

Mašininis vertimas tarp artimų kalbų

Petr Homola

*Praha Karlo universiteto Formaliosios ir taikomosios
lingvistikos institutas*

Erika Rimkutė

*Vytauto Didžiojo universiteto Kompiuterinės lingvistikos
centras*

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 Praha Karlo universiteto Formaliosios ir taikomosios lingvistikos institute sukurta vertimo tarp slavų kalbų sistema *Česilko*, 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.

1. 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.

Praha Karlo universiteto Formaliosios ir taikomosios lingvistikos institute devintajame dešimtmetyje buvo sukurtos 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 *Česilko* (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 *Česilko* (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 būsimu darbu.

2. 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 veiksmažodžių 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 veiksmažodžių sistema – ji sudėtingesnė nei slavų kalbose. Pavyzdžiui, slavų kalbose nėra būsimąjo laiko dalyvių. Sakinį iš Gamut 1991

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

reikėtų išversti į čekų (ar kitą slavų) kalbą panaudojant šalutinį sakinį su prijungiamuoju žodžiu (lietuvių kalboje prijungiamasis žodis nėra būtinas) ir būsimąjo 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 *valdytu*). Ši problema aprašyta 5 skyriuje.

Ne mažiau problemiška yra padalyvių su savo veiksmu vartoseną, 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*.

3. 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 proc., t. y., 95 proc. pažymų grupių priskirti visiškai teisingai. Dažniausia klaida yra blogai atpažintas linksnis (kai sutampa dvi ar daugiau linksnų 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 struktūrų 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 linksnams).

Vertimo procesą paaiškina pavyzdys (1):

(1)	<i>Tuto</i>	<i>knihu</i>	<i>začnu</i>	<i>číst</i>	<i>později.</i>
	<i>Šiř</i>	<i>knyřa</i>	<i>praděsiu</i>	<i>skaiityti</i>	<i>vėliau.</i>

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.

Linksnų pakeitimą paaiškina pavyzdys (2).

- (2) *Jedu do Prahy – kilmininkas*
Važiuoju i Prahu – 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 **į Prahos*). Panašiai keičiama giminė ar skaičius išvertus pagrindinę formą, kai šios kategorijos abiejose kalbose skiriasi, pvz.:

- (3) *obydlený ostrov – vyr. g*
gyvenama sala – 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.

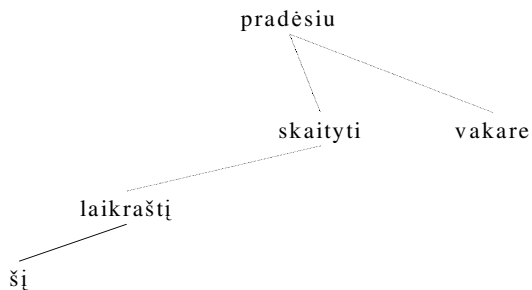
4. DALINĖ SINTAKSINĖ ANALIZĖ

Mašininiam vertimui tarp labai artimų kalbų, pvz., čekų ir slovakų, 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) *Ši 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 → A 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).

5. 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 prielinksni *kvůli*, vartojamą su naudininku, atitinka lietuvių kalboje *dėl*, 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 veiksmažodis *skaityti*. Šios poros įtraukimas į žodyną yra (1):

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

Kitų pažymių 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* (įvardžiuotinis būdvardis) + *traukinys*
rychlostní + silnice → *greitkelis* – vyr.g.

Tokios frazės verčiamos iš karto, atpažinus jų sintaksinę struktūrą.

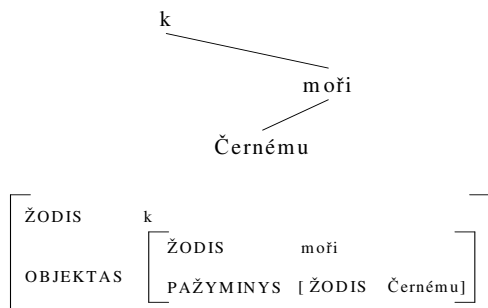
6. SINTAKSINIŲ STRUKTŪRŲ MODIFIKAVIMAS

Sintaksinių struktūrų modifikavimo fazė vartojama šiuo metu tikta 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 veiksmažodžių valdymo).

Morfologiniai požymiai keičiami cikliška. Frazės

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

sintaksinis medis ir atitinkama požymių struktūra pavaizduota 2 paveiksle:



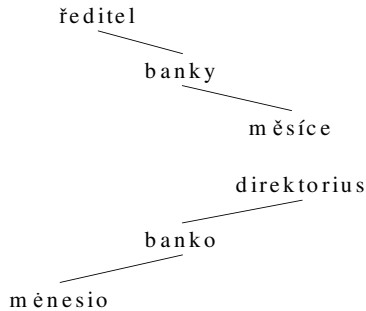
2 paveikslas. Sakinio *k Černému moři* sintaksinis medis

Pirma reikia išversti prielinksni. 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 pažymins: *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 m ěsíce* reiškia „mėnesio banko direktorius“. Žodžių tvarka lietuvių kalboje atvirkščia nei čekų. Abiejų sakinių struktūros yra:



3 paveikslas. Sakinių *ředitel banky m ěsíce* ir *m ěnesio banko direktorius* struktūros

Keli čekų kalbos prielinksniai visai neverčiami į lietuvių kalbą – užtenka tam tikro linksnio formos, pvz., *pro Petra* reiškia „Petru“; *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

7. ĮVERTINIMAS

Vertimo kokybė įvertinta naudojant programą *Trados Translator's Workbench*. Įvertinimas atliktas išvertus tekstą mašiniu (sistema *Česlko*) 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 %

Komercinė sistema *PC Translator* iš čekų kalbos į anglų išversto teksto panašumas, įvertintas tuo pačiu metodu, yra 30 proc. (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 veiksmožodžio forma.** Kadangi veiksmožodžių 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.

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

Priežastis gali būti veiksmažodžio valdymas arba nevysiškai išanalizuota daiktavardžio grupė. Pavyzdžiui, frazė *problém, který nebyl vyřešen* bus išversta **problema, kuris nebuvo išspręsta*, nes dalinė gramatika neanalizuoja šalutinių sakinių.

8. IŠVADOS

Šiame straipsnyje aprašėme eksperimentinę mašininio vertimo sistemą *Česílko* 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 mašininio vertimo iš čekų į lietuvių kalbą rezultatas – 87,6 proc. tikslumas. Žinoma, toks rezultatas nėra tobulas, bet šiuo metu neturime gero sintaksinio čekų kalbos sintaksinio anotatoriaus (geriausias rezultatus pateikia M. Collinso sukurtas statistinis sintaksinis anotatorius (Collins et al. 1999)). Be mašininio vertimo šiame projekte sukurtą dalinę 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 Saarbrückene sukurtą sistemą *SproUT* (Becker et al. 2002; Drozdzyński et al. 2003) gramatikoms rašyti.

LITERATŪRA

- Becker M., Drozdzyński W., Krieger H. U., Piskorski J., Schäfer U. And Xu F., 2002, *SproUT – Shallow processing with typed feature structures and unification*, In *Proceedings of ICON 2002*, Mumbai, India.
- Bresnan J., 2001, *Lexical-functional syntax*, Oxford: Blackwell Publishers.
- 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.
- Colmerauer A., 1969, *Les systèmes Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur*. Montréal: Mimeo.
- Dębowski Ł., 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.
- Dikovský A., Modina L., 1770, *Dependency grammar. Problemy peredači informacii*, Moskva.
- 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.
- Gamut L. T. F., 1991, *Logic, language and meaning 2: Intentional logic and logical grammar*. Chicago: University of Chicago Press.
- Hajič J., 2001, *Disambiguation of rich inflection (computational morphology of Czech)*, Prague: Karolinum, Charles University Press.
- 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.
- Hajič J., Homola P., Kuboň V., 2003 *A simple multilingual machine translation system*. In *Proceedings of the VIII MT Summit*, New Orleans.
- Homola P., 2002, *Machine translation among Slavic languages*. In *Proceedings of the WDS*, Charles University, Prague, 2002.
- 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.
- Homola P., Tolvaj B., 2004, *Distributed translation memories and shallow MT*. Malý infromatický seminář, Josefův důl.
- Kirschner Z., 1987, *APAC3-2: An English-to-Czech machine translation system*. Explizite Beschreibung der Sprache und automatische Textbearbeitung XIII, MFF UK, Prague.
- Kuboň V., 2001, *Problems of robust parsing – PhD thesis*, Faculty of Mathematics and Physics, Charles university, Praha.
- Oliva K., 1989, *A parser for Czech implemented in Systems Q*. Explizite Beschreibung der Sprache und automatische Textbearbeitung XVI, MFF UK, Prague.
- Panevová J., 1980, *Formy a funkce ve stavbě české věty*. Studie a práce lingvistické, Praha: Academia.
- Pecina P., Holub M., 2002, *Sémanticky signifikantní kolokace*. Technical report TR-2002-13, ÚFAL/CKL, Faculty of Mathematics and Physics, Charles University, Praha.
- Sgall P. Et al., 1974, *Úvod do algebraické lingvistiky*. Univerzita Karlova, SNP, Praha.
- 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.
- Žáčková E., 2002, *Parciální syntaktická analýza (češtiny) – PhD thesis*. Fakulta informatiky Masarykovy University, Brno.

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 such project is the system *Česílko* 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, explain similarities and differences between Czech (source language) and Lithuanian and the structure of the translation dictionary. Moreover we explain how we have evaluated translation quality.

APIE AUTORIUŠ

Erika Rimkutė – Vytauto Didžiojo universiteto Lietuvių kalbos katedros doktorantė, Kompiuterinės lingvistikos centro jaunesnioji mokslo darbuotoja (darbovietės adresas: Donelaičio g. 52-206, Kaunas). Mokslinių interesų sritys: tekstynų lingvistika, kompiuterinė lingvistika, automatinė morfologinė analizė bei sintezė, mašininis vertimas.

El. paštas: e.rimkute@hmf.vdu.lt

Petr Homola – Prahos Karlo universiteto Formaliosios ir taikomosios lingvistikos instituto doktorantas ir darbuotojas (darbovietės adresas: Malostranské náměstí 25, Praha). Mokslinių interesų sritys: kompiuterinė lingvistika, automatinė sintaksinė analizė bei sintezė, analitiniai ir tektogramatiniai tekstynai, mašininis vertimas, vertimo atminties sistemos.

El. paštas: homola@ufal.mff.cuni.cz