

Artimų kalbų mašininis vertimas

Peter Homola, Erika Rimkutė

Anotacija. Mašininis vertimas yra viena iš didžiausių kompiuterinės lingvistikos užduočių. Vertimo rezultatai tuo geresni, kuo siauresnė vertimo duomenų tema. Kitas svarbus kriterijus yra abiejų kalbų artumas. Kaip paaiškėjo ankstesniuose projektuose, vertimui tarp artimų kalbų (pvz., tarp slavų) nereikia pilnos lingvistinės analizės – užtenka, pavyzdžiui, tik daiktavardžių ir panašių grupių analizės. Tai patvirtino Prahos Karlo universiteto Formaliosios ir taikomosios lingvistikos institute sukurta vertimo tarp slavų kalbų sistema *Česílko*, kuri jau pritaikyta ir kito tipo kalboms. Straipsnyje pristatysime eksperimentinį šios sistemos modelį, pritaikytą lietuvių kalbai. Aprašysime vartojamas paprastas gramatikas, analizuojančias daiktavardžių ir prielinksnių grupes, čekų ir lietuvių kalbų panašumus bei skirtumus, dvikalbį žodyną, pagal kurį verčiamos visos pagrindinės formos ir linksniuojami arba asmenuojami kaitomi žodžiai. Be to, paaiškinsime, kaip įvertinamas vertimo rezultatas.

Įvadas

Mašininis vertimas yra labai sudėtinga sritis. Šios kompiuterinės lingvistikos šakos metodai skiriami į dvi grupes: *klasikinius* (vartojamos iš taisyklių sudarytos gramatikos) ir *statistinius*. Šis straipsnis aprašo pirmosios grupės metodus.

Tipiška mašininio vertimo sistema susideda iš trijų fazių: *analizės*, *sintaksinių struktūrų modifikavimo* ir *sintezės*. Pirmą analizuojamas verčiamas sakinytis. Šios fazės rezultatas yra sintaksinė sakinio struktūra. Kitame etape modifikuojama ši struktūra, kad atitiktų kalbos, į kurią verčiama, taisykles. Pagaliau generuojamos taisyklingos kaitomų žodžių formos.

Prahos Karlo universiteto Formaliosios ir taikomosios lingvistikos institute devintajame dešimtmetyje buvo sukurta dvi mašininio vertimo sistemos: *APAC3*, verčianti iš anglų kalbos į čekų (Kirschner, 1987), ir *RUSLAN*, verčianti iš čekų kalbos į rusų (Oliva, 1989). Kuriant pastarąją sistemą iškilo klausimas, ar reikia pilnos sintaksinės analizės ir sintaksinių struktūrų modifikavimo. Sistema *Česílko* (žr. 3 skyrių) parodė, kad vertimui tarp artimų kalbų užtenka dalinės analizės ir sintaksinių struktūrų modifikavimo, nes panaudojami abiejų kalbų panašumai, ypač tipologiniai. Kituose skyriuose aprašyti įvairūs tokio vertimo aspektai.

2 straipsnio skyriuje paaiškiname kalbų panašumus ir tai, kokią įtaką jie daro vertimo kokybei. 3 skyriuje aprašyta mašininio vertimo sistema *Česílko* (sukurta slavų kalboms) ir jos pritaikymas lietuvių kalbai. 4 dalyje aprašomas sintaksinės analizės modulis. Dvikalbių žodynų struktūrai ir vartojimui skirtas 5 skyrius. 6 dalyje išsamiai aprašyta sintaksinių struktūrų modifikavimo fazė. Vertimo kokybės įvertinimas aprašytas 7 skyriuje, o paskutiniame, 8 skyriuje, pateikiame išvadas ir tolimesnius planus, susijusius su mūsų būsimo darbu.

Kalbų panašumas

Galima skirti keturias kalbų panašumo grupes: tipologinius, leksikinius, morfologinius ir sintaksinius panašumus. Svarbiausi yra tipologiniai panašumai. Pvz., baltų ir slavų (išskyrus bulgarų ir makedonų kalbas) kalbos yra kaitomos ir

turi labai laisvą žodžių tvarką. Todėl verčiant, pavyzdžiui, iš čekų kalbos į lietuvių, beveik nereikia keisti žodžių tvarkos ir vertimo kokybei didelės įtakos neturi veiksmazodžių paradigmos. Labiau skiriasi sakinio dalių, pavyzdžiui, daiktavardžių grupių, struktūra (svarbiausi skirtumai tarp čekų ir rusų kalbų aprašyti Homola, 2002).

Leksiniai panašumai nėra tokie svarbūs kaip tipologiniai. Didesnė problema yra semantinis daugiareikšmiškumas, t.y., kai tas pats žodis, atsižvelgiant į kontekstą, gali būti verčiamas skirtingai, pvz., žodis *jeřáb* verčiamas kaip *kranas* ar *gervė*. Teisingas vertimas dažnai priklauso nuo teksto temos ir sprendžiamas specialiuose žodynuose (Hajič et al, 2003). Tokių leksinių ir semantinių skirtumų tarp artimų kalbų gana mažai. Verčiant platesnės temos tekstą būtų galima panaudoti vieną iš metodų, aprašytų Pecina et al, 2002 straipsnyje.

Kai kalbos skiriasi tipologiškai, morfologiniai skirtumai dažniausiai būna labai dideli. Tipologinis artumas dažnai reiškia, kad kalbų morfologinė sistema yra artima, kitaip sakant, tarp morfologinių abiejų kalbų sistemų yra mažai skirtumų. Pavyzdžiui, čekų kalba turi 7 linksnius ir jie beveik tiksliai atitinka lietuvių kalbos linksnius. Čekų kalboje nėra seniau vartotų ir dabartinėje kalboje kai kur pasitaikančių linksnių, kaip, pavyzdžiui, lietuvių kalbos *iliatyvas*. Bet šio linksnio vartoseną gana aiški, todėl nekyla problemų jį analizuojant.

Vienoje kalboje dažnai morfologiškai sutampa dvi reikšmės, išreiškiamos kitoje kalboje dviem skirtingomis formomis. Pavyzdžiui, čekų kalboje neretai sutampa vyriškosios giminės gyvos būtybės vienaskaitos kilmininkas ir galininkas, pvz., *velkého* reiškia *didelio* ar *didelį*, priklausomai nuo konteksto. Panašiai sutampa padalyvių daugiskaitos formos: padalyvis *majíce* gali būti susijęs su bet kokios giminės forma (reiškia *turėdami* ar *turėdamos*). Lietuvių kalboje tai yra visai atskiros morfologinės formos. Truputį problemiškesnė yra lietuvių kalbos veiksmazodžių sistema – ji sudėtingesnė nei slavų kalbose. Pavyzdžiui, slavų kalbose nėra būsimosio laiko dalyvių. Sakinį iš Gamut (1991)

(1a) *Gimė vaikas, valdysiantis pasaulį.*

reikėtų išversti į čekų (ar kitą slavų) kalbą panaudojant šalutinį sakinį su prijungiamuoju žodžiu (lietuvių kalboje prijungiamasis žodis nėra būtinas) ir būsimojo laiko veiksmažodžiu:

(1b) *Narodilo se dítě, které bude vládnout světu.*

Šalutinis sakiny (1b) yra pažodinis šalutinio sakinio *kuris valdys pasaulį* vertimas. Pavyzdyje (1b) matomas dar vienas svarbus sintaksės skirtumas: čekų kalboje vartojamas pagalbinis veiksmažodis būsimajam laikui sudaryti (šiam pavyzdyje *bude vládnout*). Taip pat, t.y., naudojant pagalbinį veiksmažodį, sudaroma tariamoji nuosaka (pvz., *vládnul by* reiškia *valdytų*). Ši problema aprašyta 5 skyriuje.

Ne mažiau problemiška yra padalyvių su savo veiksnium vartosena, nes tokios konstrukcijos neturi tiesioginio atitikmens ir dažniausiai verčiamos nominalizuojant veiksmažodį, pvz.:

(2) *Sprogus bombai žuvo žmogus.*

Padalyvis *sprogus* verčiamas prielinksnine fraze *při výbuchu* ar *po výbuchu* (pažodžiui *per sprogimą* ar *po sprogimo*). Lietuvių kalbos padalyvio reikšmė (semantinė funkcija) yra platesnė (žr. Panevová, 1980; Sgall et al, 1974).

Didžiausias sintaksinis skirtumas yra kitokia daiktavardžių grupių žodžių tvarka, nes čekų kalboje nederinami kilmininko pažyminiai vartojami po valdančiojo žodžio, pvz., *kniha bratra* – brolio knyga. Taip pat skirtinga kelių prielinksnių grupių žodžių tvarka, pvz., frazė *k lesu* verčiama *link miško* arba *miško link*.

Mašininio vertimo sistema ČESÍLKO

Mašininio vertimo sistema artimoms kalboms *Česílko* buvo sukurta Prahos Karlo universiteto Formaliosios ir taikomosios lingvistikos institute (ÚFAL) (Hajič et al, 2000; Dębowski et al, 2002; Homola et al, 2003). Ši sistema susideda iš kelių modulių, analizuojančių kalbos, iš kurios verčiama, morfologiją arba sintaksinę struktūrą ir generuojančių kalbos, į kurią verčiama, tekstą. Iki šiol sukurtos trys kalbų poros: čekų-slovakų, čekų-lenkų ir čekų-lietuvių.

Pirmasis modulis yra čekų kalbos morfologinis anotatorius, analizuojantis morfologines žodžių kategorijas. Verčiamą tekstą reikia morfologiškai vienareikšminti, nes sintaksinės analizės moduliui būtina, kad vienam žodžiui būtų priskirta tik viena pažymų grupė. Žáčková parodė, kad daline gramatika neįmanoma morfologiškai vienareikšminti čekų kalbos tekstų (Žáčková, 2002).

Sistemoje vartojamas statistinis anotatorius, sukurtas Jano Hajičiaus (Hajič, 2001). Šio morfologinio anotatoriaus kokybė didesnė nei 95%, t.y., 95% pažymų grupių priskirti visiškai teisingai. Dažniausia klaida yra blogai atpažintas linksnis (kai sutampa dvi ar daugiau linksnių formų), kartais neteisingai parinkta pagrindinė žodžio forma (kai atsitiktinai sutampa kelių skirtingų žodžių formos, pvz., *sen* yra žodžio *sen* „sapnas“ vardininko vienaskaitos forma ir žodžio *seno* „šienas“ kilmininko daugiskaitos forma; *žena* reiškia „moteris“ bei „varant“ ir pan.).

Verčiant iš čekų kalbos į lietuvių, vartojamas sintaksinės analizės modulis, aprašytas ketvirtajame skyriuje. Po morfologinės ir sintaksinės analizės seka sintaksinių struk-

tūrų modifikavimo fazė. Šioje fazėje pagal specialius žodynus išverčiamos pagrindinės žodžių formos ir keičiami morfologiniai požymiai ir, kai to reikia, žodžių tvarka.

Galiausiai generuojamos teisingos kaitomų žodžių formos. Klaidų gali atsirasti pirmose trijose fazėse. Apie klaidas, daromas statistiniu anotatoriumi, jau rašyta. Jei naudojama sintaksinė analizė, ji yra tik dalinė, t.y., neanalizuojamas visas sakiny, bet tik jo dalys, dažniausiai daiktavardžių ir prielinksnių grupės. Daug priklausomybių lieka neatpažintų, todėl gali atsirasti klaidų, kai tokia „slapta“ priklausomybė daro įtaką morfologiniams požymiams ir jie abiejose kalbose skiriasi (pvz., veiksmažodžių valdymas turi įtakos priklausančių žodžių linksniams).

Vertimo procesą paaiškina pavyzdys (1):

<i>Tuto</i>	<i>knihu</i>	<i>začnu</i>	<i>číst</i>	<i>později.</i>
Šią	knygą	pradėsiu	skaityti	vėliau.

Čekų kalbos sakiny šiuo atveju verčiamas pažodžiui, nereikia pakeisti nei morfologinių požymių, nei žodžių tvarkos. Pakeitus daiktavardžių grupę *tuto knihu* grupe *knihu bratra* (brolio knyga) reikia keisti žodžių tvarką, nes čekų kalboje nederinami pažyminiai vartojami po valdančiojo daiktavardžio.

Linksnių pakeitimą paaiškina pavyzdys (2).

<i>Jedu</i>	<i>do</i>	<i>Prahý</i> – kilmininkas
<i>Važiuuju</i>	<i>í</i>	<i>Prahá</i> – galininkas

Nors ir šiuo atveju verčiama pažodžiui, kyla problema, kadangi čekų prielinksnis *do* „į“ reikalauja kilmininko, o ne galininko, kaip jo lietuviškas atitikmuo. Taigi reikia išanalizuoti šią prielinksninę grupę ir sintaksinių struktūrų modifikavimo fazėje linksnį reikia pakeisti į galininką (nepakeitus jo gaunamas neteisingas rezultatas **í Prahos*). Panašiai keičiama giminė ar skaičius išvertus pagrindinę formą, kai šios kategorijos abiejose kalbose skiriasi, pvz. (3):

<i>obydlený</i>	<i>ostrov</i> – vyr. g
<i>gyvenama</i>	<i>sala</i> – mot. g

Kadangi žodis *ostrov* yra vyriškosios giminės, todėl vartojama derinamojo pažyminio *obydlený* vyriškoji forma. Išvertus daiktavardį pasikeis giminė, bet to neužtenka – dėl derinimo reikia pakeisti ir dalyvio giminę. Nepakeitus derinamų priklausančių žodžių kategorijų vertimo rezultatas yra negramatiškas (šiuo atveju **gyvenamas sala*) arba turi kitą reikšmę.

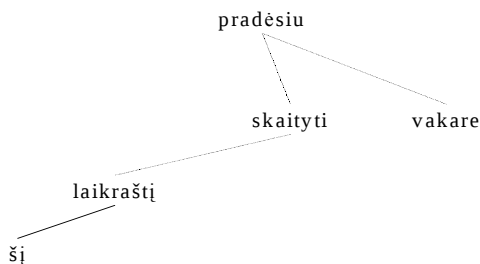
Dalinė sintaksinė analizė

Mašininiam vertimui tarp labai artimų kalbų, pvz., čekų ir slovakių, nereikalinga netgi dalinė sintaksinė analizė. Gautas gana neblogas į lenkų kalbą vertimo rezultatas, nors šiuo atveju sintaksinė analizė padėtų pagerinti vertimą.

Kai nėra jokios sintaksinės analizės, neįmanoma pakeisti žodžių tvarkos ir priklausančių žodžių morfologinių požymių, t.y., reikia didelio tipologinio bei leksinio panašumo.

Verčiant iš čekų kalbos į lietuvių, be abejo, reikia sintaksinės analizės. Sakinio (6) sintaksinė struktūra pavaizduota 1 paveiksle, slaptos briaunos pažymėtos punktyru:

(1) *Šį laikraštį pradėsiu skaityti vakare (Tyto noviny začnu číst večer).*



1 paveikslas. Sakinio *Ši laikraštį pradėsiu skaityti vakare* sintaksinė struktūra

Sakinį (1) galima versti pažodžiui, bet reikia pakeisti daiktavardžio *laikraštis* giminę į moteriškąją ir skaičių į daugiskaitą – *noviny*. Taip pat reikia pakeisti atitinkamas įvardžio *šis* kategorijas. Minėtas sakinytis ir jo vertimas yra neprojekcinis (apie neprojekcinius sakinius išsamiau rašo Kuboň, 2001).

Sakiniams analizuoti vartojama laisvojo konteksto gramatika (nors vėliau vartojamos priklausomybių struktūros; apie skirtumus žr. Dikovski et al, 1970), susidedanti iš specialių taisyklių, išreiškiančių sintaksines priklausomybes tarp žodžių. Pavyzdžiui, paprasta taisyklė, analizuojanti daiktavardžių grupes, susidedančias iš būdvardžio (ar daugiau būdvardžių) ir daiktavardžio, yra (2):

$$(2) NP \rightarrow A \text{ NP}$$

A reiškia būdvardis (*adjective*), NP daiktavardžių grupė (*noun phrase*). Kadangi galima analizuoti tik derinamuosius būdvardžius, dar reikalingos tokios sąlygos:

Giminė (A) = giminė (NP), skaičius (A) = skaičius (NP), linksnis (A) = linksnis (NP)

Panašiai sudaromos kitos taisyklės, analizuojančios prielinksnių grupes ir pan. Gramatikai pritaikyti galima vartoti sistemas Q (žr. Colmerauer, 1969) ar formalizmą LFG (žr. Bresnan, 2002).

Žodynai

Sintaksinių struktūrų modifikavimo fazėje vartojami dvikalbiai žodynai, atliekantys dvi funkcijas: jie verčia pagrindines formas ir pakeičia morfologinius požymius. Abi funkcijos jau aprašytos 3 skyriuje. Šioje straipsnio dalyje aprašyta žodyno struktūra ir techninių problemų sprendimas.

Kiekvienas žodžių poros įtraukimas į žodyną susideda iš pagrindinės verčiamo žodžio formos ir jo vertimo į kalbą, į kurią verčiama. Be to, kiekvienai porai įmanoma priskirti morfologinių požymių sąrašą. Požymiai gali skirtis dėl dviejų priežasčių:

1. Žodžio vertimas, susijęs su kitomis morfologinėmis vertėmis, dažnai skiriasi, pavyzdžiui, daiktavardžiai giminėmis: čekų *voda* – mot. g. → *vanduo* – vyr. g. Panašiai skiriasi kelių prielinksnių reikalaujantis linksnis, pvz., čekų prielinksnių *kvůli*, vartojamą su naudininku, atitinka lietuvių kalboje *del*, vartojamas su kilmininku. Pirmajame pavyzdyje skirtumas turi įtakos

kaitomoms formoms, paskutiniame yra svarbus keičiant prielinksnio objekto linksnį.

2. Požymių vertės skiriasi dėl morfologinių abiejų kalbų skirtumų.

Pirmojo tipo skirtumai svarbūs tam, kad vertimas būtų gramatiškas. Šie skirtumai nėra vien tik kalbiniai, jie svarbūs ir dėl techninių priežasčių. Pvz., lietuvių dalyviai čekų kalbos anotatoriumi pažymimi kaip būdvardžiai: frazėje *čtená kniha* „skaitoma knyga“ pirmasis žodis čekų kalboje anotuojamas kaip būdvardis, turintis pagrindinę formą *čtený* „skaitytas“, bet taisyklingai lietuvių kalbos formai sudaryti reikia pažymos *dalyvis*, kurios pagrindinė forma yra veiksmazodis *skaityti*. Šios poros įtraukimas į žodyną yra (1):

(1) *čtený* → *skaityti* (kalbos dalis: veiksmazodis, veiksmazodžio forma: dalyvis, rūšis: neveikiamoji, laikas: esamasis)

Kitų požymių vertės (linksnis, giminė, skaičius, laipsnis) lieka tos pačios arba modifikuojamos pagal kitus kriterijus.

Kartais vienas žodis gali būti verčiamas keliais žodžiais, pvz., *rychlík* – *greitasis traukinys*. Ir atvirkščiai: keli žodžiai atitinka vieną, pvz., *rychlostní silnice* – *greitkelis*. Šių porų įtraukimas į žodyną yra toks:

(2) *rychlík* → *greitas* (įvardžiutinis būdvardis) + *traukinys*
rychlostní + *silnice* → *greitkelis* – vyr.g.

Tokios frazės verčiamos iš karto, atpažinus jų sintaksinę struktūrą.

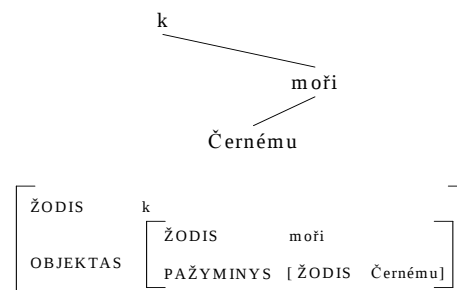
Sintaksinių struktūrų modifikavimas

Sintaksinių struktūrų modifikavimo fazė vartojama šiuo metu tiksliai verčiant į lietuvių kalbą, nes slovakų bei lenkų kalbos gana panašios ir vertimo rezultatas net be šios vertimo fazės yra geros kokybės. Šiame vertimo etape keičiama verčiamo sakinio sintaksinė struktūra, kad atitiktų lietuvių kalbos gramatinės taisyklės. Svarbiausi pakeitimai yra žodžių tvarka ir linksnis (dažniausiai dėl kitokio veiksmazodžių valdymo).

Morfologiniai požymiai keičiami cikliška. Frazės

(1) *k Černému moři* (naudininkas) – *prie Juodosios jūros*

sintaksinis medis ir atitinkama požymių struktūra pavaizduota 2 paveiksle:



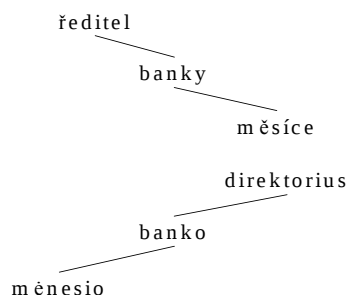
2 paveikslas. Sakinio *k Černému moři* sintaksinis medis

Pirma reikia išversti prielinksnių. Prielinksnis *k* žodyne pažymimas taip:

(2) *k* → *prie* – kilmininkas

Kadangi prielinksnių reikalaujantys linksniai abiejose kalbose skiriasi, reikia pakeisti daiktavardžio linksnį į kilmininką: *jūrai* → *jūros*. Bet po šio pirmojo ciklo dar reikia pakeisti priklausančio būdvardžio linksnį, nes šis žodis yra derinamasis pažyminis: *Juodajai* → *Juodosios*. Taigi po dviejų ciklų gautas toks rezultatas: *prie Juodosios jūros*. Be sintaksinio modifikavimo frazėje liktų naudininko formos ir būtų verčiama taip: **prie Juodajai jūrai*.

Žodžių tvarka dažniausiai keičiama frazėse su nederinamais kilmininko pažyminiiais. Pvz., frazė *ředitel banky mėsice* reiškia „mėnesio banko direktorius“. Žodžių tvarka lietuvių kalboje atvirkščia nei čekų. Abiejų sakinių struktūros yra:



3 pavykslas. Sakinių *ředitel banky mėsice* ir *mėnesio banko direktorius* struktūros

Keli čekų kalbos prielinksniai visai neverčiami į lietuvių kalbą – užtenka tam tikro linksnio formos, pvz., *pro Petra* reiškia „Petrai“; *ve městě* reiškia „mieste“ ir pan. Atitinkamas įtraukimas į žodyną yra toks:

- (3) *pro* → [be prielinksnio] – naudininkas
v → [be prielinksnio] – vietininkas
*ve*¹ → [be prielinksnio] – vietininkas

Įvertinimas

Vertimo kokybė įvertinta naudojant programą *Trados Translator's Workbench*. Įvertinimas atliktas išvertus tekstą mašininio (sistema *Česilko*) ir rankiniu būdu. Programa *Trados Translator's Workbench* palygina kiekvieną automatiškai išverstą sakinį su rankomis išverstu atitikmeniu ir pateikia abiejų variantų panašumą. Viso teksto panašumas yra visų sakinių panašumo verčių (pagal sakinio ilgį, t.y. žodžių skaičių) vidurkis. Visų trijų kalbų porų rezultatai (Dębowski et al, 2002; Homola et al, 2003) pateikti 1 lentelėje:

1 lentelė. Vertimo kokybės įvertinimas

Kalbų pora	Vertimo tikslumas
Čekų-slovakų	90 %
Čekų-lenkų	71,4 %
Čekų-lietuvių	87,6 %

¹ Žodelis *ve* yra prielinksnio *v* vokalizuos variantas, vartojamas prieš žodžius, prasidedančius keliais priebalsiais ar priebalsių kombinacijomis (pvz., *v-, sv-, f-, sf-*).

Komercinė sistema *PC Translator* iš čekų kalbos į anglų išversto teksto panašumas, įvertintas tuo pačiu metodu, yra 30% (Hajič et al, 2003).

Dažniausios vertimo klaidos yra:

- Blogai parinkta pagrindinė forma. Kartais dėl semantinių priežasčių blogai išversta pagrindinė forma, ypač kai žodis daugiareikšmis. Pavyzdžiui, *jméno otce* reiškia *tėvo vardas* arba *tėvavardis*. Vartojant tik dalinę analizę neįmanoma išversti kiekvieno tokio žodžio teisingai.
- Neteisinga veiksmažodžio forma. Kadangi veiksmažodžių formos (pvz., padalyviai) kartais neatitinka kitos kalbos formų, todėl neretai nesutampa ir morfologiniai požymiai. Pavyzdžiui, *čtená kniha* reiškia ir *skaitoma knyga*, ir *skaityta knyga*.
- Neteisingas linksniavimas. Verčiant automatiškai kartais nepakeičiamas linksnis, kur to reikia. Priežastis gali būti veiksmažodžio valdymas arba neviseškai išanalizuota daiktavardžio grupė. Pavyzdžiui, frazė *problém, který nebyl vyřešen* bus išversta **problema, kuris nebuvo išspręsta*, nes dalinė gramatika neanalizuoja šalutinių sakinių.

Išvados

Šiame straipsnyje aprašėme eksperimentinę mašininio vertimo sistemą *Česilko* ir parodėme, kokie kalbos aspektai yra svarbūs verčiant iš vienos kalbos į kitą. Nors pilna sintaksinė analizė su viso sakinio modifikavimu leidžia pasiekti geresnius rezultatus, vertimui tarp artimų kalbų turbūt, užtenka dalinių metodų. Gautas neblogas mašinio vertimo iš čekų į lietuvių kalbą rezultatas – 87,6% tikslumas. Žinoma, toks rezultatas nėra tobulas, bet šiuo metu neturime gero sintaksinio čekų kalbos sintaksinio anotatoriaus (geriausius rezultatus pateikia M. Collinso sukurtas statistinis sintaksinis anotatorius (Collins et al, 1999)). Be mašininio vertimo šiame projekte sukurta dalinė gramatiką būtų įmanoma panaudoti, pavyzdžiui, pagerinant statistinių sintaksinių anotatorių rezultatus (Zeman, 2001) ar išplečiant vertimo atminties sistemas (Homola et al, 2004).

Tolimesni darbai bus dalinės gramatikos tobulinimas ir išplėtimas kitoms slavų kalboms. Be to, norėtume panaudoti Vokietijos Dirbtinio intelekto institute Saarbrükene sukurta sistemą *SproUT* (Becker et al, 2002; Drozdzyński et al, 2003) gramatikoms rašyti.

Literatūra

1. Becker, M., Drozdzyński, W., Krieger, H. U., Piskorski, J. & Schäfer, U. (2002). And Xu F., *SproUT – Shallow Processing with Typed Feature Structures and Unification*, In Proceedings of ICON 2002, Mumbai, India.
2. Bresnan, J. (2001). *Lexical-functional Syntax*, Oxford: Blackwell Publishers.
3. Collins, M. et al. (1999). A Statistical Parser for Czech. In Proceedings of the 37th ACL '99, University of Maryland, College park, MD, USA, pp. 505-512.
4. Colmerauer, A., (1969). *Les systèmes Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur*. Montréal: Mimeo.

5. Dębowski, L., Hajič, J. & Kuboň, V. (2002). Testing the Limits — Adding a New Language to an MT System. *Prague Bulletin of Mathematical Linguistics*, pp. 95–102, Prague.
6. Dikovsky, A. & Modina, L. (1770). *Dependency Grammar*. Problemy peredači informacii, Moskva.
7. Drozdzyński, W., Homola, P., Piskorski, J. & Zinkevičius, V. (2003). Adapting SproUT to Processing Baltic and Slavonic Languages, In *Proceedings of Information Extraction for Slavonic and other Central and Eastern European Languages*, Borovets, Bulgaria.
8. Gamut, L. T. F. (1991). *Logic, Language and Meaning 2: Intentional Logic and Logical Grammar*. Chicago: University of Chicago Press.
9. Hajič, J. (2001). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*, Prague: Karolinum, Charles University Press,.
10. Hajič, J., Hric, J. & Kuboň, V. (2000). Machine Translation of Very Close Languages. In *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, Washington, USA, April 2000, pp. 7–12.
11. Hajič, J., Homola, P. & Kuboň, V. (2003). A Simple Multilingual Machine Translation System. In *Proceedings of the VIII MT Summit*, New Orleans.
12. Homola, P. (2002). Machine Translation Among Slavic Languages. In *Proceedings of the WDS*, Charles University, Prague, 2002.
13. Homola, P. & Rimkutė, E. (2003) Shallow Machine Translation — in Between of Two Extremes. In *Proceedings of the 5th International Symposium of Logic, Language and Computation*, Tbilisi State University, Georgia.
14. Homola, P. & Tolvaj, B. (2004) Distributed Translation Memories and Shallow MT. *Malý informatický seminář*, Josefův dům.
15. Kirschner, Z. (1987) APAC3-2: An English-to-Czech Machine Translation System. *Explizite Beschreibung der Sprache und automatische Textbearbeitung XIII*, MFF UK, Prague.
16. Kuboň, V. (2001). *Problems of Robust Parsing* – PhD thesis, Faculty of Mathematics and Physics, Charles university, Praha.
17. Oliva, K. (1989). A Parser for Czech Implemented in Systems Q. *Explizite Beschreibung der Sprache und automatische Textbearbeitung XVI*, MFF UK, Prague.
18. Panevová, J. (1980). *Formy a funkce ve stavbě české věty*. Studie a práce lingvistické, Praha: Academia.
19. Pecina, P. & Holub, M. (2002). *Sémanticky signifikantní kolokace*. Technical report TR-2002-13, ÚFAL / CKL, Faculty of Mathematics and Physics, Charles University, Praha.
20. Sgall, P. et al. (1974). *Úvod do algebraické lingvistiky*. Univerzita Karlova, SNP, Praha.
21. Zeman, D. (2001). How Much Will a RE-based Preprocessor Help a Statistical Parser, In *Proceedings of the Seventh International Workshop on Parsing Technologies*, Beijing Daxue, Beijing: Tsinghua University Press.
22. Žáčková, E. (2002). *Parciální syntaktická analýza (češtiny)* – PhD thesis. Fakulta informatiky Masarykovy University, Brno.

Peter Homola, Erika Rimkutė

Mechanical Translation between Contiguous Languages

Summary

The results of machine translation as one of the biggest challenges of today's computational linguistics depend on many various criteria such as domain specificity and source and target language similarity. Recent projects have shown that machine translation among related languages (e.g., Slavonic) can be performed without a full-fledged analysis; good results can be achieved by analyzing only simpler constituents. One of such projects is the system *Česilko* developed at Charles University in Prague at the Institute of Formal and Applied Linguistics that has been extended from Slavonic languages to another language family, the Baltic languages. This paper presents the architecture of this system adapted for Lithuanian, describes used partial parser, explains similarities and differences between Czech (source language) and Lithuanian and the structure of the translation dictionary. Moreover the authors explain how they have evaluated the quality of translation.

Straipsnis įteiktas 2004 05

Parengtas spaudai 2004 12

Autoriai

Erika Rimkutė, Vytauto Didžiojo universiteto Lietuvių kalbos katedros doktorantė, Kompiuterinės lingvistikos centro jaunesnioji mokslo darbuotoja. *Mokslinių interesų sritys*: tekstynų lingvistika, kompiuterinė lingvistika, automatinė morfologinė analizė bei sintezė, mašininis vertimas.

Adresas: Donelaičio g. 52-206, Kaunas, Lietuva

El. paštas: e.rimkute@hmf.vdu.lt

Petr Homola, Prahos Karlo universiteto Formaliosios ir taikomosios lingvistikos instituto doktorantas ir darbuotojas.

Mokslinių interesų sritys: kompiuterinė lingvistika, automatinė sintaksinė analizė bei sintezė, analitiniai ir tektogramatiniai tekstynai, mašininis vertimas, vertimo atminties sistemos.

Adresas: Malostranské náměstí 25, Praha, Čekijos respublika.

El. paštas: homola@ufal.mff.cuni.cz

