Znotiņa, Inga. Otrās baltu valodas apguvēju korpusa morfoloģiska anotēšana. *Via scientiarum* : starptautiskās jauno lingvistu konferences rakstu krājums. 3. laidiens. Sastādītājas I. Laizāne, I. Znotiņa. Ventspils, Liepāja : Ventspils Augstskola, Liepājas Universitāte, 2016, 148.–160. lpp.

Inga ZNOTIŅA (Liepājas Universitāte, Rīgas Stradiņa universitāte)

OTRĀS BALTU VALODAS APGUVĒJU KORPUSA MORFOLOĢISKA ANOTĒŠANA

Ievads

Viena no korpuslingvistikas metožu galvenajām priekšrocībām ir datora spēja liela apjoma datu kopumā ātri atrast noteiktus elementus un veikt ar tiem saistītos aprēķinus. Taču, tā kā dators jebkādus datus uztver kā simbolu virkni, ierīce nespēj tos interpretēt, kā to spēj cilvēks. Tāpēc arī iespējas apstrādāt tekstu korpusu ir ierobežotas, ja vien netiek pievienota papildu informācija, kas datoram ļauj atpazīt noteiktas parādības arī tad, ja tās ir grūtāk nosakāmas vai formāli neatšķiras.

Šāda papildinformācija var būt divējāda: marķēšana (angļu val. *markup*) un anotēšana (angļu val. *annotation*). Termini *marķēšana* un *anotēšana* Latvijā dažkārt tiek lietoti sinonīmiski, taču šajā rakstā tie tiek šķirti, vadoties pēc angļu valodniecībā nostiprinājušās tradīcijas: par *marķēšanu* uzskatāma meta-informācijas pievienošana tekstam (piem., teksta valoda, autora dzimtā valoda utt.), turpretim *anotēšana* ir teksta interpretēšana – tā ir lingvistiskas analīzes rezultātu (piem., vārda sastāva, palīgteikumu veida) pievienošana korpusā iekļautajam tekstam (par šo šķīrumu sk. McEnery, Hardie 2012, 29). Šajā rakstā par marķēšanu sīkāk netiek runāts, uzmanība pievērsta tikai anotēšanai.

Datus (korpuslingvistikā – korpusu) var anotēt pēc dažādiem principiem. Tāpat kā lingvistisku analīzi var veikt dažādos valodas līmeņos, arī tekstus var anotēt atbilstoši šiem līmeņiem. Anotējot korpusa datus, tiek izmantota tagu (angļu val. – *tag*) sistēma atbilstoši veiktajai analīzei. Jebkādas lingvistiskas analīzes, kas paredz veikt kādu klasifikāciju, rezultāti ir transformējami anotējumā. Klasifikācijai nav jābūt pilnai un visaptverošai: piem., ja attiecīgajā korpusā paredzēts pētīt darbības vārdus, tad arī var anotēt tikai tos, nevis visas vārdšķiras – tā var ietaupīt resursus, ko patērē anotēšana. Tātad par klasifikāciju ir uzskatāma arī opozīcija „piemīt kāda pazīme” : „nepiemīt šī pazīme” un dažādas tās variācijas – piem., „vārds ar priedēkli” : „vārds bez priedēkļa”. Valodas elementi, kuriem attiecīgā pazīme nepiemīt, var netikt anotēti – sal.:

1. <bez pried>***Blakus*** <ar pried>***sapucētajai*** <bez pried>***mājai*** <ar pried>***neglīti*** <ar pried>***izskatījās*** <bez pried>***maza*** <bez pried>***būdiņa***[1]**.**
2. ***Blakus*** <ar pried>***sapucētajai mājai*** <ar pried>***neglīti*** <ar pried>***izskatījās maza būdiņa***.

Balstoties valodas līmeņu principā, ļoti populāra korpuslingvistikā ir morfoloģiskā anotēšana. Šim anotēšanas veidam un īpaši tā izmantojumam otrās baltu valodas apguvēju korpusā ir veltīts šis raksts. Tas aplūko morfoloģisko anotēšanu korpuslingvistikā un min, kas jāņem vērā, šādi anotējot valodas apguvēju korpusus. Izrietot no minētā, tiek skaidrota morfoloģiskās anotēšanas paveidu izvēle otrās baltu valodas apguvēju korpusam.

1. Morfoloģiskās anotēšanas paveidi

Morfoloģiskā anotēšana balstās morfoloģijā kā valodniecības apakšnozarē, tāpēc, lai noskaidrotu, pēc kādiem principiem šī anotēšana tiek veikta, jāzina, ko aplūko morfoloģija. Saskaņā ar valodas līmeņu teoriju morfoloģijas līmenis ir „valodas apakšsistēma, kuru veido morfēmu un gramatisko formu kopums to savstarpējās attieksmēs" (VPSV 2007, 241). Savukārt morfoloģija ir „valodniecības apakšnozare, gramatikas daļa, kurā pēta valodas morfoloģisko sistēmu – morfēmu struktūru un funkcijas, gramatiskās kategorijas, gramatisko formu veidošanu" (turpat, 240–241). Tiesa, morfoloģijas robežas nav īsti skaidri nosakāmas. Tradicionāli pieņemts, ka morfoloģijā ietilpst arī vārddarināšana un morfēmika (MLLVG 1959, 75–76; Kalnača 2004, 8), tomēr par šo jautājumu tiek diskutēts (sk. Urbutis 1978, 35–39; Miliūnaitė 2003, 20–21). Kamēr daži autori min, ka tās var izdalīt arī atsevišķi atkarībā no pētījumu rakstura (Miliūnaitė 2003, 21), citi uzskata, ka gan vārddarināšana, gan morfēmika būtu skatāmas kā patstāvīgas valodniecības apakšnozares līdzās morfoloģijai (sk. LVG 2013, 138, 190, 300). Interesanti, ka, piem., LVG minēts, ka „iezīmējas arī tendence nošķirt morfēmiku kā patstāvīgu valodniecības apakšnozari", kā piemēru minot Andras Kalnačas darbu „Morfēmika un morfonoloģija", taču minētajā grāmatā – gluži pretēji – uzsvērts, ka „Mūsdienu lingvistikā morfēmika **netiek** [izcēlums mans – I. Z.] uzskatīta par patstāvīgu valodniecības nozari, tā ietilpst morfoloģijā kā īpaša tās daļa [..]" (Kalnača 2004, 8). Šī raksta nolūks nav iesaistīties diskusijā, tāpēc šeit ievērots tradicionālais iedalījums, proti, vārddarināšana un morfēmika ir uztvertas par morfoloģijas daļām. Toties kā

[1] Šeit un turpmāk piemēros sniegti vienkāršoti tagi, lai piemērs būtu pēc iespējas vieglāk saprotams. Šie tagi neietilpst nekādā esošā sistēmā. Visi tagi attēloti stūrainajās iekavās: <tags>. Tā kā mērķis ir tikai anotēšanas principu ilustrēšana, nolemts neizmantot esošās sistēmas, kurās lietoto tagu nozīme būtu atsevišķi jāskaidro.

morfoloģijas daļa netiek aplūkota morfonoloģija, kurai latviešu un lietuviešu valodniecībā līdz šim nav skaidras piederības fonētikai, morfoloģijai vai savam atsevišķam valodas līmenim.

Korpuslingvistikā morfoloģiskajai anotēšanai tāpat nav skaidru robežu, tās ir atkarīgas no izpratnes un korpusa valodas. Piem., seno portugāļu valodas tekstu korpuss tiek raksturots kā morfoloģiski anotēts, piebilstot, ka anotējumā ir šķirtas: 1) vārdšķiras; 2) gramatiskās kategorijas, piem., dzimte un skaitlis; 3) diakritiskās zīmes; 4) interpunkcija (Britto u. c. 1999). Atbilstoši latviešu valodas aprakstiem tikai pirmie divi šķīrumi ietilptu morfoloģiskajā anotējumā. Savukārt Silviāne Greindžere (*Sylviane Granger*) kļūdu anotēšanai franču valodas apguvēju korpusā piedāvā klasifikāciju, kurā starp citām kļūdu grupām ir arī trīs atsevišķas grupas: morfoloģijas kļūdas, gramatikas kļūdas un sintakses kļūdas. Tādas kategorijas kā dzimte, skaitlis, persona u. c. šeit nav iekļautas morfoloģijā, bet gan atsevišķi gramatikā (Granger 2003, 468). Latviešu valodnieki savukārt morfoloģiju un sintaksi uzskata par gramatikas daļām (sk., piem., Kalme, Smiltniece 2001, 4), tāpēc šāds dalījums nebūtu lietojams. Lai arī S. Greindžeres piedāvātais nav morfoloģiskas anotēšanas, bet gan kļūdu anotēšanas kategoriju kopums, tas labi ilustrē dažādību uzskatos par to, kas būtu anotējams kā morfoloģiska parādība.

Atbilstoši morfoloģijas aptvertajam parādību lokam, kā to raksturo latviešu un lietuviešu valodnieki, morfoloģiskajai anotēšanai var tikt izdalīti šādi paveidi: morfēmu anotēšana; pamatformu anotēšana; vārdšķiru anotēšana; vārdformu anotēšana; vārddarināšanas celmu anotēšana. Tālāk tie raksturoti sīkāk.

1.1. Morfēmu anotēšana

Ja, kā skaidrots iepriekš, morfēmiku uzskata par morfoloģijas daļu, tad viens no morfoloģiskās anotēšanas paveidiem ir morfēmu anotēšana. Korpusu anotējot pēc šāda principa, tags tiek piešķirts nevis vārdam, bet morfēmai, vai nu tikai to nosaucot, vai arī raksturojot pēc kādām pazīmēm, piem.:

<sakn>***Vis***</sakn><gal>***s***</gal> <sakn>***mež***</sakn><gal>***s***</gal>
<pried>***ne***</pried><pried>***no***</pried><sakn>***deg***<sakn><gal>***a***</gal>**.**

Tiesa, morfēmu anotēšana nav sevišķi izplatīta. Valodās, kurās vārda sadalīšana morfēmās ir sarežģīta, šādi anotēt korpusu būtu nesamērīgi grūti attiecībā pret ieguvumu no anotētā korpusa.

1.2. Pamatformu anotēšana

Pamatformu jeb lemmu anotēšana ir ļoti izplatīts anotēšanas paveids. Tas nosaka, ka katram anotējamam vārdam tiek noteikta pamatforma – darbības vārda nenoteiksme, lietvārda vienskaitļa nominatīvs utt., piem.:

<kur>***Kur*** <tu>***tu*** <likt>***liki*** <mans>***manu*** <krekls>***kreklu?***

Šeit gan jāpiebilst, ka pamatformu anotēšana arī ne vienmēr tiek uzskatīta par morfoloģiskās anotēšanas paveidu un dažkārt tiek minēta atsevišķi. Tomēr morfoloģiskai anotēšanai ar to ir ciešs sakars – morfoloģiskā analīze, kas tālāk ved pie morfoloģiskās anotēšanas, sākas ar pamatformas noteikšanu, tāpēc nereti, ja korpusā ir anotētas morfoloģiskās vārda pazīmes, ir anotētas arī pamatformas (sk., piem., Daudaravičius u. c. 2007).

Tomēr, atšķirībā no pārējiem morfoloģiskās anotēšanas paveidiem, kuru noderīgums galvenokārt parādās morfoloģiskas dabas pētījumos (kā, piem., Arkadiev, Pakerys [*manuskripts*]), pamatformu anotēšanai ir ļoti plašs lietojuma lauks, jo tā atvieglo datu meklēšanu, atlasi un aprēķinus pēc meklējamā vārda, neliekot pētniekam atsevišķi noteikt un meklēt visas šī vārda iespējamās formas. Piem., meklējot šādi anotētā korpusā vārdu *vecis*, tiktu atrasti visi piemēri, kuros ir šis vārds neatkarīgi no locījuma: *veci, večos, vecim, večus* utt., turklāt, ja anotēšana veikta pareizi, netiktu atrasts homonīms *veci* (īpašības vārda *vecs* vīriešu dzimtes daudzskaitļa nominatīvs). Tas nozīmē, ka neatkarīgi no tā, kādā valodas līmenī tiek veikts pētījums, ja ir nepieciešams atrast visus kāda noteikta vārda lietojumus, šis anotējums darbu krietni atvieglo.

1.3. Vārdšķiru anotēšana

Vārdšķiru anotēšana paredz noteikt vārda atbilstību noteiktai vārdšķirai un to atbilstoši anotēt. Pēc izvēles var anotēt arī sīkākas apakšgrupas, piem., vietniekvārdus iedalot personu vietniekvārdos, piederības vietniekvārdos, norādāmajos vietniekvārdos utt.; sal.:

1. <vietnv>***Es*** <darbv>***nedošu*** <vietnv>***viņam*** <vietnv>***tavu*** <lietv>***grāmatu!***
2. <vietnv-pers>***Es*** <darbv>***nedošu*** <vietnv-pers>***viņam*** <vietnv-pieder>***tavu*** <lietv>***grāmatu!***

Lai noteiktu vārdšķiras, ir nepieciešams noteikt katra vārda pamatformu, tāpēc vārdšķiru anotēšana bieži vien tiek veikta kopā ar pamatformu anotēšanu (sk., piem., Panunzi u. c. 2004).

Vārdšķiru anotēšana nereti netiek dēvēta par morfoloģisko anotēšanu, bet gan minēta atsevišķi (pētījumos angļu valodā – *part of speech* jeb *POS annotation*). Latviešu un lietuviešu valodniecībā tomēr vārdšķiras tiek uzskatītas par morfoloģiskām kategorijām (sk., piem., Paegle 2003, 25; LVG 2013, 317; DLKG 1997 55–59; u. c.), un nav skaidra pamata tās nodalīt atsevišķi. Jāpiebilst arī – parasti netiek uzsvērts, ka vārdšķiru anotēšana nebūtu morfoloģiskās anotēšanas paveids, tāpēc var pieņemt, ka vismaz daļā gadījumu tā par tādu tiek uzskatīta, lai arī virskategorija netiek pieminēta.

1.4. Vārdformu anotēšana

No pirmā acu uzmetiena vārdformu anotēšana šķiet tā pati morfēmu anotēšana, tomēr, lai arī pastāv cieša saistība, atšķirības starp abiem anotēšanas

veidiem ir ievērojamas. Anotējot morfēmas, tiek atzīmētas vārdu daļas un/vai to pazīmes, bet vārdformu anotēšanas gadījumā runa ir par visai vārdformai piemītošām pazīmēm. Nereti tās ir ierasti nosakāmās gramatiskās kategorijas, piem., lietvārda gadījumā – dzimte, skaitlis un locījums. Tātad atšķirība starp morfēmu un vārdformu anotēšanu ir uzskatāma, sal.:

1. <sakn>***Andr***</sakn><gal>***is***</gal> <sakn>***ir***</sakn> <sakn>***Ilz***</sakn>-<gal>***es***</gal> <sakn>***vīr***</sakn><gal>***s***</gal>.
2. <vīrdz vsk nom>***Andris*** <tagadne īstenizt viensk 3pers>***ir*** <sievdz vsk ģen>***Ilzes*** <vīrdz vsk nom>***vīrs***.

Tāpat kā citos gadījumos, arī anotējot vārdformas, nav noteikti jāanotē pilnīgi visas kādam vārdam iespējamās kategorijas, ja pētījumam tas nedod labumu vai dod maz labuma. Tomēr, ja korpusu ir iecerēts izmantot arī nākotnē, ir vērts to apsvērt, jo bieži vien papildus ieguldāmais darba apjoms ir salīdzinoši neliels. Šī paša iemesla dēļ reizēm korpusu anotētāji apvieno vārdformu anotēšanu ar vārdšķiru un pamatformu anotēšanu, jo ir jānosaka arī pamatforma, lai noteiktu vārdformas kategorijas, savukārt vārdšķira ir jānosaka kaut vai tikai tādēļ, lai saprastu, kuras gramatiskās kategorijas uz konkrēto vārdu ir attiecināmas.

1.5. Vārddarināšanas celmu anotēšana

Ja par morfoloģijas daļu uzskata arī vārddarināšanu, tad viens no morfoloģiskās anotēšanas paveidiem ir arī vārddarināšanas celmu anotēšana (angļu val. – *stemming*). Tas nozīmē, ka, atšķirībā no pamatformu anotēšanas, tiek anotēti celmi, kas var būt kopīgi vairākiem vārdiem, piem.:

<brauk>***Braukātājs*** <brauk>***nobrauca no*** <brauk>***uzbrauktuves uz*** <brauk>***brauktuves***.

Arī šim anotēšanas veidam, līdzīgi kā pamatformu anotēšanai, pielietojums nereti tiek rasts ārpus vārddarināšanas un arī morfoloģijas robežām – vārddarināšanas celmu anotēšana, īpaši automātiska anotēšana, var tikt izmantota, piem., tematiskas meklēšanas sistēmu uzlabošanā (Kreslins 1996, 23).

2. Morfoloģiskā anotēšana latviešu un lietuviešu korpuslingvistikā

Latviešu korpuslingvisti morfoloģiskajai anotēšanai ir pievērsušies darbā ar *Līdzsvaroto mūsdienu latviešu valodas tekstu korpusu*. Šis korpuss ir automātiski morfoloģiski anotēts ar īpašu programmrīku, kas katram vārdam pievieno tā pamatformu un attiecīgās vārdformas morfoloģiskās pazīmes (Levāne-Petrova 2011, 188; u. c.). Latviešu valodai ir tapis arī morfoloģisko pazīmju kopums, kas ir izveidots Latvijas Universitātes Matemātikas un

informātikas institūtā[1]. Lietuviešu valodnieki līdzīgi strādā ar *Mūsdienu lietuviešu valodas korpusu* (Rimkutė u. c. 2009; u. c.).

Anotēšanu var veikt automātiski, pusautomātiski vai manuāli. Tā kā morfoloģiskā sistēma katrai valodai atšķiras, arī morfoloģiskā anotēšana ir valodspecifiska, un automātiskie un pusautomātiskie anotēšanas rīki jāpielāgo vai no jauna jārada katrai valodai. Daži rīki ir radīti arī latviešu valodas apstrādei (Skadiņa 2010). Ir pieejams gan automātisks rīks vārddarināšanas celmu anotēšanai (Kreslins 1996, Eger, Sējāne 2010), gan morfoloģiskais analizators, kurš nosaka gan katra vārda pamatformu, gan vārdšķiru, gan arī attiecīgās gramatiskās kategorijas (Paikens 2008). Arī lietuviešu valodai ir morfoloģiskās analīzes rīks *Lemuoklis* (Zinkevičius 2000).

Valodas apguvēju korpusos morfoloģiskā anotēšana nav plaši izplatīta. Galvenokārt uzmanība pievērsta vārdšķirām – Zigrīda Vinčela ar angļu valodas automātisko anotēšanas rīku *CLAWS* vārdšķiras ir anotējusi savā angļu valodas apguvēju korpusā (Vinčela 2011), savukārt latviešu valodas apguvēju korpusā tās anotētas LVASA pētījumā (Kalnbērziņa u. c. 2011). Ir runāts arī par morfosintaktisku anotējumu franču valodas apguvēju korpusā (Kazlauskienė 2015), taču nav precizēts, pēc kādiem kritērijiem tas veikts.

3. Valodas apguvēju korpusu morfoloģiska anotēšana

Viens īpaši būtisks faktors valodas apguvēju korpusu morfoloģiskajā anotēšanā ir tas, ka valodas apguvēju korpusu anotēšana parasti jāveic manuāli. Tā kā apguvēju tekstos mēdz būt īpaši daudz dažādu noviržu no sagaidāmajiem un algoritmos ietvertajiem likumiem, automātiskās anotēšanas rīku pielāgošana var prasīt iegūstamajam labumam neadekvāti daudz resursu. Tas lielā mērā attiecas arī uz morfoloģisko anotēšanu, jo, apgūstot kādu valodu, īpaši ja šai valodai ir plaša un daudzveidīga morfoloģiskā sistēma, liela daļa kļūdu ir tieši nepareizu vārdformu lietojumā (piem., *es nācīju* ‘nācu’ *uz istabu*) vai nepareizā pareizu vārdformu lietojumā (piem., *es nedzeru piena* ‘pienu’).

Kļūdas mēdz būt ļoti dažādas, nereti – grūti paredzamas, līdz ar to sarežģījot tādu anotēšanas rīku pielāgošanu, kas ir paredzēti pareiza teksta anotēšanai. Tiesa, ar laiku tas varētu mainīties: iespējams, balstoties valodas apguvēju korpusu datos, kļūdu tipus un to rašanos varētu izpētīt tik pilnīgi, lai ievērojami uzlabotu automātisku vai pusautomātisku morfoloģiskās anotēšanas rīku efektivitāti arī darbā ar apguvēju tekstiem.

Vineta Rūtenberga un Vita Kalnbērziņa, runājot par valodas apguvēju korpusa manuālu anotēšanu[1], piebilst, ka „tas ir ārkārtīgi sīkumains un

[1] Pieejams tiešsaistē: *http://www.semti-kamols.lv/doc_upl/TagSet.pdf*

laikietilpīgs uzdevums, it sevišķi zemākajos valodas apguves līmeņos(..)[2]" (Kalnbērziņa, Rūtenberga 2013, 124). Grūtību iemesls ir īpaši augsts variāciju skaits. Jau anotējot literārā valodā rakstītus tekstus, jārisina, piem., homonīmijas radīti jautājumi (sk., piem., Rimkutė 2006). Anotējot kļūdainus tekstus, nereti nav viegli izprast, kāda morfēma vai forma domāta vai pat ko autors vēlējies pateikt. Tāpēc, anotējot valodas apguvēju korpusu, īpaši jāņem vērā anotēšanai pamatā piemītošā subjektivitāte – jebkura anotēšana zināmā mērā ir anotētāja interpretācija. To gan var ierobežot, piesaistot vairākus anotētājus un aprēķinot to vienprātību (sk., piem., Bermingham, Smeaton 2009).

Katrs no šiem faktoriem daļēji nosaka to, ka valodas apguvēju korpusiem, sevišķi zemākā valodas prasmju līmenī, ir vēlams izvēlēties vienkāršāk veicamus anotēšanas paveidus, piem., pamatformu anotēšanu (pretstatā, piem., morfēmu anotēšanai).

4. Otrās baltu valodas apguvēju korpusa morfoloģiska anotēšana

Otrās baltu valodas apguvēju korpusā[3] ir iekļauti teksti, kurus rakstījuši pieauguši studenti, mācīdamies otro baltu valodu (latvieši – lietuviešu, lietuvieši – latviešu valodu) iesācēju līmenī. Korpuss nav paredzēts kādam vienam noteiktam pētījumam, bet gan pēc iespējas vispusīgai baltu starpvalodas[4] pētniecībai. Tas ietekmē arī morfoloģiskās anotēšanas paveidu izvēli.

Tālāk skaidrotas katra morfoloģiskās anotēšanas paveida ieviešanas iespējas otrās baltu valodas apguvēju korpusā, ilustrējot tos ar tekstu fragmentiem no šobrīd publiski pieejamā neanotētā paraugkorpusa lietuviešu valodā.

Morfēmu anotēšana otrās baltu valodas apguvēju korpusā būtu grūti veicama. Jau pats par sevi tas, kā minēts, ir samērā sarežģīts morfoloģiskās anotēšanas paveids, taču minētajā korpusā turklāt daļa tekstu ir tapuši, autoriem vēl esot pašā valodas apguves sākumposmā – pirmajā semestrī. Līdz ar to ir bieži sastopami gadījumi, kuros noteiktu morfēmu robežas nav skaidri nosakāmas, piem.:

Aš mėgstu <u>šviežių</u>s produktus. 'Man patīk svaigi produkti.'

---

[1] Tiesa, šajā gadījumā pētnieces runā par sintaktisku anotēšanu, taču šo apgalvojumu var attiecināt arī uz morfoloģisku pazīmju manuālu anotēšanu.

[2] Šeit un turpmāk autores tulkojums – I.Z.

[3] Pieejams tiešsaistē: *http://www.esamkorpuss.wordpress.com*

[4] Ar *baltu starpvalodu* šeit domāta starpvaloda, kas veidojas, personai ar vienas baltu valodas zināšanām, apgūstot otru baltu valodu.

Šajā teikumā vārdā *šviežiųs* ir izplūdusi galotnes robeža: vai šeit ir viena nepareizi veidota vīriešu dzimtes daudzskaitļa akuzatīva galotne? Varbūt – tāpat kļūdaina ģenitīva galotne? Varbūt divas galotnes, kāds akuzatīva un ģenitīva hibrīds? Droši vien var atrast dažādus argumentus par labu vienam vai citam variantam un atrast šķietami atbilstošāko variantu, taču tas prasa laiku un iedziļināšanos. Tālab, ja šādu piemēru ir daudz, kā tas ir iesācēju tekstos, anotēšanas process tiek ļoti apgrūtināts. Sevišķi, ja to neprasa kāda konkrēta pētījuma vajadzības, ir vēlams izvairīties no anotēšanas, kas prasa izvērstu analīzi lielā skaitā gadījumu. Tāpēc, par spīti potenciālai lietderībai, pieņemts lēmums otrās baltu valodas apguvēju korpusā morfēmas vismaz sākotnēji neanotēt.

Pamatformu anotēšana ir vieglāk veicama. Lai arī bieži sastopamas kļūdas, parasti tomēr nav grūti noteikt, kāds vārds tekstā ir domāts. Tekstus rakstot, studentiem ir bijušas pieejamas vārdnīcas un cita veida palīgmateriāli, tomēr vārdu krājums iesācēju līmenī nav sevišķi plašs, toties ir daudz dažādu formu – gan pareizu, gan kļūdainu, piem., lūk, daži teikumi no viena teksta:

1. <mano>***Mano*** <mama>***mamma*** <būti>***yra*** <labai>***labai*** <teigiamas>***teigiama***, <retai>***retai*** <matyti>***matau*** <ji>***ją*** <liūdnas>***liūdną***. ‘Mana mamma ir ļoti pozitīva, reti redzu viņu bēdīgu.’
2. <su>***Su*** <savo>***savo*** <mama>***mamo*** <galėti>***galiu*** <kalbėti>***kalbėti*** <apie>***apie*** <visas>***visą***, <aš>***aš*** <ta>***tai*** <patikėti>***patikiu*** <paslaptis>***paslaptis***, <nes>***nes*** <žinoti>***žinau,*** <kad>***kad*** <ji>***ji*** <niekas>***nieką*** <tas>***tuos*** <papasakoti>***nepapasakįs.*** ‘Ar savu mammu varu runāt par visu, es tai uzticu noslēpumus, jo zinu, ka viņa nekam tos nepastāstīs.’
3. <žinoti>***Žinau,*** <kad>***kad*** <mama>***mama*** <žinoti>***žina***, <kad>***kad*** <ji>***ji*** <irgi>***irgi*** <aš>***man*** <galėti>***gali*** <patikėti>***patikėti*** <paslaptis>***paslaptis.*** ‘Zinu, ka mamma zina, ka viņa arī man var uzticēt noslēpumus.’
4. <aš>***Aš*** <visāda>***visada*** <imti>***imu*** <galva>***galvoje*** <mama>***mamo*** <požiūris>***požiūrį***, <nes>***nes*** <tai>***tai*** <dažnai>***dažniausiai*** <būti>***yra*** <tiesa>***tiesa.*** ‘Es vienmēr ņemu galvā mammas viedokli, jo tas visbiežāk ir taisnība.’[1]

Kā redzams, četros teikumos četras reizes lietots vārds *mama* ‘mamma’ trīs dažādās formās: *mamma*, *mamo* un *mama*. Turklāt tikai pēdējā no minētajām formām atbilst lietuviešu literārajai valodai, pārējās uzskatāmas par kļūdainām. Protams, tās vienalga var atrast, ievadot programmas meklēšanas logā *mam* ar atbilstošu aizstājējzīmi (*wildcard*), taču šādi var atrast arī

[1] Šajos piemēros īsuma dēļ pamatformas ir norādītas kā tagi, nevis viena vispārīga taga atribūti.

neatbilstošus rezultātus, piem., ja students vārda *mane* ‘mani’ vietā kļūdīdamies rakstījis *mame.* Ja korpusā ir anotētas pamatformas, formu dažādība un kļūdas meklēšanā grūtības nerada.

Tāpat gan šajā piemērā labi redzams, ka, kā jau minēts, nozīmīga loma anotēšanā ir arī interpretācijai. 2. un 4. teikumā vārdiem *tai* ir pievienotas dažādas pamatformas – *ta* ‘tā’ un *tai* ‘tas (vispārīgā nozīmē, nekatrā dzimtē)’. Turklāt 2. teikumā, iespējams, nozīme bijusi iecerēta tuvāk *tai* formai vai pat *jai* – personas vietniekvārda *ji* ‘viņa’ datīva formai. Tātad šeit jāpaļaujas uz anotētāja intuīciju.

Ņemot vērā plašās izmantojuma iespējas, pēc šī principa otrās baltu valodas apguvēju korpusu anotēt būtu vēlams, taču, anotēto materiālu izmantojot, noteikti jāņem vērā interpretācijas subjektivitātes faktors, jo šī korpusa izveidei pieejamie resursi vismaz sākumā neļauj piesaistīt vairākus anotētājus, kā minēts iepriekš.

Vārdšķiru anotēšana, līdzīgi kā pamatformu anotēšana, ir diezgan vienkārša. Kā jau minēts – ja tiek anotētas pamatformas, tad vārdšķiras noteikt nav sarežģīti. Piem., izmantojot iepriekš pamatformu anotēšanu ilustrējušo tekstu, nav grūti aizstāt pamatformas ar vārdšķiru tagiem:

1. <vietniekv>***Mano*** <lietv>***mamma*** <darbv>***yra*** <apstv>***labai*** <īpv>***teigiama***, <apstv>***retai*** <darbv>***matau*** <vietniekv>***ją*** <īpv>***liūdną***.
2. <priev>***Su*** <vietniekv>***savo*** <lietv>***mamo*** <darbv>***galiu*** <darbv>***kalbėti*** <priev>***apie*** <vietniekv>***visą***, <vietniekv>***aš*** <vietniekv>***tai*** <darbv>***patikiu*** <lietv>***paslaptis***, <saikl>***nes*** <darbv>***žinau,*** <saikl>***kad*** <vietniekv>***ji*** <vietniekv>***nieką*** <vietniekv>***tuos*** <darbv>***nepapasakįs.***

Šāds anotējums ļauj veikt pētījumus, meklējot pēc vārdšķiras. Ja vajadzīgs, var arī apvienot abus anotējuma veidus vienā datnē. Lai to izdarītu, veido sarežģītākus tagus, kurās ietilpst dažāda informācija, piem.:

<pamatf=mano vārdšķ=vietniekv>***Mano*** <pamatf=mama vārdšķ=lietv>***mamma*** <pamatf=būti vārdšķ=darbv>***yra*** <pamatf=labai vārdšķ=apstv>***labai*** <pamatf=teigiamas vārdšķ=īpv>***teigiama***, <pamatf=retai vārdšķ=apstv>***retai*** <pamatf=matyti vārdšķ=darbv>***matau*** <pamatf=ji vārdšķ=vietniekv>***ją*** <pamatf=liūdnas vārdšķ=īpv>***liūdną***.

Tā kā katrs vārds pieder vienai vārdšķirai, šāds anotējums visbiežāk nesašaurinās meklējumu pēc vārda (piem., pēc vaicājuma *mano* tiks atrasts vienāds rezultātu skaits neatkarīgi no tā, vai būs norādīts, ka jāmeklē tikai vietniekvārdi). Tomēr anotējums var palīdzēt homonīmijas gadījumos, piem., latviešu valodas vārdu pārī *plāns* (lietvārds; piem., *dzīvokļa plāns*) un *plāns* (īpašības vārds; piem., *plāna grāmata*).

Otrās baltu valodas apguvēju korpusā vārdšķiru anotēšana būtu vēlama līdz ar pamatformu anotēšanu. Izveidojot īsāku un parocīgāku tagu sistēmu, abus anotējumus var arī apvienot.

Vārdformu anotēšana ir sarežģītāka. Kā jau norādīts šī darba 2.4. apakšnodaļā, tā ir cieši saistīta ar morfēmu noteikšanu, un gadījumos, kad morfēmas nav skaidri nosakāmas, arī vārdformu noteikšana ir apgrūtināta. Piem., jau minētajā teikumā:

Aš mėgstu šviežiųs produktus.

Tā kā nav skaidrības, kādai galotnei šeit būtu jābūt, arī locījums nav īsti skaidrs. Tiesa, šķiet, autors ir zinājis, ka *mėgstu* lieto ar akuzatīvu, jo šādā locījumā ir lietojis vārdu *produktus*, taču it sevišķi iesācēju rakstītos tekstos ir iespējams arī nekonsekvents locījumu lietojums.

Jāatzīst, ka daļā tekstu, kas rakstīti otrajā valodas apguves semestrī, ir maz šāda tipa kļūdu vai pat to nav vispār, tomēr, anotējot visu otrās baltu valodas apguvēju korpusu, šis anotēšanas paveids nebūtu uzskatāms par primāru.

Vārddarināšanas celmu anotēšana valodas apguvēju korpusos nav ierasta prakse. Iespējams, tam iemesls ir neskaidrība par ieguvumu no šāda darba. Tas gan daļēji izriet no pamatformu anotēšanas, tāpēc nav tik sarežģīts kā, piem., morfēmu anotēšana, tomēr nav īsti skaidrs, kāds nozīmīgs labums tiktu gūts, anotējot, piem., šādi:

<staig>***Staiga*** <vis>***visas*** <suk>***pasisuka*** <prieš>***priešingai.*** ‘Pēkšņi viss apgriežas otrādi.’

Tā kā otrās baltu valodas apguvēji mācību iestādēs, kurās tapuši korpusā iekļautie teksti, leksiku apgūst, kā pamatvienību uztverot leksēmu, nevis celmu, nez vai celms būtu nozīmīgākā anotējamā kategorija. Lai arī noteikti ir rodams ieguvums arī no šāda anotēšanas veida, tas nešķiet primārs otrās baltu valodas apguvēju korpusa gadījumā.

Secinājumi

Morfoloģiskā anotēšana, tāpat kā citu veidu anotēšana, balstās korpusa tekstu analīzē. Tātad morfoloģiskās anotēšanas gadījumā tā ir morfoloģiskā analīze. Anotēšanas paveidu saraksts ir atkarīgs no tā, kas tiek uzskatīts par morfoloģijas pētāmo objektu. Pieņemot, ka līdz ar vārdšķirām un formveidošanu morfoloģijā ietilpst arī morfēmika un vārddarināšana, var izdalīt piecus morfoloģiskās anotēšanas paveidus: morfēmu, pamatformu, vārdšķiru, vārdformu un vārddarināšanas celmu anotēšanu.

Gan Latvijā, gan Lietuvā morfoloģiskajai anotēšanai pētnieki ir pievērsušies; ir gan korpusi, kas anotēti pēc morfoloģijas principiem, gan arī rīki automātiskai anotēšanai. Tiesa, šāds anotējums nav sevišķi izplatīts

minētajās valstīs tapušajos valodas apguvēju korpusos. Plašā nestandarta variāciju klāsta dēļ valodas apguvēju korpusu anotēšana jāveic manuāli.

Otrās baltu valodas apguvēju korpusam vispiemērotākie morfoloģiskās anotēšanas paveidi ir pamatformu anotēšana un vārdšķiru anotēšana. Tā kā korpuss satur iesācēju tekstus, kuros ir daudz noviržu, lietojot šādi anotētu korpusu, jāņem vērā subjektīvās interpretācijas ietekme.

Saīsinājumi

LVASA – Latvijas valodu skolotāju asociācija

LVG – Latviešu valodas gramatika

Bibliogrāfija

1. **Arkadiev, Pakerys [*manuskripts*] – Arkadiev, Peter, Pakerys, Jurgis.** Lithuanian morphological causatives: A corpus based study. To appear in *Argument Structure and Diathesis in Baltic.* Eds. Axel Holvoet, Nicole Nau.
2. **Bermingham, Smeaton 2009 – Bermingham, Adam, Smeaton, Alan F.** A study of inter-annotator agreement for opinion retrieval. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. New York : ACM, 2009, pp. 784–785.
3. **Britto u. c. 1999 – Britto, Helena, Galves, Charlotte, Ribeiro, Ilza, Augusto, Marina, Scher, Ana Paula.** Morphological Annotation System for Automatic Tagging of Electronic Textual Corpora: from English to Romance Languages [skatīts 02.02.2015.]. *Proceedings of the 6th International Symposium of Social Communication*. Santiago de Cuba : University of Oriente, 1999. Pieejams tiešsaistē: *www.tycho.iel.unicamp.br/~tycho/pesquisa/artigos/BRITTO_HetalFase1a.pdf*
4. **Daudaravičius u. c. 2007 – Daudaravičius, Vidas, Rimkutė, Erika, Utka, Andrius.** Morphological annotation of the Lithuanian corpus. *ACL '07 Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*. Prague : Association for Computational Linguistics, 2007, pp. 94–99.
5. **DLKG 1997 –** *Dabartinės lietuvių kalbos gramatika*. Trečiasis pataisytas leidimas. Vilnius : Mokslo ir enciklopedijų leidybos institutas, 1997.
6. **Eger, Sējāne 2010 – Eger, Steffen, Sējāne, Ineta.** An Ensemble of Classifiers Methodology for Stemming in Inflectional Languages: Using the Example of Latvian. *Human Language Technologies – The Baltic Perspective*. Amsterdam : IOS Press, 2010, pp. 217–224.
7. **Granger 2003 – Granger, Sylviane.** Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* 20 (3), pp. 465–480.

8. **Kalme, Smiltniece 2001 – Kalme, Vilma, Smiltniece, Gunta.** *Latviešu literārās valodas vārddarināšana un morfoloģija.* Lokāmās vārdšķiras. Liepāja : LiePA, 2001.
9. **Kalnača 2004 – Kalnača, Andra.** *Morfēmika un morfonoloģija.* Rīga : LU akadēmiskais apgāds, 2004.
10. **Kalnbērziņa, Rūtenberga 2013 – Kalnbērziņa, Vita, Rūtenberga, Vineta.** Syntactic indicators of language acquisition levels in English and French written language learner corpora [viewed on 17.01.2014.]. *Lublin Studies in Modern Languages and Literature* 37. Available online: *http://www.lsmll.umcs.lublin.pl/issues/37-2013/8kalnberzina.pdf*
11. **Kalnbērziņa u. c. 2011 – Kalnbērziņa, Vita, Lokmane, Ilze, Kunda, Tatjana, Vinčela, Zigrīda, Baiža, Kristīne.** *Latviešu valodas apguves kvalitāte mazākumtautību skolās*. Rīga : LVASA, 2011.
12. **Kazlauskienė 2015 – Kazlauskienė, Vitalija.** Daiktavardinis žodžių junginys kaip gramatinės kompetencijos įsisavinimo elementas prancūzų kalbos baigiamojo egzamino rašto darbuose. Santrauka konferencijai *Darnioji Daugiakalbystė: kalba, kultūra, visuomenė* 2015 m. gegužės 29–30 d. Kaunas : Vytauto Didžiojo universitetas, 2015.
13. **Kreslins 1996 – Kreslins, Karlis.** *A stemming algorithm for Latvian.* Doctoral thesis. Loughborough : Loughborough University, 1996.
14. **Levāne-Petrova 2011 – Levāne-Petrova, Kristīne.** Morfoloģiski marķēta valodas korpusa izmantošana valodas izpētē. *Vārds un tā pētīšanas aspekti* : rakstu krājums, 15 (1), Liepāja, LiePA, 2011, 187.–193. lpp.
15. **LVG 2013 –** *Latviešu valodas gramatika*. Rīga : LU Latviešu valodas institūts, 2013.
16. **McEnery, Hardie 2012 – McEnery, Tony, Hardie, Andrew.** *Corpus Linguistics: Method, Theory and Practice*. Cambridge, New York : Cambridge University Press, 2012.
17. **Miliūnaitė 2003 – Miliūnaitė, Rita.** *Lietuvių kalbos gramatikos norminimo pagrindai*. Vilnius : Lietuvių kalbos instituto leidykla, 2003.
18. **MLLVG 1959** – *Mūsdienu latviešu literārās valodas gramatika I.* Fonētika un morfoloģija. Rīga : LPSR Zinātņu akadēmijas izdevniecība, 1959.
19. **Paegle 2003 – Paegle, Dzintra.** *Latviešu literārās valodas morfoloģija*. 1. daļa. Rīga : Zinātne, 2003.
20. **Paikens 2008 – Paikens, Pēteris.** Lexicon-Based Morphological Analysis of Latvian Language. *Proceedings of the 3rd Baltic Conference on Human Language Technologies.* Vilnius : Vytautas Magnus University, Institute of the Lithuanian Language 2008, pp. 235–240.
21. **Panunzi u. c. 2004 – Panunzi, Alessandro, Picchi, Eugenio, Moneglia, Massimo.** Using PiTagger for Lemmatization and PoS Tagging of a

Spontaneous Speech Corpus: C-Oral-Rom Italian [viewed on 17.01.2014.]. *Proceedings of the 4th LREC Conference* Paris : ELRA, 2004. Available online: *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.7175&rep=rep1 &type=pdf*

22. **Rimkutė 2006 – Rimkutė, Erika.** *Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne*. Daktaro disertacija. Kaunas : Vytauto Didžiojo universitetas, 2006.
23. **Rimkutė u. c. 2009 – Rimkutė, Erika, Valskys, Vidas, Vaskelienė, Jolanta.** Lietuvių kalbos leksemų morfologinis anotavimas: ypatumai ir sunkumai. *Kalbų studijos*, Nr. 15, 63–70 psl.
24. **Skadiņa 2010 – Skadiņa, Inguna.** Vienota valodas resursu un tehnoloģiju infrastruktūra *Clarin*. *Vārds un tā pētīšanas aspekti* : rakstu krājums, 14 (2). Liepāja : LiePA, 2010, 299.–305. lpp.
25. **Urbutis 1978 – Urbutis, Vincas.** *Žodžių darybos teorija.* Vilnius : Mokslas, 1978.
26. **Vinčela 2011 – Vinčela, Zigrīda.** Linguistic Variation in EFL Students-Composed Virtual Texts in Different Registers. *Corpus Linguistics Conference 2011*, Birmingham, 20–22 July 2011 : Proceedings.
27. **VPSV 2007** – *Valodniecības pamatterminu skaidrojošā vārdnīca.* Rīga : LU Latviešu valodas institūts, 2007.
28. **Zinkevičius 2000 – Zinkevičius, Vytautas.** *Lemuoklis* – morfologinei analizei. *Darbai ir dienos,* Nr. 24. 2000, 245–274 psl.

MORPHOLOGICAL ANNOTATION OF A LEARNER CORPUS OF THE SECOND BALTIC LANGUAGE

Summary

One of the most popular annotation types in corpus linguistics is morphological annotation. It is also popular in Latvia and Lithuania, albeit not much used in learner corpora here. General corpora of both languages have been morphologically annotated, and automatic tools for morphological analysis of both languages have been created. Five subtypes of morphological annotation can be divided. Lemmatization and part of speech annotation would also be a good choice for the learner corpus of the second Baltic language, while morphemic annotation, inflection (word form) annotation and stemming are not seen as a priority for it.

Some important factors include the large amount of various errors which make this kind of corpora rather unfitting for automatic annotation. Besides, every kind of annotation can be, to some extent, subjective, and, the more varieties of errors there are, the more intuitive the annotation process may become. This, in turn, can make the resulting annotated corpus less objectively reliable for research.