 923

Tabuľka 3: Štruktúra podkorpusu voľne dostupných textov podľa zdrojov, veľkosti v tokenoch

Cieľom našej práce je využiť tento paralelný korpus na získanie dvojjazyčných krátkych ekvivalentov (v terminológii systému strojového prekladu MOSES nazývané „frázy“, pozri (Koehn., Och, et Marcu 2003); v našom článku budeme používať termín „prekladové ekvivalenty“), ktoré by mohli dopĺňať dvojjazyčné slovníky, prípadne zahŕňať aj slová a kontexty, ktoré sa v dvojjazyčných slovníkoch bežne nespracovávajú. Prekladové ekvivalenty z prekladov vytvorených profesionálnymi prekladateľmi odrážajú reálne vzťahy medzi používaním slov v oboch jazykoch v ekvivalentných konštrukciách, hoci samozrejme niekedy sa nevyhneme chybám a nejednoznačnostiam[2].

2 Spracovanie korpusu

2.1 Normalizácia textu

Pri spracovaní francúzskych textov narážame na drobný, ale podstatný detail – mnohé z nich pochádzajú spred éry masového rozšírenia Unicode a UTF-8 kódovania,

[2]Notoricky známy je nesprávny preklad francúzskeho *mauvaises terres* ako *nezdravé končiny*.

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

a často sú v kódovaní ISO8859-1 alebo Windows-1252. Kódovanie ISO8859-1 neobsahuje francúzske písmená *Ÿ* (vyskytujúce sa vo francúzštine veľmi zriedkavo) a *œ* (vyskytujúce sa pomerne často). Hoci Windows-1252 pridáva obe tieto písmená, texty označené (napr. v kontexte http protokolu) týmto kódovaním často obsahujú len znaky z podmnožiny ISO8859-1. *œ* býva v týchto textoch nahradzované dvojicou znakov *o* a *e*. Z 58 dokumentov v podkorpuse beletrie obsahuje ligatúru *œ* dokumentov 39; ďalších 19 dokumentov používa prepis dvojicou písmen *oe*. Z týchto dôvodov sme ligatúru *œ* oželeli a aj v tých zdrojoch, v ktorých sa vyskytuje, ju nahrádzame dvojicou písmen *oe*.

Francúzske texty sú lematizované a morfologicky označkované systémom TreeTagger (Schmidt 2003; Stein 2003); slovenské sú lematizované a označkované systémom MorphoDita (Straková et al. 2014), natrérovaným na slovenskom ručne lematizovanom korpuse a morfologickom tagsete (Garabík, Šimková 2012).

2.2 Filtrovanie korpusu

V zarovnanom korpuse sme v prvom kroku ponechali len tie vety, ktoré sú dlhé maximálne 40 tokenov, aby sme predišli príliš komplikovaný vetným štruktúram, v ktorým sa problematický určuje ekvivalencia slov medzi jazykmi. Príliš dlhé vety sú často aj výsledkom chybného spracovania textu. Taktiež sme ponechali len tie vety, pri ktorých bola medzijazyková korelácia 1:1 – t. j. jednej slovenskej vete zodpovedala práve jedna francúzska veta (a naopak).

Po tejto filtrácii zostalo v podkorpuse beletrie 239 490 viet (rovnaké množstvo v oboch jazykoch), z pôvodných 289 515 viet zo slovenskej časti (t. j. 82.75 % pôvodnej veľkosti) a 280 120 z francúzskej (85.5 % pôvodnej veľkosti); pre pomery vyjadrené v tokenoch, v slovenskej časti zostalo 77.4 % pôvodného množstva tokenov, vo francúzskej 74.7 %. V podkorpuse voľne dostupných textov zostalo 10 006 807 viet (rovnaké množstvo v oboch jazykoch), z pôvodných 14 024 793 viet v slovenskej časti (71.4 % pôvodnej veľkosti) a 13 232 351 vo francúzskej (75.6 % pôvodnej veľkosti); v tokenoch 45.4 % pôvodnej veľkosti slovenskej časti, 43.5 % francúzskej časti.

2.3 Trénovanie a filtrovanie prekladových ekvivalentov

Na účely vytvorenia databázy fráz používame osobitný trénovaci režim, pri ktorom faktorizujeme tokeny vstupných textov na slovný tvar+lema+slovný druh (takúto faktorizáciu potom využijeme pri indexovaní prekladových ekvivalentov). Zarovnanie na úrovni slov poskytuje softvér GIZA++ (Och & Ney, 2000; Koehn, Och & Marcu, 2003).

Proces trénovania každej dvojici extrahovaných prekladových ekvivalentov priradí nasledovné skóre (*f* znamená zdrojový text, *e* preložený):

a) inverznú pravdepodobnosť prekladu ϕ(f|e)
b) inverznú lexikálnu váhu lex(f|e)
c) priamu pravdepodobnosť prekladu ϕ(e|f)
d) priamu lexikálnu váhu lex(e|f)

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

Jednoduché skóre, ktoré berie do úvahy všetky tieto skóre, ktoré je symetrické voči výmene jazykov a odzrkadluje pravdepodobnostný charakter je

$$g = \phi(f|e) \cdot \phi(e|f) \cdot lex(f|e) \cdot lex(e|f)$$

Keďže výsledné ekvivalenty chceme ponúknuť používateľom vo forme pseudoslovníkového rozhrania, pomerne dôležitou záležitosťou sa stáva výber správnych ekvivalentov. Nejde pritom len o odstránenie nesprávne zarovnaných textov (ktoré dokáže celkom spoľahlivo odstrániť hranica pre lexikálnu pravdepodobnosť), ale aj o taký výber, pri ktorom budú podľa možnosti syntaktický ucelené kúsky textu, hľadané slovo bude v rôznych relevantných kontextoch a výsledok bude okrem formálnej správnosti prekladu pôsobiť aj esteticky ucelene a vyvážene. Nami vybrané kritériá sú nasledovné: výsledný prekladový ekvivalent by mal byť dlhý štyri až šesť slov (neberúc do úvahy interpunkciu), mal by začínať aj končiť písmenom (t. j. nie znakom interpunkcie ani číslicou) a nemal by obsahovať bodku, úvodzovky ani zátvorky. Posledným tokenom prekladového ekvivalentu by nemala byť predložka, spojka ani (vo francúzskej časti) člen. Tieto kritériá boli stanovené empiricky, po preskúmaní vzoriek prekladových ekvivalentov a selekcii najčastejších dôvodov pre nevhodné prekladové ekvivalenty.

V korpuse sa niekedy nachádzajú časti viet, kde je francúzska časť identická so slovenskou, prípadne sa líši iba v drobnostiach (v diakritike alebo interpunkcii). Buď ide o propriá, texty v iných jazykoch než francúzština a slovenčina (obzvlášť v podkorpuse voľne dostupných textov sa niekedy vyskytujú, prevažne omylom – buď anglické texty, ktoré mali byť zrejme preložené, alebo niekedy texty v slovinčine, ktoré zrejme mali byť preložené do francúzštiny, omylom boli klasifikované ako slovenské a prekladatelia ich celkom pochopiteľne nepreložili). V podkorpuse beletrie sa v niektorých dokumentoch vyskytujú transkribované citáty z ruštiny, ktoré ale nie sú identické – je tu použitý voľný fonetický prepis francúzskym alebo slovenským pravopisom. Preto prvý krok filtrovania vhodných prekladových ekvivalentov zahŕňa odstránenie identických a veľmi podobných slovenských a francúzskych ekvivalentov. Za „veľmi podobné“ považujeme také, ktoré po odstránení diakritiky majú rozdielne menej než tri slová (tokeny; rozdiel je braný ako symetrická diferencia množín slov francúzskeho a slovenského ekvivalentu, t. j. bez ohľadu na poradie slov), alebo ktorých Levenštejnova vzdialenosť je menšia než päť znakových operácií.

Celkový počet natrénovaných prekladových ekvivalentov z podkorpusu beletrie je 2 256 440, z podkorpusu voľne dostupných textov 52 370 669. Po odfiltrovaní identických alebo takmer identických ekvivalentov je počet dvojíc z podkorpusu beletrie 1 977 932, z podkorpusu voľne dostupných textov 29 337 581. Po odfiltrovaní ekvivalentov obsahujúcich znaky mimo abecedy daného jazyka, s dĺžkou inou než štyri, päť alebo šesť tokenov, obsahujúcich neželané znaky a končiacich predložkami, spojkami alebo členmi, a s odstránením takých dvojíc ekvivalentov, ktoré majú prekladové skóre $g < 3.3 \cdot 10^{-9}$ dostaneme z podkorpusu beletrie 31 614 prekladových ekvivalentov, z podkorpusu voľne dostupných textov 528 788 ekvivalentov. Aby sme podľa možnosti odstránili texty v iných jazykoch (ako už spomínaná angličtina, slovinčina alebo foneticky prepísaná ruština), testujeme slovenské prekladové ekvivalenty voči

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

jazykovému modelu postavenému na veľkom písanom korpuse slovenčiny[3] (Slovenský národný korpus – prim-6.1-public-sane). Jazykový model je vytvorený s použitím sytému KenLM Language Toolkit (Heafield 2011). Ako diskriminačnú veličinu sme zvolili

$$log\ \mathrm{p} = \mathrm{n}^{-1}\ \cdot\ log\ p(w_n/w_{n-1}w_{n-2}\ldots)$$

kde $w_n/w_{n-1}w_{n-2}\ldots$ je prekladový ekvivalent, n jeho dĺžka v tokenoch a $\mathrm{p}(w_n/w_{n-1}w_{n-2}\ldots)$ pravdepodobnosť prekladového ekvivalentu podľa jazykového modelu – inými slovami, priemerný logaritmus pravdepodobnosti na jeden token textu (čím dosiahneme porovnateľnosť medzi ekvivalentami s rôznou dĺžkou). Empiricky zvolená hodnota pre diskrimináciu „dobrého“ jazyka bola $log\ \mathrm{p} > 10^{-4}$.

Po odfiltrovaní ekvivalentov na základe vhodnosti podľa jazykového modelu je počet ekvivalentov z podkorpusu beletrie 29 202, z podkorpusu voľne dostupných textov 455 170.

Posledným typom filtrácie je odstránenie prílišnej schematickosti a uniformity výsledných ekvivalentov, spôsobených hlavne textami z podkorpusu voľne dostupných textov. V tomto kroku odstránime všetky ďalšie výskyty tých ekvivalentov, v ktorých bolo každé slovo indexované aspoň päťkrát, a neindexujeme slová, ktoré boli už 15-krát indexované. Z výsledných slov indexujeme len tie, ktoré su dlhé aspoň tri znaky. Po filtrácii počet prekladových ekvivalentov je 19 889 z podkorpusu beletrie a 108 216 z podkorpusu voľne dostupných textov.

Pre zaujímavosť uvádame prekladový ekvivalent s najvyšším skóre *g*, francúzsky originál *nfepmqkyuuexto gz gkyuumfv ijdqdpzjq* a zodpovedajúci slovenský preklad[4] *xnuxl rnth xrvt qumcpfoli*, čo je z translatologického hľadiska nesporne zaujímavé, hoci pre slovníkové účely neveľmi relevantné.

2.4 Odhad presnosti a indexovanie slovníka

Presnosť extrahovaných prekladových ekvivalentov ohodnotíme na základe náhodne vybranej vzorky veľkosti 100 prvkov, ktorú sme ručne skontrolovali. Počet nie celkom ekvivalentných prekladov je pomerne malý, takže na tomto mieste ich môžeme všetky zreprodukovať (tabuľky 4 a 5). Z tabuliek vidno, že hoci tieto ekvivalenty nepovažujeme za „správne“, v skutočnosti nejde o výrazné chyby – buď je ekvivalent kratší než mal byť, alebo ide o správne detekovaný preklad, ktorý bol ale voľnejší a nedá sa považovať za ekvivalent (väčšinou sa ale také preklady nedostanú do množiny ekvivalentov s dostatočne vysokou lexikálnou pravdepodobnosťou).

Môžeme teda odhadnúť chybovosť výsledných fráz na 3 % v podkorpuse beletrie a 9 % v podkorpuse voľne dostupných textov; alebo za predpokladu hypergeometrického

[3] Ako ukázalo, tento jazykový model nefiltruje celkom spoľahlivo v podkorpuse beletrie – dôvodom je skutočnosť, že mnohé texty z neho sa nachádzajú aj v korpusoch prim- a teda úryvky textu z nich majú vysokú pravdepodobnosť, bez ohľadu na ich jazyk. Avšak v podkorpuse voľne dostupných textov je táto filtrácia pomerne spoľahlivá a účinná (hoci samozrejme nie stopercentná).

[4] Z románu « Verne, Jules: La Jangada – Huit cents lieues sur l'Amazone. 1881 ». Slovenský preklad: « Verne, Jules: 800 míľ po Amazone. Bratislava: Mladé letá 1988. Preklad: Dominik Jarábek ». Ide o prekladový ekvivalent, ktorý nikde inde nie je preložený inými slovami. Preto mu trénovací algoritmus priradí (správne) stopercentnú pravdepodobnosť.

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

rozdelenia pravdepodobnosti ekvivalentov vo vzorke a jeho aproximáciou rozdelením binomickým bude 95%-ný interval spoľahlivosti chybovosti v rozmedzí 0.7 % až 7.9 % v podkorpuse beletrie a 4.5 % až 15.9 % v podkorpuse voľne dostupných textov (v prípade Clopperovho-Pearsonovho odhadu intervalu spoľahlivosti).

slovenský ekvivalent	francúzsky ekvivalent
nástrojov ceremoniálu a dávam	instruments du cérémonial et
hĺbke ukazoval dvanásť stupňov	marqua douze degrés centigrades
ju nepobozkať na ústa	l' embrasser sur la bouche

Tabuľka 4: Nie celkom správne ekvivalenty vo vzorke náhodne vybraných 100 prvkov, podkorpus beletrie

slovenský ekvivalent	francúzsky ekvivalent
v ohniskovej rovine obrazu	dans un plan focal image de
frakcie hydrogenovaného ricínového olej	fractions d' huiles de ricin
za tepla alebo ryby gravad	chaud ou gravad lax
prídavky a ďalšie príspevky	allocations familiales et indemnités
byvolieho mlieka , v nádobách	de bufflonne , en récipients
trhom a musí byť	marché commun et doit être
sulfanylamid a jeho deriváty	para-amino benzène sulfonamide et ses dérivé ?
a reakcia strany spotreby	et effacements de consommation ?
parciálnej soli , ktorá je	sel partiel de thermoplastique qui est

Tabuľka 5: Nie celkom správne ekvivalenty vo vzorke náhodne vybraných 100 prvkov, podkorpus voľne dostupných textov

V tabuľkách 6 a 7 je počet indexovaných lem a slov z oboch korpusov; slová a lemy môžu byť samozrejme indexované veľakrát, ak sa vyskytujú vo viacerých ekvivalentoch, v stĺpcoch *rôzne lemy* a *rôzne slová* je zachytený ich unikátny výskyt (počet indexovaných slov a lem bez ohľadu na počet indexácií jednotlivého slova alebo lemy). Okrem toho je každé slovo obsahujúce znak *œ* alebo dvojicu písmen *oe* indexované dvomi spôsobmi (s *œ* aj *oe*), bez ohľadu na originálny spôsob zápisu (tieto výskyty nie sú zahrnuté v počtoch uvedených v tabuľkách). Vzhľadom na veľkosť podkorpusu beletrie je lexikálne pokrytie veľmi dobré; na druhej strane, hoci podkorpus voľne dostupných textov je značne väčší, lexikáne pokrytie síce výrazne narástlo, ale menej než by zodpovedalo veľkosti korpusu, čo vyplýva z jeho menšej jazykovej rozmanitosti.

	ekvivalenty	lemy	slová	rôzne lemy	rôzne slová
slovenské	19 889	87 356		7 625	14 327
francúzske		96 245		5 791	9 725

Tabuľka 6: Počet indexovaných lem a slov v slovníku prekladových ekvivalentov vytvorených z korpusu beletrie

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

	ekvivalenty	lemy	slová	rôzne lemy	rôzne slová
slovenské	108 216	469 418		21 756	45 185
francúzske		534 160		17 358	25 798

Tabuľka 7: Počet indexovaných lem a slov v slovníku prekladových ekvivalentov vytvorených z korpusu voľných textov

Keďže charakter textov v korpuse voľne dostupných textov je do značnej miery schematický a jednotvárny, používateľovi je vhodnejšie prezentovať prekladové ekvivalenty z korpusu beletrie, ak také pre dané hľadané slovo existujú. V slovníkovom rozhraní zobrazujeme maximálne osem výskytov (prekladových ekvivalentov) indexovaných kľúčových slov. Týchto osem výskytov vyberáme primárne zo slovníka ekvivalentov podkorpusu beletrie; ak je počet vyhovujúcich ekvivalentov obsahujúcich dané slovo menší než osem, doplníme ich z podkorpusu voľne dostupných textov.

3 (Po)užívateľské rozhranie

Prekladové ekvivalenty sú indexované v slovníkovom systéme založenom na dict protokole (RFC 2229) na serveri, ktorý slúži ako základ pre Webové slovníkové rozhranie Slovníkového portálu Jazykovedného ústavu Ľ. Štúra SAV[5]. Kľúčové (indexované) slova sú tvorené zjednotením slovných tvarov a lem oboch fráz (francúzskej aj slovenskej). Toto zabezpečuje jednoduché a intuitívne vyhľadávanie slov buď v základnom alebo aj konkrétnom tváre v oboch jazykoch, bez nutnosti voľby konkrétneho typu hľadania, čím sa snažíme osloviť okrem odborných užívateľov aj široké vrstvy verejnosti používajúce slovníkové rozhranie. Rozhranie podporuje vyhľadanie celého slova, začiatku slova, konca slova, podreťazca alebo regulárneho výrazu. V rozhraní je možné vyhľadávať slová aj bez diakritiky, diakritika pritom nie je vždy ignorovaná – v prípade zadania slová s diakritikou bude zohľadnená, iba v prípade vyhľadávania bez diakritiky bude ignorovaná. Zarovnanie na úrovni slov je zobrazené, ak používateľ nadíde kurzorom myši nad príslušné slovo, spárovanie pritom zohľadňuje komplikovanejšie korešpondencie medzi slovami (N:M), vrátane rudimentárnej detekcie fixných viacslovných spojení (ktoré sú navzájom zarovnané s ekvivalentným slovným spojením v druhom jazyku). Výsledné zobrazenie rekonštruuje prítomnosť či neprítomnosť medzier medzi tokenami na základe heuristickej analýzy medzier pri rôznych interpunkčných znakoch, aby sa dosiahol „prirodzený“ vzhľad zobrazených prekladových ekvivalentov.

4 Záver

Extrahované prekladové ekvivalenty majú dobré lexikálne pokrytie oboch jazykov, francúzskeho aj slovenského. Ich sprístupnenie vo forme slovníkového rozhrania môže byť dobrou doplňujúcou pomôckou k existujúcim dvojjazyčným, ako aj jednojazyčným výkladovým slovníkom. Zobrazenie hľadaných výrazov v prekladovom kontexte ponúka dodatočnú informáciu, ktorá môže byť užitočnou vo všetkých druhoch pre-

[5] https://slovnik.juls.savba.sk/.

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN: 9782813004260, doi: 10.17184/eac.9782813004260

Frázy z paralelného slovensko-francúzskeho korpusu
jazyk nemôže nič uchopiť ≈ langage ne peut rien saisir
jazyky vysvetlia ti svet ≈ langages t'expliquent le monde
i do racionalistického jazyka ≈ même en langage rationaliste
jednoduchým jazykom a zmenšením ≈ langage simple et simplification
na seba vyplazujú jazyk ≈ se tirent la langue
na seba vždy vyplazujú jazyk ≈ toujours se tirent la langue
odpovedať v jeho jazyku ≈ répondre en sa langue
predmetu v rozličných jazykoch ≈ objet dans divers langages

Tabuľka 8: Príklad vyhľadávania v slovníkovom rozhraní, slovo *jazyk*. Zobrazených je osem prekladových ekvivalentov, v prvom riadku (po nadídení kurzorom myši) vidíme zvýraznené zarovnanie na úrovni slov – slovo *nemôže* je sprárované so slovným spojením *ne peut*.

kladateľských a bilinguálnych slovensko-francúzskych potrieb. Slovníkové rozhranie je dostupné na adrese https://www.juls.savba.sk/skfr.html, ako súčasť slovníkového portálu Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied.

Literatúra

FAITH, R., MARTIN, B. (1997), Request for Comments: 2229, « A Dictionary Server Protocol ». Network Working Group.

GARABÍK, R., DIMITROVA, L. (2015) « Extraction and Presentation of Bilingual Correspondences from Slovak-Bulgarian Parallel Corpus », *Études cognitives*, Varsovie, SOW Publishing House, n° 15, p. 327–334.

GARABÍK, R, ŠIMKOVÁ, M. (2012), « Slovak Morphosyntactic Tagset », *Journal of Language Modelling*, Vol 0, n° 1, p. 41–63.

HEAFIELD, K. (2011), « KenLM: Faster and smaller language model queries », *Proceedings of the sixth workshop on statistical machine translation*. p. 187–197.

KOEHN, P., OCH, F. J., et MARCU, D. (2003), « Statistical phrase-based translation », *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,* Association for Computational Linguistics, p. 48–54.

OCH, F. J., NEY, H. (2000), « Improved statistical alignment models », *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, p. 440–447.

SCHMID, H. (2003), « TreeTagger – a part-of-speech tagger for many languages », http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

Slovenský národný korpus – prim-6.1-public-sane. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2013. Dostupný z WWW: https://korpus.juls.savba.sk.

STEIN, A. (2003), « French TreeTagger Part-of-Speech Tags » (2003). http://www.ims.uni-stuttgart.de/schmid/french-tagset.html.

STRAKOVÁ J., STRAKA M., et HAJIČ, J. (2014) « Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition », *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, p. 13–18.

VASILIŠINOVÁ, D., GARABÍK, R., « Parallel French-Slovak Corpus », *Computer Treatment of Slavic and East European Languages. Proceedings of the Conference Slovko 2007*. Éds. Jana Levická, Radovan Garabík. Brno: Tribun 2007, p. 261–266.

Published in : JAMBOROVA LEMAY, Diana ; PEREIRA, Christophe ; ŠMILAUER, Ivan (dir.), *Des langues calculables à l'homme incalculable. Hommage à Patrice Pognan*, Editions des archives contemporaines, ISBN : 9782813004260, doi : 10.17184/eac.9782813004260