LIETUVIŲ KALBOS ŽODŽIO DALIŲ ANALIZĖS ALGORITMAS

Saulius Kazilionis, Egidijus Paliulis
Šiaulių universitetas, Technologijos fakultetas

Įvadas

Tarptautiniam bendravimui labai svarbus užsienio kalbų mokėjimas. Todėl šiuo metu pasaulyje itin didelę paklausą turi daugiakalbės automatinio (mašininio) vertimo sistemos, kurių veikimas remiasi skirtingų kalbų gramatikos ir leksikos analize bei jų transformacijomis iš vienos kalbos į kitą. Automatinio vertimo sistemų vertimas pakankamai greitas, tačiau ne visada kokybiškas ir tikslus. Tai susiję su įvairių kalbų tarpusavio gramatiniais skirtumais, kurie sukelia netikslias transformacijas. Lietuvių kalbos gramatika ir leksika sudėtinga ir labai skiriasi nuo kai kurių tarptautiniu mastu vartojamų bendravimo kalbų (pvz., anglų kalbos).

Tyrimo tikslas – sukurti lietuvių kalbos žodžio dalių analizės algoritmą, jį realizuoti ir praktiškai patikrinti.

Uždaviniai:

- Išnagrinėti lietuvių kalbos morfologiją.
- Atlikti egzistuojančių lietuvių kalbos gramatikos ir leksikos analizės sistemų apžvalgą.
- Sukurti lietuvių kalbos žodžio dalių analizės algoritmą.
- Atlikti sukurto algoritmo realizaciją ir praktinį patikrinimą.

Lietuvių kalbos žodžio sandara

Automatinio vertimo sistemose labai svarbi konkrečios kalbos gramatikos ir leksikos analizė. Kuriant žodžio dalių analizės algoritmą, svarbu išstudijuoti tos kalbos morfologiją. Morfologija (*gr. morphe* – forma, *logos* – mokslas) yra gramatikos šaka, nagrinėjanti kalbos dalis (žodžių klases) ir jų sudėtį – kaitybą, darybą [1].

Visi žodžiai sudaryti iš morfemų [2]. Morfema yra paprasčiausias, mažiausias kalbos vienetas, turintis reikšmę. Pavyzdžiui, žodyje *stalelis* yra trys morfemos: *stal-* , *-el-* , *-is*. Visos jos turi vienokią ar kitokią reikšmę: *stal–* sudaro žodžio leksinės reikšmės branduolį, *-el-* rodo, kad žodis turi mažybinę reikšmę, *-is* – kad jis yra vyriškosios giminės vienaskaitos vardininko linksnio. Lietuvių kalbos morfemos pagal jų funkcijas skirstomos [1, 2, 3]:

- šaknis, arba šakninė morfema,
- afiksas, arba afiksinė morfema.

Afiksas (*lot. affixus* – pritvirtintas) yra reikšminė žodžio dalis (morfema), išskyrus šaknį (pvz., priešdėlis, priesaga, intarpas, galūnė ir pan.) [1, 2, 3].

Morfema negali būti skaidoma į smulkesnius vienetus, nepažeidžiant jos reikšmės [1, 2]. Analizuojant žodį morfemiškai, svarbu išsiaiškinti, ar jis skaidomas į morfemas, ar neskaidomas, ir jei skaidomas, – kurios morfemos jį sudaro, kaip jos pasiskirsčiusios. Morfemiškai skaidomi tokie žodžiai, kurių atskiros morfemos, turinčios tą pačią reikšmę, pasikartoja kituose žodžiuose. Skaidomas žodis gretinamas su bendrašakniais žodžiais ir su tais, kurie turi tuos pačius darybos ar kaitybos afiksus. Pvz.: *išvežimas* skaidomas į tokias morfemas: *vež-* šakninė morfema, pasikartojanti žodžiuose *vež-ti*, *vež-imas*, *iš-* priešdėlis, su kuriuo gali būti sudaryta daugybė priešdėlinių veiksmažodžių (*iš-nešti*, *iš-vykti*), *-im-* priesaga, būdinga daugeliui lietuvių kalbos daiktavardžių, pavadinančių veiksmą (*draud-im-as*, *lėk-im-as*). Galūnė *-as* rodo vyriškosios giminės vienaskaitos vardininką.

Visi kaitomieji lietuvių kalbos žodžiai yra morfemiškai skaidomi [1, 2], nes juose galima aiškiai išskirti kaitybos morfemas (*mišk-as*, *ger-as*, *vien-as*). Morfemiškai skaidomi ir kai kurie nekaitomi žodžiai [2]. Juose galima išskirti kituose žodžiuose pasikartojančias šaknines ir darybines morfemas (*vis-ada*, *kit-ur*). Morfemiškai neskaidomais vadinami tokie žodžiai, kuriuos sudaro viena morfema.

Sunkesnis uždavinys yra morfeminis kamieno skaidymas [1]. Žodžio kamienas laikytinas skaidomu tada, kai jo morfemos pasikartoja kituose kamienuose. Žodžio kamieną gali sudaryti: šaknis ir priesaga (*alksn-yn-as*), šaknis ir dvi priesagos (*plėš-ik-av-o*), šaknis ir trys priesagos (*penk-et-uk-inink-as*), priešdėlis ir šaknis (*pa-mišk-ė*), priešdėlis, šaknis ir priesaga (*per-vež-im-as*), priešdėlis, šaknis ir dvi priesagos (*iš-aug-in-im-as*), priešdėlis sangrąžinė morfema ir šaknis (*iš-si-praus-ti*), priešdėlis, sangrąžinė morfema, šaknis ir priesaga (*iš-si-praus-im-as*), dvi šaknys (*juod-alksn-is*), dvi šaknys ir jungiamoji morfema (*saul-ė-lyd-is*), dvi šaknys ir priesaga (*žmog-žudž-iau-ti*), dvi šaknys ir dvi priesagos (*nak-v-yn-pinig-iai*) ir t. t. Skaidomais kamienais laikomi ir tie, kurių viena morfema (ar kelios) su ta pačia reikšme pasikartoja kituose žodžiuose, o viena yra labai reta, arba visiškai unikali [1, 2]. Toks kamienų skaidumas vadinamas daliniu. Unikali gali būti arba šaknis arba afiksas. Morfemiškai neskaidomais laikomi tokie kamienai, kurių visos numatomos išskirti morfemos yra unikalios.

Panašių sistemų apžvalga

Panašiai veikiančios sistemos – tai automatinio (mašininio) vertimo sistemos. Automatinio vertimo sistemų (toliau – AVS) kūrimui reikalingas kelių sričių išmanymas – kompiuterinės ir matematinės lingvistikos, vertimo teorijos, informacinių technologijų, kalbos filosofijos. AVS vartojamos tada, kai reikalingas greitas, nors ne visada kokybiškas ir tikslus vertimas.

Internetu pasiekiamų AVS skaičius yra daugiau nei 30. Tačiau jų veikimo principai nėra visiškai skirtingi, turi panašumų. Pvz., *Babelfish*, *Apple*, *Worldlingo*, *Yahoo!*, *Google* vertimo portalai yra palaikomi *Systran* vertimo sistemos. Interaktyvi rusų – anglų programa *MagicGooddie* yra senos rusų kompanijos *Promt* produktas. Viena pirmųjų AVS, dirbančių su lietuvišku tekstu, yra VDU projektas (*Anglų – lietuvių mašininio vertimo sistema*), taip pat sukurtas *Promt* kompanijos sistemos pagrindu, kuris anksčiau buvo skirtas rusų – anglų kalbų krypties teksto vertimui [6].

Aptarsime keletą panašiai veikiančių sistemų. Šiuo metu didžiausia mašininio vertimo sistema yra *Google Translator* [5]. Sistemos galimybės:

- Greitas vertimas į 57 skirtingas kalbas ir iš jų.
- Verčiami žodžiai, sakiniai, dokumentai ir tinklalapiai iš bet kurios palaikomos kalbos ir į bet kurią palaikomą kalbą.
- Išversto žodžio įgarsinimas, norimo išversti žodžio pasiūlymas.
- Tinklalapio vertimas vienu pelės mygtuko spustelėjimu.

Realizacija: sprendžiant iš adreso struktūros, ši sistema yra realizuota *PHP* programavimo kalba kartu su *JavaScript*.

Lietuviška anglų – lietuvių mašininio vertimo sistema [6]. Galimybės:

- Rišlus teksto vertimas, atsižvelgiant į teksto morfologiją, sintaksę ir semantiką.
- *DOC*, *RTF, HTML* ir *TXT* formatu pateiktų dokumentų vertimas.
- Vertimo funkcijos įtraukimas į *Microsoft Word*, *Internet Explorer* ir *Mozilla Firefox* (pasitelkiant atitinkamus papildinius).
- Tekstų vertimas bet kurioje *Windows* taikomojoje programoje (pasitelkiant specialias taikomąsias programas).
- Vertimo kokybės gerinimas, panaudojant papildomus specializuotus žodynus ir temų šablonus;
- Vertimų statistikos peržiūra.

Didžiausias šios vertimo sistemos trūkumas yra tas, kad ji verčia žodžius tik iš anglų kalbos.

Realizacija: Ši sistema yra realizuota *ASP. NET* tinklalapio struktūros technologijos pagrindu.

Morfologinis lietuvių kalbos anotatorius [7]. Galimybės:

- Įvesto žodžio ar sakinio pilna morfologinė analizė.
- Nemažas nustatymų pasirinkimas.
- Galimybė rezultatus išsaugoti tekstiniame faile (= rinkmenoje).

Veikimas: pritaikius statistinius modelius ir panaudojus 1 mln. žodžių pusiau automatiškai parengtą morfologiškai anotuotą tekstyną, sukurtas vienareikšminimo įrankis, kurio tikslumas apie 94 procentai. Antraštinių lietuvių kalbos žodžių formų nustatymo tikslumas netgi 99 procentai.

Realizacija: *PHP* programavimo kalba kartu su *JavaScript*.

1 lentelė. ***Analogiškų sistemų palyginimas***

	Google vertėjas [5] <http://translate.google.lt>	Teksto vertimas [6] <vertimas.vdu.lt>	*Tildė* biuras <www.tilde.lt>
Nemokama	taip	taip	ne
Žodžio vertimas	yra	yra	yra
Sakinio vertimas	yra	yra	yra
Galimybė atsispausdinti išverstą tekstą	yra	yra	yra
Dokumento vertimas	yra	nėra	nėra
Išsami pagalba	nėra	yra	yra
Internetinio puslapio vertimas	yra	yra	nėra

***PROMT* vertimo sistema**

Promt sistemos vertimo algoritmai paremti ne nuosekliomis transformavimo procedūromis, bet hierarchiniu pagrindu, kuriame vertimo procesai suskaidyti į susisiekiančias transformavimo procedūras skirtinguose analizės etapuose [8].

Sistemą galima išskirstyti į tokius lygmenis:

Leksikos vienetų lygmuo. Leksikinis vienetas yra žodis, arba kolokacija, priklausanti žemiausiam lygiui. Žodis aprašytas kaip kamieno ir galūnės kompozicija. Iš vienos pusės, tai leidžia atpažinti pirmos kalbos (toliau – PK) žodį ir morfologiškai jį nagrinėti, iš kitos pusės, patogus kelias nustatyti vertimui pagal susijusius morfologinius duomenis (kamienas, kaitymo tipas, ir galūnės adresas tokios kaitybos tipo galūnių masyve). Taigi, jei galima pritaikyti keitimo taisykles iš PK morfologinių duomenų į antros kalbos (toliau – AK) morfologinius duomenis, tada

galima vykdyti transformavimo procedūras morfologiniame lygmenyje.

Grupių lygmuo. Jis atsako už sudėtingesnes struktūras: daiktavardžių grupes, būdvardžius, prieveiksmius ir sudėtines veiksmažodžių formas. Šio lygmens pagrindas yra formalios ryšių gramatikos. Analizės metu tai leidžia jungti grupes į sintaksinius vienetus. Kiekvieną vienetą charakterizuoja susintezuoti struktūriniai duomenys ir pagrindinis jungimo vienetas.

Paprasto sakinio lygmuo. Paprastasis sakinys traktuojamas kaip struktūra, susidedanti iš sintaksinių vienetų. Jo analizė vykdoma pagal karkasines tarinio struktūras. Paprastuose sakiniuose, pagrindinis elementas yra veiksmažodis, o jo junglumas (valentingumas) nulemia aktyvaus karkaso užpildymą.

Sudėtinio sakinio lygmuo. Analizė reikalinga, kai reikia suderinti laikus ir teisingai išversti jungtukus.

Šie procesai tarpusavyje siejasi pagal tekstinio vieneto hierarchiškumą, keičiasi susintezuotais ir paveldėtais atributais. Toks algoritmo sudarymas leidžia panaudoti formaliuosius metodus algoritmams aprašyti skirtinguose lygiuose.

SYSTRAN vertimo sistema

Iš pradžių sistema suprojektuota tiktai vertimui iš rusų kalbos į anglų kalbą, dabar apima 80 kalbų porų, verčia iš 22 kalbų ir yra daugelio didžiųjų portalų variklis [8]. *Systran* sistema tradiciškai traktuojama kaip taisyklėmis pagrįsta sistema. Projektas pradėtas kurti daugiau nei prieš 30 metų. Pirmieji vartotojai – JAV vyriausybinės organizacijos, o po to ir Europos ekonomikos sąjunga, kuri *Systran* pertvarkė ir gausios dokumentacijos vertimas tapo įmanomas į daugelį jai priklausančių šalių.

Systran toliau tobulinamas įtraukiant ir statistinius metodus. Čia aprašomi formalieji metodai ir vertimo algoritmas, nesusijęs su statistiniais metodais. *Systran* sistemos dizainas yra agreguotas ir aukšto modalumo. Jame yra dviejų tipų programos:

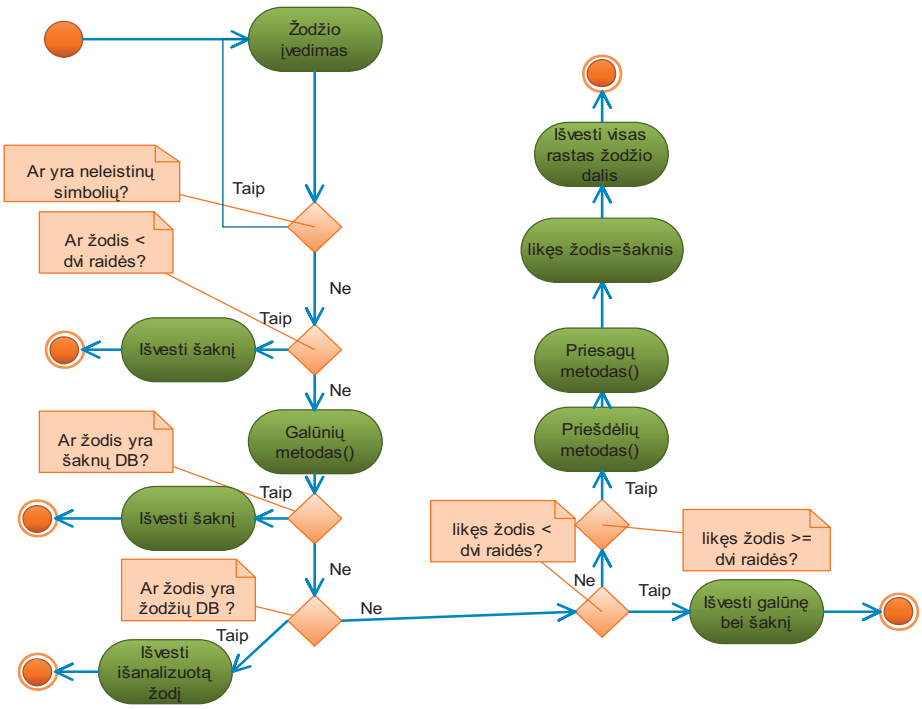
Sisteminės, aprašytos asemblerio kodu, nepriklausančios nuo kalbos; jos, pavyzdžiui, atsakingos už žodyno peržiūros paprogrames.

Vertimo programos, sudarytos iš daug atskirų modulių. Vertimo programos skirtos analizei ir generavimui. Analizės modulis pirmai kalbai yra pastovus nepriklausomai nuo antros kalbos, o generavimo moduliai yra pastovūs antrai kalbai, nesvarbu, kokia bebūtų pirma kalba.

Pagrindinė sistemos dalis – didžiulis dvikalbis žodynas, talpinantis leksikos ekvivalentus, gramatikos ir semantikos informaciją, vartojamą analizei ir generacijai. Didelė šios informacijos dalis yra algoritmų formos, jie iškviečiami įvairiuose vertimo proceso etapuose. Pagrindiniai vertimo procesai valdomi sudėtingo dvikalbio žodyno.

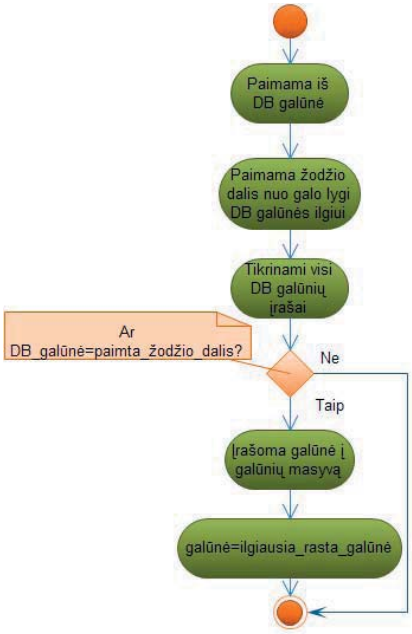
Lietuvių kalbos žodžio analizės algoritmas

Atlikus lietuviško žodžio dalių analizę ir išnagrinėjus panašias sistemas, buvo sudarytas lietuvių kalbos žodžio dalių analizės algoritmas. Diagramoje (1 pav.) pateiktas bendras algoritmo veikimas.



1 pav. Bendras žodžio analizės algoritmas

Įvedus žodį, pirmiausia tikrinama, ar nėra draudžiamų simbolių (leidžiamos tik raidės ir tarpai). Tikrinama, ar žodis nėra trumpesnis už dvi raides. Jei taip yra, visas žodis yra šaknis, rezultatai išvedami į ekraną. Jei ne, tikrinama, ar žodis yra tarp DB įvestų šaknų, ir jei taip, visas žodis yra šaknis, rezultatai išvedami į ekraną. Jei žodis nebuvo rastas tarp DB šaknų, kviečiamas galūnių radimo metodas (2 pav.). Po to tikrinama, ar likęs žodis yra tarp DB įvestų, jau išanalizuotų žodžių. Jei yra, išvedamas išanalizuotas žodis su rasta galūne. Jei nėra, tikrinama, ar likęs žodis yra trumpesnis už dvi raides. Jei trumpesnis, likusi žodžio dalis yra šaknis, išvedama šaknis bei rasta galūnė. Jei ne, kviečiami galūnių (2 pav.), priešdėlių ir priesagų radimo metodai (žr. 3 pav.). Įvykdžius šiuos metodus, likusi žodžio dalis priskiriama šakniai. Pabaigoje išvedamos visos rastos žodžio dalys.

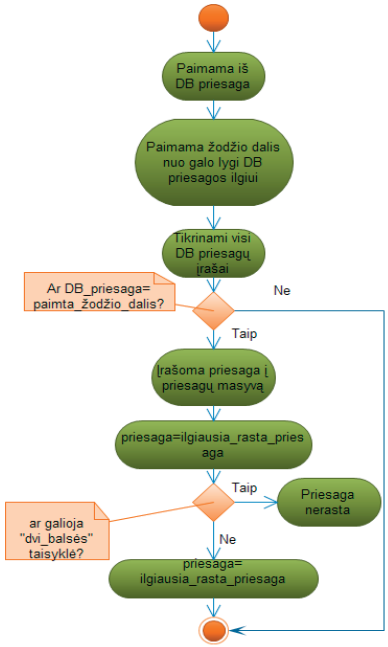


2 pav. Galūnių analizė algortimas

Detaliau aprašysime galūnės radimo algoritmą (2 pav.). Paieška atliekama taip:

- Nuskaitoma galūnė iš duomenų bazės.
- Paima tiek raidžių nuo žodžio galo, kiek turi galūnė, esanti duomenų bazėje;
- Tikrinama, ar paimta žodžio dalis yra tapati galūnei, paimtai iš DB. Jei taip, įrašoma į galūnių masyvą.
- Taip patikrinamos visos galūnės, esančios duomenų bazėje.
- Jei buvo rastos kelios galūnės, tikroji bus ta, kuri sudaryta iš daugiau raidžių.

Priesagų analizės algoritmas (3 pav.) analogiškas galūnių analizei – tik pabaigoje, radus priesagą, dar patikrinama, ar rastos priesagos pirma raidė ir likusio žodžio (atmetus rastą priesagą) paskutinė raidė nėra balsės. Jei nėra balsės, tada tai yra tikroji įvesto žodžio priesaga.



3 pav. Priesagų analizė algortimas

Tyrimo rezultatai

Algoritmo veiksmingumui išsiaiškinti buvo sukurta sistema ir atlikta įvairių lietuvių kalbos žodžių dalių analizė. Tikrinti daiktavardžiai, būdvardžiai, veiksmažodžiai, įvardžiai, skaitvardžiai, prieveiksmiai, prielinksniai. 2 lentelėje parodyta, kaip sistema apdorojo skirtingų lietuvių kalbos žodžių dalis. Brūkšnelis šalia morfemos pažymi kitos morfemos vietą.

2 lentelė. ***Žodžio dalių analizės tikrinimo rezultatai***

Reikalavimai / tikslai	**Įvykis / įvestis**	**Laukiamas rezultatas**	**Gautas rezultatas**	**T/N**	**Pastabos**
Daiktavardžių analizė	Analizuojamas žodis *dangus*	Šaknis *dang-* Galūnė *-us*	Šaknis *dang-* Galūnė *-us*	T	
	Analizuojamas žodis *namelis*	Šaknis *nam-* Priesaga *-el-* Galūnė *-is*	Šaknis *nam-* Priesaga *-el-* Galūnė *-is*	T	
	Analizuojamas žodis *kaimas*	Šaknis *kaim-* Galūnė *-as*	Šaknis *kaim-* Galūnė *-as*	T	

2 lentelės tęsinys

Daiktavardžių analizė	Analizuojamas žodis *trintukas*	Šaknis *trin-* Priesaga *-tuk-* Galūnė *-as*	Šaknis *trin-* Priesaga *-tuk-* Galūnė *-as*	T	
	Analizuojamas žodis *kalnelis*	Šaknis *kaln-* Priesaga *-el-* Galūnė *-is*	Šaknis *kaln-* Priesaga *-el-* Galūnė *-is*	T	
Būdvardžių analizė	Analizuojamas žodis *baltas*	Šaknis *bal-* Priesaga *-t-* Galūnė *-as*	Šaknis *bal-* Priesaga *-t-* Galūnė *-as*	T	
	Analizuojamas žodis *svarbus*	Šaknis *svar-* Priesaga *-b-* Galūnė *-us*	Šaknis *-svar-* Priesaga *-b-* Galūnė *-us*	T	
	Analizuojamas žodis *sraunus*	Šaknis *srau-* Priesaga *-n-* Galūnė *-us*	Šaknis *srau-* Priesaga *-n-* Galūnė *-us*	T	
	Analizuojamas žodis *didelis*	Šaknis *did-* Priesaga *-el-* Galūnė *-is*	Šaknis *did-* Priesaga *-el-* Galūnė *-is*	T	
	Analizuojamas žodis *aukštas*	Šaknis *auk-* Priesaga *-št-* Galūnė *-as*	Šaknis *auk-* Priesaga *-št-* Galūnė *-as*	T	
Veiksmažodžių analizė	Analizuojamas žodis *bėgti*	Šaknis *bėg-* Priesaga *-ti-*	Šaknis *bėg-* Priesaga *-ti-*	T	
	Analizuojamas žodis *miegoti*	Šaknis *mieg-* Priesaga *-o-* Priesaga *-ti*	Šaknis *miego-* Priesaga *-ti-*	N	Priesaga *-o-* neįvesta į duomenų bazę
	Analizuojamas žodis *pereiti*	Priešdėlis *per-* Šaknis *-ei-* Priesaga *-ti*	Priešdėlis *per-* Šaknis *-ei-* Priesaga *-ti*	T	
	Analizuojamas žodis *nuvažiavo*	Priešdėlis *nu-* Šaknis *-važ-* Priesaga *-iav-* Galūnė *-o*	Priešdėlis *nu-* Šaknis *-važ-* Priesaga *-iav-* Galūnė *-o*	T	
	Analizuojamas žodis *nešdamas*	Šaknis *neš-* Priesaga *-dam-* Galūnė *-as*	Šaknis *neš-* Priesaga *-dam-* Galūnė *-as*	T	
Įvardžių analizė	Analizuojamas žodis *jis*	Šaknis *j-* Galūnė *-is*	Šaknis *j-* Galūnė *-is*	T	
	Analizuojamas žodis *kas*	Šaknis *k-* Galūnė *-as*	Šaknis *k-* Galūnė *-as*	T	
	Analizuojamas žodis *kuris*	Šaknis *kur-* Galūnė *-is*	Šaknis *kur-* Galūnė *-is*	T	
	Analizuojamas žodis *mano*	Šaknis *man-* Galūnė *-o*	Šaknis *man-* Galūnė *-o*	T	
	Analizuojamas žodis *patiems*	Šaknis *pat-* Galūnė *-iems*	Šaknis *pat-* Galūnė *-iems*	T	
Skaitvardžių analizė	Analizuojamas žodis *vienas*	Šaknis *vien-* Galūnė *-as*	Šaknis *vien-* Galūnė *-as*	T	
	Analizuojamas žodis *dveji*	Šaknis *dv-* Priesaga *-ej-* Galūnė *-i*	Šaknis *dvej-* Galūnė *-i*	N	Taisyklės „šaknis dvi priebalsės“ išimtis
Skaitvardžių analizė	Analizuojamas žodis *penktas*	Šaknis *penk-* Priesaga *-t-* Galūnė *-as*	Šaknis *penk-* Priesaga *-t-* Galūnė *-as*	T	
	Analizuojamas žodis *ketveri*	Šaknis *ket-* Priesaga *-v-* Priesaga *-er-* Galūnė *-i*	Šaknis *ket-* Priesaga *-v-* Priesaga *-er-* Galūnė *-i*	T	

2 lentelės tęsinys

Skaitvardžių analizė	Analizuojamas žodis *penketas*	Šaknis *penk-* Priesaga *-et-* Galūnė *-as*	Šaknis *penk-* Priesaga *-et-* Galūnė *-as*	T	
Prieveiksmių analizė	Analizuojamas žodis *puikiai*	Šaknis *puik-* Priesaga *-iai*	Šaknis *puik-* Galūnė *-iai*.	N	
	Analizuojamas žodis *svaiginamai*	Šaknis *svaig-* Priesaga *-in-* Priesaga *-am-* Priesaga *-ai*	Šaknis *svaiginam-* Galūnė *-ai*	N	Neskiriama galūnė nuo priesagos, jei jos tapačios
	Analizuojamas žodis *nesunkiai*	Priešdėlis *ne-* Šaknis *-sunk-* Priesaga *-iai*	Priešdėlis *ne-* Šaknis *-sunk-* Galūnė *-iai*	N	Neskiriama galūnė nuo priesagos, jei jos tapačios
	Analizuojamas žodis *jaunai*	Šaknis *jaun-* Priesaga *-ai*	Šaknis *jaun-* Galūnė *-ai*	N	Neskiriama galūnė nuo priesagos, jei jos tapačios
	Analizuojamas žodis *švariai*	Šaknis *švar-* Priesaga *-iai*	Šaknis *šavar-* Galūnė *-iai*	N	Neskiriama galūnė nuo priesagos, jei jos tapačios
Prielinksnių analizė	Analizuojamas žodis *per*	Šaknis *per*	Šaknis *per*	T	
	Analizuojamas žodis *už*	Šaknis *už*	Šaknis *už*	T	
	Analizuojamas žodis *prieš*	Šaknis *prieš*	Šaknis *prieš*	T	
	Analizuojamas žodis *ant*	Šaknis *ant*	Šaknis *ant*	T	
	Analizuojamas žodis *su*	Šaknis *su*	Šaknis *su*	T	

Sistema gerai analizuoja daiktavardžius ir būdvardžius, sunkiau sekasi su veiksmažodžiais, įvardžiais ir skaitvardžiais, o prieveiksmių analizė nėra tinkama. Taip yra dėl to, kad sistema nesugeba nustatyti, kuri kalbos dalis yra žodis. Be to, ne visos žodžių dalys yra suvestos į duomenų bazę.

Išvados

1. Atlikta lietuvių kalbos žodžių analizė ir nustatyta, kad tai – sudėtingas mokslas: norint nustatyti žodžio „sudėtį“, reikia žinoti jo kilmę, o tai ne visada įmanoma išsiaiškinti.
2. Atlikta analogiškų sistemų analizė ir nustatyta, kad panašiai veikiančios sistemos – automatinio vertimo sistemos, kurios taip pat analizuoja žodžius, netgi nustato žodžio galūnę, tačiau šios analizės rezultatų vartotojams neišveda. Artimiausias analogas sukurtai sistemai – morfologinis anotatorius [7], analizuojantis žodžius – tik platesniu, morfologiniu aspektu.
3. Išsiaiškinta lietuvių kalbos žodžių sandara ir sukurtas lietuviškų žodžių analizės algoritmas bei sistema. Sistema naudoja duomenų bazę, kurioje saugomos žodžio dalys. Kadangi lietuvių kalboje šaknų aibė yra gausiausia ir ją sudėtinga nustatyti, todėl buvo nuspręsta į duomenų bazę įvesti tik tas šaknis, kurios sudaro visą žodį. Šaknų duomenų bazė nuolatos papildoma atlikus žodžių analizę.
4. Sukurta sistema geriausiai analizuoja daiktavardžius, būdvardžius, prasčiau – veiksmažodžius, įvardžius ir skaitvardžius, o prieveiksmių analizė dažniausiai klaidinga. Taip yra dėl to, jog sistema neskiria, kuriai kalbos daliai priklauso įvestas žodis.
5. Sistemos rezultatams gerinti sukurta galimybė įvesti į DB jau išanalizuotą žodį. Siekiant dar geresnių rezultatų, sistemą reikėtų susieti su anksčiau minėtu morfologiniu anotatoriumi, kuris padėtų nustatyti kalbos dalį, giminę, skaičių, linksnį ir t. t. Tada būtų galima patobulinti algoritmą.

Literatūra

1. Ružė A., 2008, *Lietuvių kalbos morfologija* (I dalis). Vilniaus universitetas.
2. Paulauskienė A., 2006, *Lietuvių kalbos morfologijos pagrindai*. Kaunas.
3. Rimkutė E., 2006, Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne. *Daktaro disertacija*. Kaunas: VDU Lietuvių kalbos institutas.
4. Keinys S., 1999, *Bendrinės lietuvių kalbos žodžių daryba*. Šiauliai: VšĮ Šiaulių universiteto leidykla.
5. Google automatinio vertimo sistema [interaktyvus] [žiūrėta 2011–12–02]. Prieiga per internetą: <http://translate.google.lt>.
6. Anglų–lietuvių kalbos vertimo sistema [interaktyvus] [žiūrėta 2011–12–02]. Prieiga per internetą: <http://vertimas.vdu.lt>.

7. Morfologinis anotatorius internete [interaktyvus] [žiūrėta 2011–12–02]. Prieiga per internetą: <http://donelaitis.vdu.lt/>.

8. Paliulis E., Milisevičiūtė D., 2009, Lietuvių – anglų kalbų vertimo sistema. *Jaunųjų mokslininkų darbai*. Nr. 2 (23). P. 51–56. Šiauliai: VšĮ Šiaulių universiteto leidykla.

ALGORITHM FOR ANALYSIS OF PARTS OF LITHUANIAN WORD

Saulius Kazilionis, Egidijus Paliulis

Summary

Multilingual automatic (machine) translation systems are now in great demand in the world. These systems are based on analysis of grammar and lexis of different languages, and their conversions from one language into another. Translations produced by automatic translation systems are obtained fast, but they do not always have quality and accuracy. Lithuanian grammar and lexis are very complex and very different from those of some of the internationally used languages (e.g., English).

This morphology of the Lithuanian language was analyzed and the algorithm for analysis of Lithuanian words and their parts was created in this study. The algorithm was used in the system for analysis of Lithuanian word parts. The system was tested. The designed system performs best in analysing nouns and adjectives, worse in analysing verbs, pronouns and numerals, but analysis of adverbs is usually erroneous. When developing the algorithm in future it is necessary for the analysis of word parts to be preceded by identification of part of speech of the word under analysis.

Keywords: morphology of Lithuanian language, analysis of word parts, algorithm for analysis of words.

LIETUVIŲ KALBOS ŽODŽIO DALIŲ ANALIZĖS ALGORITMAS

Saulius Kazilionis, Egidijus Paliulis

Santrauka

Šiuo metu pasaulyje labai didelę paklausą turi daugiakalbės automatinio (mašininio) vertimo sistemos. Jų veikimas remiasi skirtingų kalbų gramatikos ir leksikos analize bei jų transformacijomis iš vienos kalbos į kitą. Automatinio vertimo sistemų vertimas pakankamai greitas, tačiau ne visada kokybiškas ir tikslus. Lietuvių kalbos gramatika ir leksika sudėtinga ir labai skiriasi nuo kai kurių tarptautiniu mastu vartojamų bendravimo kalbų (pvz., anglų kalbos).

Šiame darbe nagrinėjama lietuvių kalbos morfologija ir kuriamas lietuviškų žodžių ir jo dalių analizės algoritmas. Sukurtas algoritmas panaudotas žodžio dalių analizės sistemoje. Atliktas sistemos patikrinimas. Sukurta sistema geriausiai analizuoja daiktavardžius, būdvardžius, prasčiau – veiksmažodžius, įvardžius ir skaitvardžius, o prieveiksmių analizė dažniausiai klaidinga. Ateityje tobulinant algoritmą, būtina pirmiausia nustatyti, kuriai kalbos daliai priklauso nagrinėjamas žodis, po to atlikti analizę žodžio dalimis.

Prasminiai žodžiai: lietuvių kalbos morfologija, žodžio dalių analizė, žodžio analizės algoritmas.

Įteikta 2011-12-02