O jednej skratke

Radovan Garabík

JÚĽŠ SAV

813 64 Bratislava, Slovakia

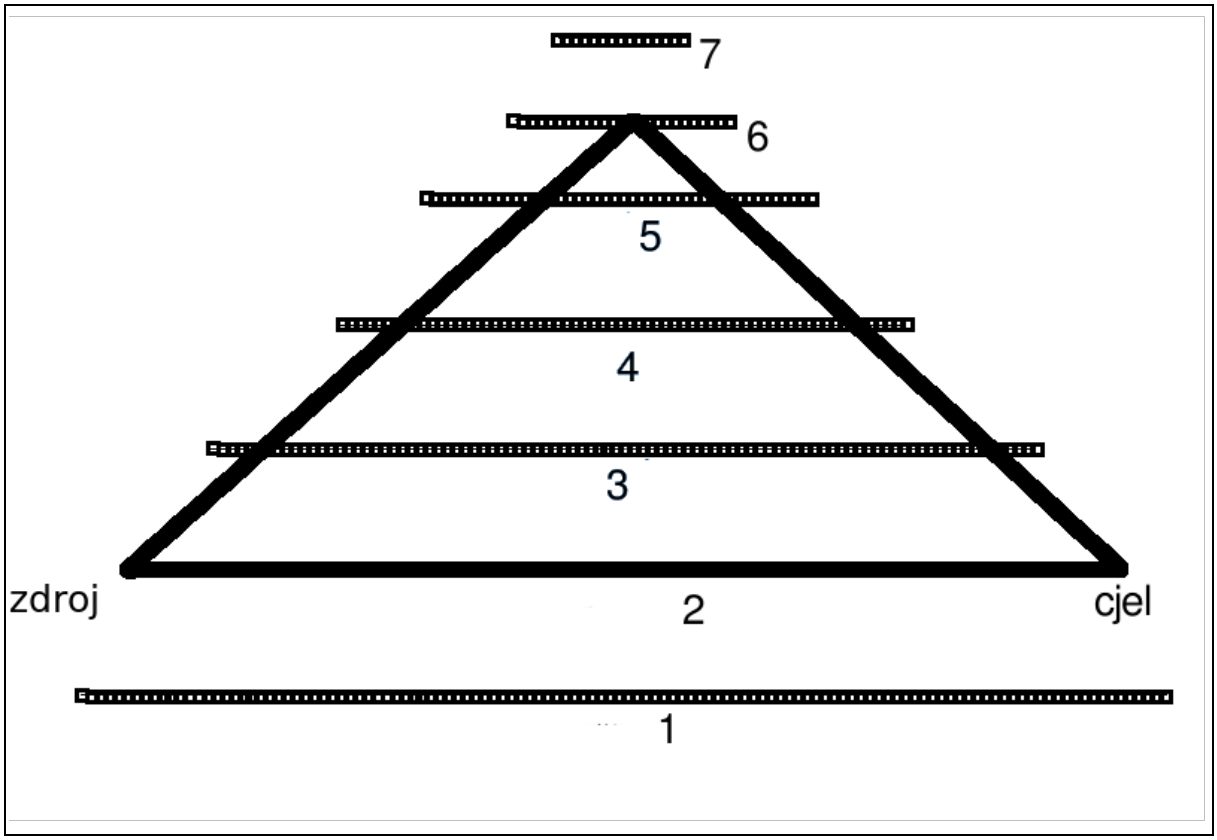
korpus@korpus.juls.savba.sk, http://korpus.juls.savba.sk

Abstract. Machine translation systems tend to be rather complicated and the results are often disappointing. However, the difficulties involved in a successful translation diminish when dealing with a pair of very close languages, and the translation can be ameliorated by strategic use of common morphological, grammar and lexical features of the languages involved. Presented system can be used for translation on the orthographic and lexical level between very close languages and was successfully applied to translation from standard Slovak into the L. Štúr's Slovak language.

Úvod

Sistemi automatickjeho prekladu patrja k najkomplikovaňejším aplikácijám v oblasťi počítačovjeho spracuvaňja prirodzenjeho jazika. Toto viplíva z potrebi urobiť hĺbkovú analízu zdrojovjeho nárečja a transformovať zmisel puovodnjeho textu do cjelovjeho nárečja. Schematicki muožeme proces prekladu znázorňiť diagramom podobním tomu na obr. 1. Plocha trojuholňíka vijadruje oblasť, v ktorej pracujú tipickje sistemi automatickjeho prekladu. Každá vodorovná čjara zodpovedá abstraktnej úrovňi transferu medzi zdrojovím a cjelovím nárečím. Čím višša úroveň, tím abstraktňejší transfer sa uskutočňuje, a vísledok je tím bližší prirodzenjemu nárečju. Úroveň 1vá zodpovedá fonetike a na obrázku je uveďená len kvuoli úplnosťi, pretože vo večšiňe sistemou automatickjeho prekladu (ako aj v našom článku) iďe o písaní text. Úroveň 2há zodpovedá ortografii, transfer na tejto úrovňi znamená len zmenu ortografickjeho sistemu (takíto transfer je použiťelní napríklad pri zmeňe pravopisu jedného nárečja, alebo preklad medzi nárečjami, ktorje sa líšja iba ortografiou). Úroveň 3ťja zodpovedá morfologii a je použiťelná pre preklad medzi nárečjami, ktorje sa odlišujú maximálňe morfologiou (s istími obmedzeňjamí muože ísť o dve velmi blízkje príbuzňje nárečja). Pri odlišňejších nárečjach dostaňeme na vístupe sintakticki a semanticki ňezmiselní text. Úroveň 4tá zodpovedá sintaxi, na vístupe dostaňeme text sintakticki správni, aj keď možno s ňezmiselním víznamom (alebo s víznamom ňezodpovedajúcim originálnemu textu). Moderňje špičkovje sistemi automatickjeho prekladu sa k tejto úrovňi iba približujú. Úroveň 5ta, semantika, zodpovedá pochopeňju víznamu slov a slovních spojení originálnjeho textu a ich preklad na slovňje spojeňja s rovnakím víznamom – na tejto úrovňi pracujú prekladaťelá-luďja. Úroveň 6ta, na diagrame znázorňená vrcholom trojuholňíka zodpovedá užiťju interlingvi (medzireči), pri ktorom preklad prebehou už po stranách trojuholňíka a transfer sa zredukovau na identicku operácju, pretože všetki črti puovodnjeho aj preloženjeho textu sú obsjahnutje v medziprodukťe. Do diagramu sme ešťe doplňili sjedmu úroveň, ležjacu nad vrcholom

trojuholňíka. Táto úroveň bi sa dala opísať ako „pochopeňja toho, čo chceu autor povedať“ a jej znázorňeňja je vjacmeňej iba akademickje, pretože k dosjahnuťju tejto úrovňe dochádza velmi zrjedka.



Obrázok 1ví: Schematicki znázorňení trojuholňík prekladu

Preklad medzi velmi blízkimi nárečjami

Blízkje (geneticki aj štrukturňe) nárečja majú vela podobních čŕt. Pri vzďalovaní nárečí rozďjeli medzi ňimi celkom dobre sledujú úrovňe v uveďenom trojuholňíku – najprú sa zjavja rozďjeli v fonetike (aj v rámci jednjeho nárečja či dokonca rozličnorečja), potom v ortografii (pri kodifikácii alebo odšťjepení nárečja, často s politickou motiváciou). Pri morfologickích rozďjeloch sme už oprávňení hovoriť o ruoznich nárečjach. Sintax často zostáva kompatibilná aj pri nárečjach od seba značňe vzďjaleních, a v prípaďe dramatickích rozďjelou v lexike už ňemuožeme hovoriť o blízkich nárečjach v našom poňímaní. Z automatickích prekladovích sistemou medzi blízkimi nárečjami muožeme spomenúť preklad medzi Češťinou a Slovenčinou[1] a preklad medzi Turečťinou a krimskou Tatárčinou[2].

Štúrovská Slovenčina

Spisovnuo Slovenskuo nárečje, tak ako ho definovau Ludevít Štúr v [3] sa od modernej Slovenčini [4] líši na prví pohlad prevažňe ortografiou, pričom rozďjeli sú lahko algoritmicki popísaťelňje. Hlavňje ortografickje rozďjeli spočívajú v absencii grafemi „y“, v inej realizácii dvojhlások a v explicitnom povinnom značení mekkosťi spoluhlások d, t, n.
Lexikálňje rozďjeli sú subtílňejšje, na prví pohlad badaťelňje len v ňjektorích najčasťejších slovách, ale v skutočnosťi mjerňe posúvajúce semantickí víznam celích trjed slov.

Technická realizácia

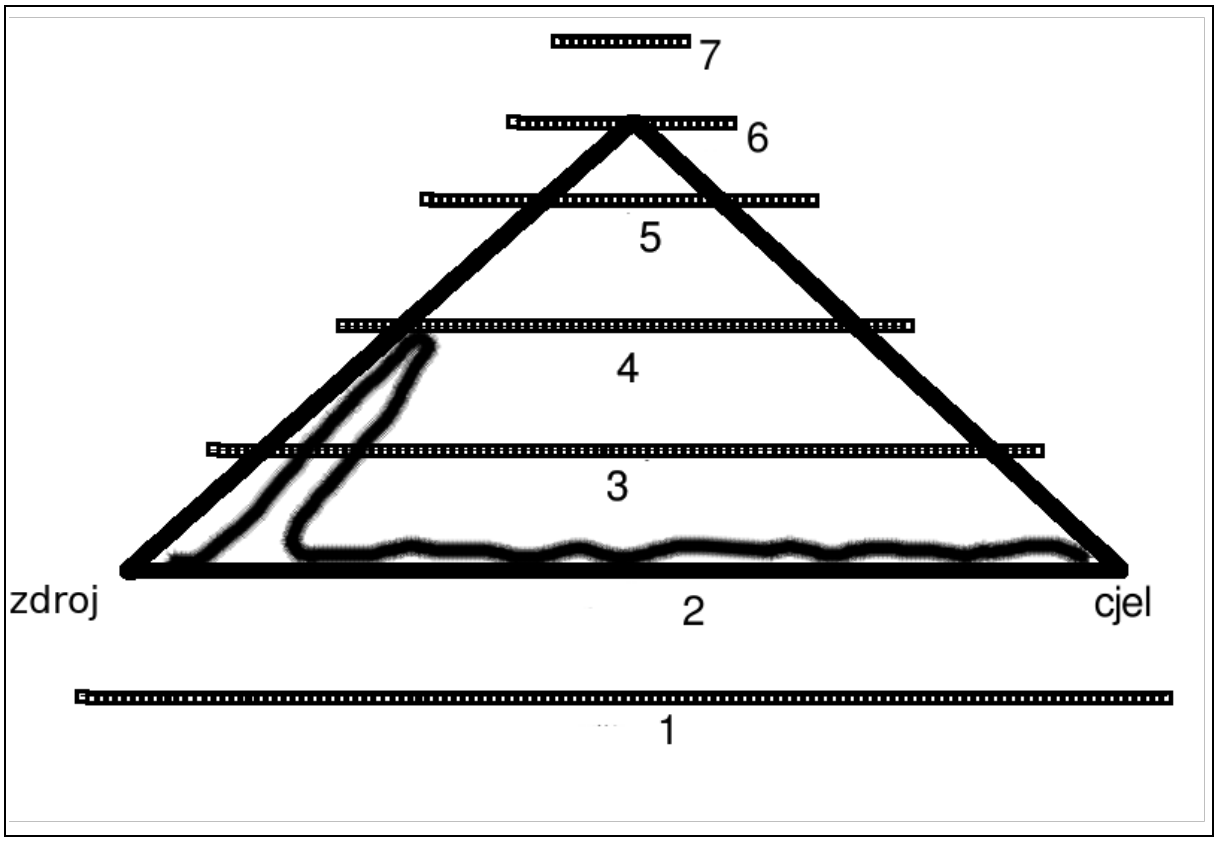
Pred prekladom je text najprú skonvertovaní zo vstupnjeho kódovaňja do Unicode, potom normalizovaní na NFKC normalizácju Unicode a všetki ďalšje operácje prebjehajú duosledňe v Unicode. Text je tokenizovaní na základňje jednotki – tokeni (slová), ku každjemu tokenu je priraďená informácja o prípadních bjelich znakoch (whitespace; Leerraum) pred slovom, abi sa po preklaďe mohlo zrekonštruovať vernuo rozloženja textu. Po preklaďe je velkosť písmen preloženjeho slova upravená tak, abi kopírovala velkosť písmen puovodnjeho slova – ak je preloženuo slovo dlhšje ako puovodnuo, velkosť „nadbitočních“ písmen kopíruje velkosť poslednej písmeni puovodnjeho slova. Toto zabezpečí verní preklad vlastních mjen a prípadních slov písaních kapitálkami. Ako víňimka sú koreňe slov „Sloven(čina, ská)“ a „Vlád(a)“ vždi v preklaďe písaňje so začjatočním velkím písmenom, podla úzu užívanjeho L. Štúrom.

Samotní preklad prebjeha v dvoch fázach: najprú sa aplikuje lexikálna transformácja, pri ktorej sa nahrádzajú slová, ktorje sú v štúrovej Slovenčiňe inak reprezentovaňje. Víhodňe sa dá viužiť prevažná ekvivalencja morfem medzi súčasnou a štúrovskou Slovenčinou a v prekladovej tabulke stačí povečšiňe uvjesť iba prekladi koreňových morfem, iba ňjekedi je potrebnuo uvjesť preklad celích tvarov slov.

Druhá fáza prebjeha na ortografickej úrovňi. Prekladi v oboch fázach sú realizovaňje jednoduchím nahrádzaňím originálnich reťazcou prekladovimí. Začjatki a konce slov sú označeňje špecijálnimí znakmí (^ začjatok, $ koňjec), čo umožňuje efektívňe spracuvať transformácje v príveskách slov a zabraňuje možním ňesprávnim nahraďeňjam. Vzhladom na duosledňje značeňja palatalizovaních spoluhlások v štúrovskej Slovenčiňe je potrebnuo duokladňe rozlišovať tvrdje a mekkje „i“ podla víslovnosťi (čo veďe k prekladom

politi→polity, diplo→dyplo, poézia→poesya) a takťjež bolo potrebnuo zavjesť tvrdje „e“ na označeňja ňepalatalizujúceho „e“ (toto písmeno sme arbitrárňe označili znakom „ë“ – U+00EB LATIN SMALL LETTER E WITH DIAERESIS, príkladi prekladou: internet→intërnët). V druhej úrovňi budú tjeto slová transformovaňje na štúrovskí pravopis (polity→politi, dyplo→diplo, poesya→poesia, intërnët→internet).

Naša skratka v prekladovom trojuholňíku potom sleduje transfer na ortografickej úrovňi, s krátkim vibočeňím do oblasťi semantiki (vlastňe iba zámena ňjektoríchlexikálnich jednoťjek).



Obrázok 2hí: Trojuholňík prekladu so znázorňeňím našej skratki

lexikálni preklad		ortografickí preklad	
u'grék'	: u'rék',	u'ov$'	: u'ou$',
u'gréc'	: u'réc',	u'né$'	: u'ňje$',
u'gréč'	: u'réč',	u'é$'	: u'je$',
u'maďar$'	: u'uher$',	u'ého$'	: u'jeho$',
u'maďar'	: u'uhr',	u'ému$'	: u'jemu$',
u'maďara'	: u'uhra',	u'é'	: u'e',
u'talian'	: u'talyan',	u'ý'	: u'í',
u'^ľudovít'	: u'^ludëvít',	u'y'	: u'i',
u'slávneho'	: u'slávnjëho',	u'ô'	: u'uo',
		u'ľ'	: u'l',
# pieseň -> peseň		u'ä'	: u'e',
u'^pies'	: u'^pes',	u'ë'	: u'e',
u'vidieť$'	: u'videť$',		
u'vedieť$'	: u'vedeť$',	u'ia'	: u'ja',
u'vedie'	: u'vede',	u'dia'	: u'ďja',
u'erie$'	: u'ere$',	u'diakon'	: u'diakon',
u'erieš$'	: u'ereš$',	u'tia'	: u'ťja',
u'^zmenši'	: u'^umenši',	u'nia'	: u'ňja',

Tabulka 1vá – časť prekladovej lexikálnej a ortografickej tabulky

Popis funkcií programu

Program (nazvaní ludevít) je napísaní v programovacom jaziku Python. Hlavnuo zameraňja programu je pre unixovje sistemi, ašak, súc napísaní len s užiťím štandardních pythonovskích kňižňíc, funguje na velmi širokej množiňe sistemou a platform. Program funguje ako filter, čítajúc štandardní vstup a zapisujúc preložení text na štandardní výstup.

Výstup je možno modifikovať ďalšimí argumentamí k programu:

-o súbor alebo --output-file súbor – výstup zapíše do súboru mjesto na štadardní výstup

-D alebo --nfkd – výstup buďe v NFKD normalizácii

-d alebo --nfkd-hack – písmeni ď a ť budú v NFKD normalizácii, ostatnje v NFKC

-e ENCODING alebo --encoding ENCODING – mjesto štandardnjeho koduvaňja utf-8, predpokladá vstup a výstup v koduvaňí ENCODING, ktoruo muože biť hocijaké koduvaňje podporovanuo pythonom, ale pravďepodobňe víznam má len jedno z utf-8, iso8859_2, cp1250, cp852 alebo mac_latin2. Koduvaňja inuo ňež utf-8 ňje je kompatibilnuo s volbamí -D a -d.

Kďe sa zvuki mekko vislovujú takjeto sa zmekčujúcou čjarkou viznačujú, ale písmeni „d“ a „t“ ju v dobe modernej inakšje označujú, značka táto skoro ako dlhá čjarka má podobu. Abi sa historická vernosť zachovala, tjeto dve písmeni je možno normalizovať na unicodovskí

„NFKD“ spuosob (parameter -d, prípadne normalizovať všetki písmeni parametrom -D), to značí že zmekčujúce čjarki sú ako samostatnje kombinujúce písmeni (combining characters, kombinierende diakritische Zeichen) reprezentovanje, keď sa s predchádzajúcou písmenou vjažu, v renderovacích sistemoch zrjedka bíva úplná podpora kombinujúcich písmen, a tak sa často písmena nad predchádzajúcou ňezmeňená zobrazí, čo vizerá temer ako puovodní historicki správni spuosob písaňja. Žjal, mnohokrát sú tjeto čjarki zle zobrazenje, alebo naopak tak ako majú biť skombinuvanje dobre a správňe (na moderní spusob) zobrazenje, a teda tento spuosob ňje vždi dobrje vísledki dáva.

Nedostatki

Ňedostatki uveďenjeho sistemu prekladu (či prepisu) muožeme rozďeliť na dve skupini. Prvú skupinu tvorja ňedostatki teoreticki ňepodstatňje, ktorje je aspon teoreticki možnuo lahko odstráňiť aplikovaňím dostatočnjeho množstva ludskej práce. Sem patrí hlavňe malí rozsah prekladovjeho slovňíka (na lexikálnej úrovňi) a chibi v slovňíkoch na lexikálnej aj ortografickej úrovňi. Chibi je možnuo odstráňiť duokladním skontrolovaňím slovňíkou, a malí rozsah slovnej zásobi samozrejme doplňeňím – okrem potrebi ňevihnutnej ludskej práce tu ňje sú žjadnje problemi, ktorje tomuto principijálňe bráňja.

Ňedostatki teoreticki podstatňje sú horšje, pretože viplívajú buď prjamo z návrhu prekladacjeho sistjemu, alebo z vroďeních vlastností oboch versií Slovenčini a vzťahu medzi ňimi. Tjeto ňedostatki ňje je možnuo jednoducho odstráňiť. Najzávažňejšje z ňich sú:

- Absencja kontextovjeho prekladu. Sistem sa vždi pozerá iba na jedno konkretňe slovo. Toto zabraňuje možnosťi lexikálňeho prekladu slovamí s iním pohlavím, pretože ňje je možnuo súčasňe preložiť prípadňje adjekťíva a slovesá (v menosloví) tak, abi bola zachovaná zhoda pohlaví. V programe sme urobili jeďinú víňimku, slovo „Bratislava“ prekladáme slovom „Prešporok“ (spolu s patričnimí tvarmi skloňeňja), pretože iďe o slovo dosť známe a často užívaňje, a občas sa viskitnuvšju chibu v ňesúlaďe pohlavja prípadnjeho mena prídavnjeho sme považuvali za menšje zlo ako poňechať tvar „Braťislava“ (s ňejasním užívaňím v štúrovskích dobách)
- Ňerozlíšiťelná homonímja v štandardnej Slovenčiňe. Najčasťejším príkladom sú prídavňje mená ňijakjeho pohlavja v nominative a akusative v jednotnom počťe (príveska -uo) a prídavňje mená ženskjeho a ňijakjeho pohlavja v množnom počťe (príveska -je), ktorje v štandardnej Slovenčiňe majú rovnakú prívesku (-je, alebo -e ak bola predchádzajúca silaba dlhá) a ňje je ňijakí spuosob, ako bez duokladnej semantickej analisi určiť správni preklad (taketo slovňje spojeňja bívajú často

ňerozhodnuťelňje aj skúsením čitaťelom-človekom).

- Absencja úpravi sintaxe. Štúrovská Slovenčina sa od štandardnej odlišuje aj mjerňe inou skladbou veti, náš sistem ňeobsahuje nijakje prostrjedki aňi na sintaktickú analisu originálu aňi úpravu prekladu.

Zhrnuťja

Uveďení sistem umožňuje základní preklad zo štandardnej Slovenčini do štúrovskej na ortografickej a čjastočňe lexikálnej úrovňi. Preklad ňje je celkom dokonalí a od originálnej štúrovskej Slovenčini sa odlišuje hlavňe v sintaktickej skladbe vjet a v lexike, ale je dostatočňe dobrí na občasňje užiťje, na demonštrácju štúrovskej Slovenčini a ako pomuocka pre prekladaťelou do štúrovskej Slovenčini. Po jednoduchej úprave (náhrada slovňíka) je možnuo program užiť pre preklad medzi podobňe odlišnimí jazikovimí sistemamí (napríklad v prípaďe závažňejších zmjen v Slovenskom pravopise pri prekladďe do novej normi).

Program je dostupní pod licenciou GNU GPL v. 2.0, a jeho demoversiu prístupnú cez WWW rozhraňja prostredňíctvom jednoduchjeho CGI skriptu je možnuo si prezreť na stránke Jazikovednjeho ústavu Ludevíta Štúra SAV[5]. O potrebe a užitočnosťi takjeho prekladu svedčí aj ňeočakávaná popularita, ktorej sa istú dobu uveďenuo WWW rozhraňja ťešilo[6].

Literatura

1. Hajič, Jan – Hric, Ján – Kuboň, Vladislav: Machine translation of very close languages. In: Proceedings of the sixth conference on Applied natural language processing, Seattle, Washington, s. 7 – 12. Morgan Kaufmann Publishers Inc., San Francisco, 2000.
2. Altıntaş, Kemal: Turkish To Crimean Tatar Machine Translation System, MSc Thesis, Bilkent University Computer Engineering Department, July 2001
3. Štúr, Ludevít: Nauka reči Slovenskej. Prešporok, Tatrín, 1846.
4. Morfológia slovenského jazyka. Red. J. Ružička. Bratislava, Vydavateľstvo Slovenskej akadémie vied 1966. 896 s.
5. http://vvv.juls.savba.sk/ludevit/
6. Štúrovská slovenčina. In: SME, 20. 12. 2006, s. 29