Marķētājs beta

Latviešu valodas tekstu korpusu morfoloģiskās un sintaktiskās marķēšanas rīks

Marķētājs ir pusautomātisks programmrīks, kas paredzēts latviešu valodas tekstu morfoloģiskai un sintaktiskai analīzei. Ar tā palīdzību katrai vārdformai tiek dots morfoloģiskais raksturojums un, ja iespējams, noteikta arī vārda sintaktiskā loma teikumā.

Morfoloģiskajam raksturojumam tiek izmantoti speciāli simboli. Marķētājā tiek izmantota latviešu valodas morfoloģisko pazīmju kopa, kas ir atvasināta no *MULTEXT-East* pazīmju kopas[1]. Morfoloģisko pazīmju kopu var aplūkot arī, ejot uz izvēlni [Palīgs] \ [Morfoloģisko pazīmju kopa].

Marķējamā teksta atvēršana

Marķējamo tekstu var atvērt, ejot uz izvēlni [Korpuss] \ [Atvērt marķējamo tekstu]

Marķētā teksta saglabāšana

Marķēto tekstu var saglabāt, ejot uz izvēlni [Korpuss] \ [Saglabāt marķēto tekstu]. Marķētie dati tiek saglabāti gan teksta formātā, gan .xml formātā. Saglabātajā teksta failā vārdi sakārtoti viens zem otra, un blakus tiem dots marķējums.

Teksta fragmentēšana

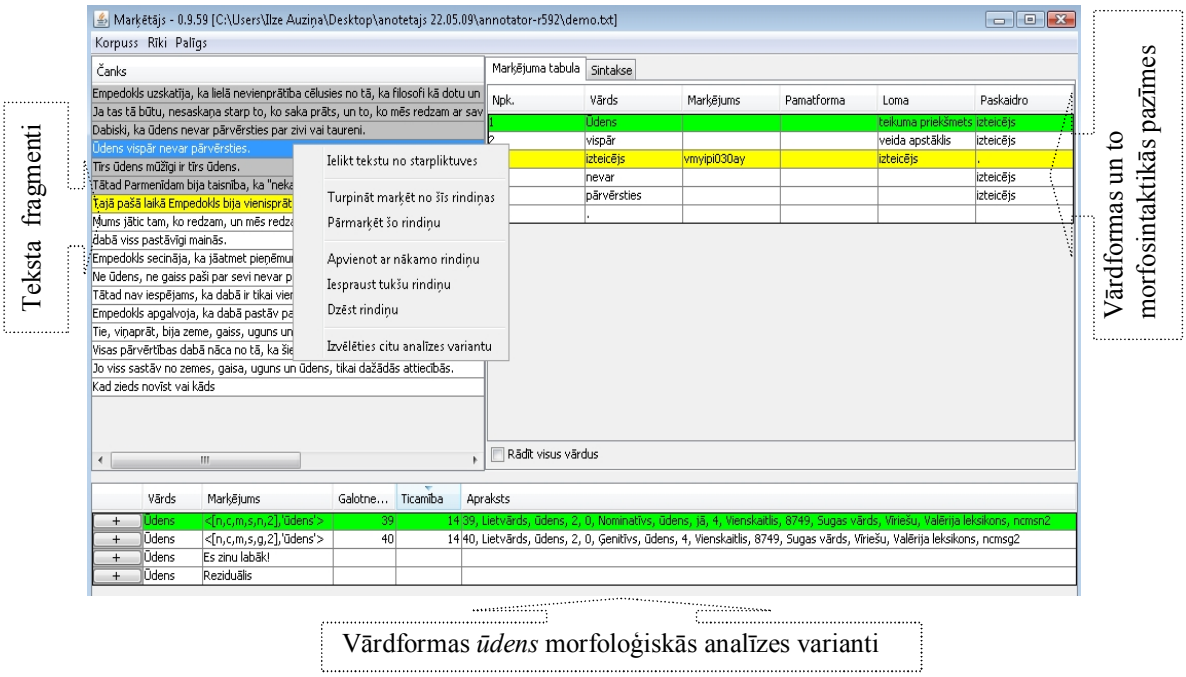
Atverot marķējamo tekstu, tas automātiski tiek sadalīts fragmentos un vārdformām tiek pievienots morfoloģiskais raksturojums/morfoloģiskās pazīmes, t. i., tiek piedāvāti visi iespējamie marķējuma varianti. Teksta fragmenti ir sakārtoti viens zem otra saskarnes kreisajā pusē (sk. 1. att.).

Teksta fragmentu krāsa. Dzeltenā krāsā iekrāsoti tie teksta fragmenti, kuru marķēšana ir iesākta, bet nav pabeigta. Savukārt pelēkā krāsā ir tie teksta fragmenti, kuru marķēšana ir pabeigta. Teksta fragments, kas konkrētajā brīdī tiek analizēts un kura vārdformas un pieturzīmes ir dotas marķējuma tabulā, ir zaļā krāsā (sk. 1. att.).

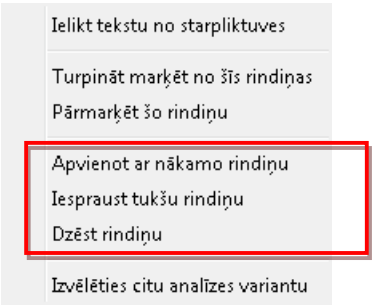
Teksta fragmentus iespējams **dzēst** [Dzēst rindiņu] un **apvienot** [Apvienot ar nākamo rindiņu], starp teksta fragmentiem var **ievietot tukšu rindiņu** [Iesprausts tukšu rindiņu]. To var izdarīt, novietojot kursoru uz teksta fragmenta un uzklikšķinot uz tā ar peles labo taustiņu. Parādās izvēlnes logs (sk. 2. att.), kurā

[1] Sk. sadaļu [Rīki] \ [Morfoloģiskais analizators] adresē http://www.semti-kamols.lv

jāizvēlas opcija [Apvienot ar nākamo rindiņu], [Dzēst rindiņu] vai [Iesprausts tukšu rindiņu].



1. att. Marķētāja saskarne

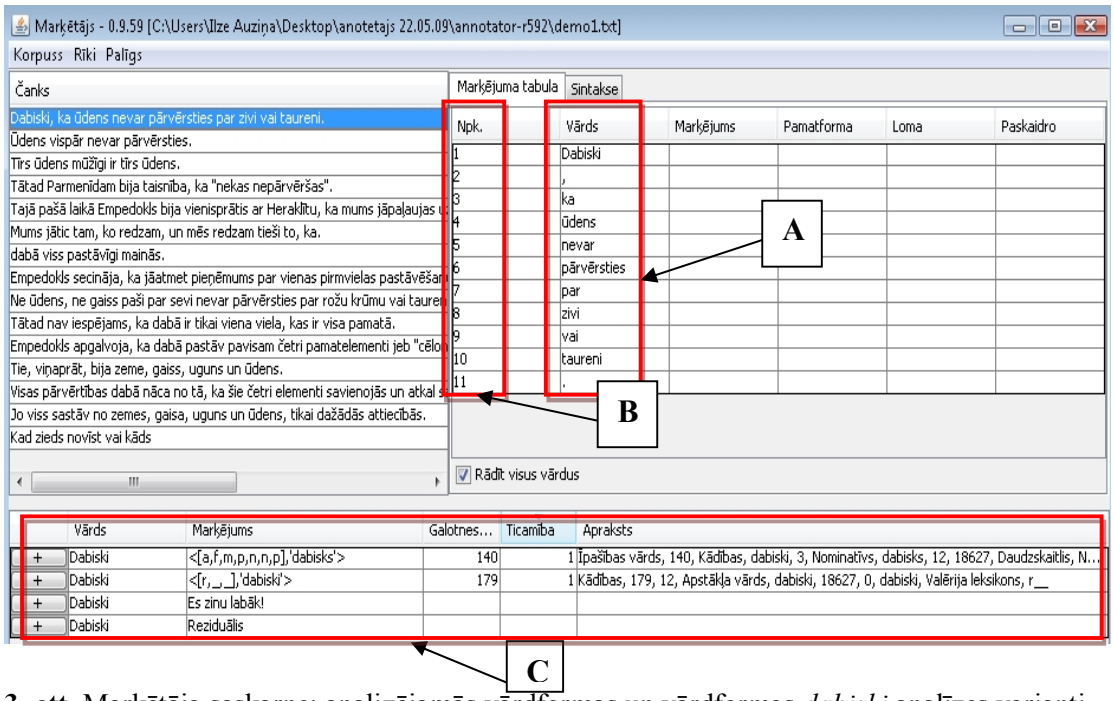


2. att. Teksta fragmentu apstrādes komandrindas

Morfoloģiskā analīze

Atverot marķējamo tekstu, visas pirmā teksta fragmenta vārdformas un pieturzīmes tiek sarakstītas *Marķējuma tabulā* saskarnes labajā pusē (sk. 3. att. A) un tām tiek piešķirts kārtas numurs (sk. 3. att. B). Automātiski tiek piedāvāti pirmās vārdformas vai pieturzīmes analīzes varianti (sk. 3. att. C)

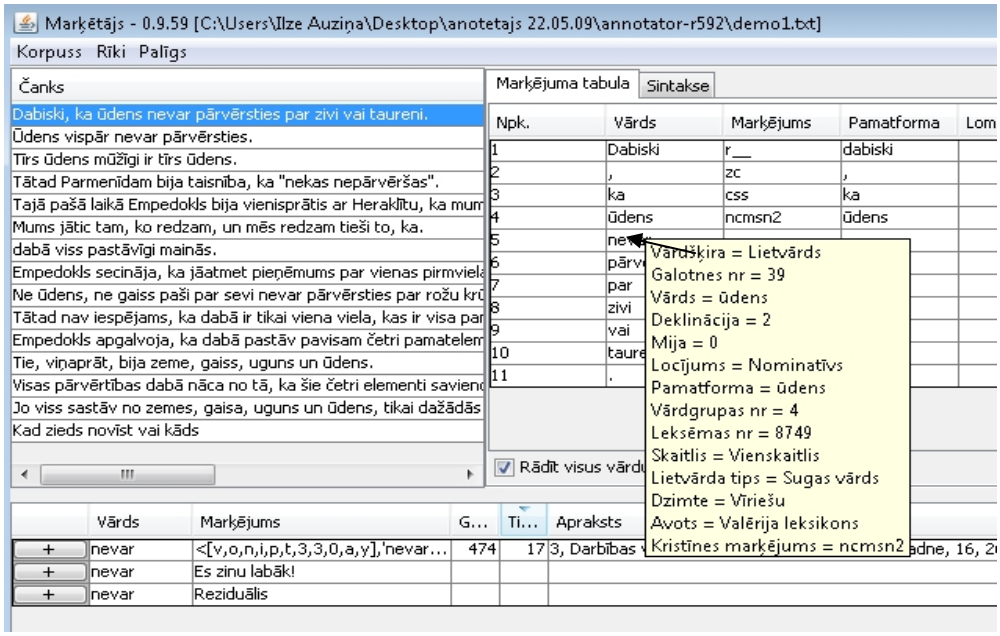
Marķējuma tabulas (sk. 3. att.) 1. ailē doti vārdformu un pieturzīmju kārtas numuri, 2. ailē *Vārds* — vārdformas un pieturzīmes, savukārt 3. aile *Marķējums* paredzēta marķējuma / morfoloģisko pazīmju virknei, 4. aile *Pamatforma* — vārda pamatformai, 5. aile *Loma* paredzēta sintaktiskās lomas atspoguļošanai, 6. aile *Paskaidro* norāda uz vārdu, kam attiecīgā vārdforma pakārtota.



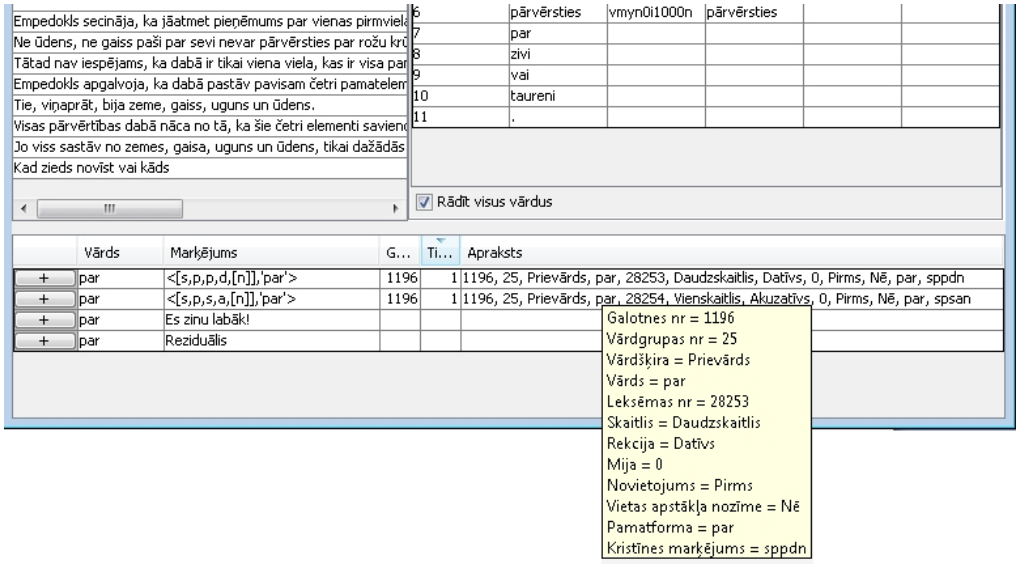
3. att. Marķētāja saskarne: analizējamās vārdformas un vārdformas *dabiski* analīzes varianti

Katru marķējamo vārdformu vai pieturzīmi var aplūkot tabulā saskarnes lejas daļā: tur redzama vārdforma (vai pieturzīme), iespējamais morfoloģiskais marķējums un marķējuma apraksts, norādīta ticamība un galotnes numurs (sk. 3. att. C). Izvēlēties atbilstošo marķējumu iespējams, uzklikšķinot uz pluszīmes rindiņas sākumā.

Novietojot peli uz jebkuras jūs interesējošā vārda ailītes marķējuma tabulā (ja vārdam jau ir pievienots morfoloģiskais raksturojums) (sk. 4. att.) vai izvēlnes daļā, kur doti iespējamie analīzes varianti (sk. 5. att.), papildu logā automātiski parādās izvērsta informācija par šo vārdu (galvenokārt morfoloģiskais raksturojums).



4. att. Papildinformācija par vārdformu *ūdens*

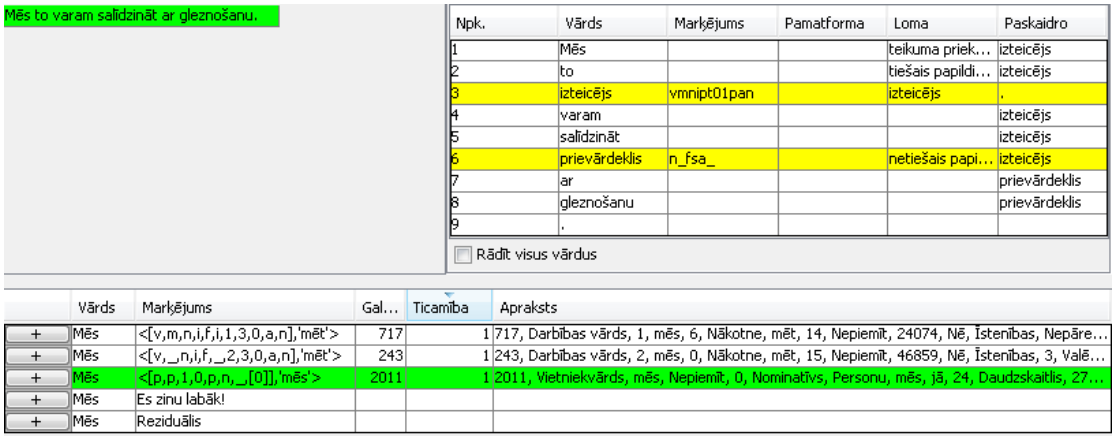


5. att. Papildinformācija par vārdu *par*

Pareizā marķējuma izvēle

Latviešu valodas marķētājs izmanto analizatora vārdnīcā (leksikonā) atrodamo informāciju par vārdiem. Ja kāds no teikumā esošajiem vārdiem neatrodas analizatora vārdnīcā, tātad nav iepriekš zināma tā vārdšķira u. c. informācija par vārdu, analizators mēģina to noteikt pēc vārdformas morfoloģiskajām pazīmēm (galotnes, priedēkļa, piedēkļa).

Tāpēc daudzos gadījumos programma piedāvā vairākus marķējuma variantus, jo, kā zināms, latviešu valodā diezgan bieži ir sastopamas homoformas, t. i., divas vai vairākas dažādas nozīmes vārdformas, kuras raksta vienādi un kuras nav iespējams atšķirt pēc ārējām pazīmēm (bez leksiskās nozīmes analīzes), piemēram, *sviesta* (lietvārds *sviests* vienskaitļa ģenitīvā) un *sviesta* (divdabis *sviests* sieviešu dzimtes vienskaitļa nominatīvā), *kur* (apstākļa vārds) un *kur* (darbības vārda *kurt* īstenības izteiksmes vienkāršās tagadnes vienskaitļa 2. personas un vienskaitļa un daudzskaitļa 3. personas forma), *mēs* (daudzskaitļa 1. personas vietniekvārds) un *mēs* (darbības vārda *mēt* īstenības izteiksmes vienkāršās nākotnes 3. personas forma) (sk. 6. att.). Marķētājs piedāvā risinājumu homoformu analīzē: lietotājam, uzklikšķinot uz pluszīmes, ir iespēja izvēlēties atbilstošo morfoloģisko raksturojumu.

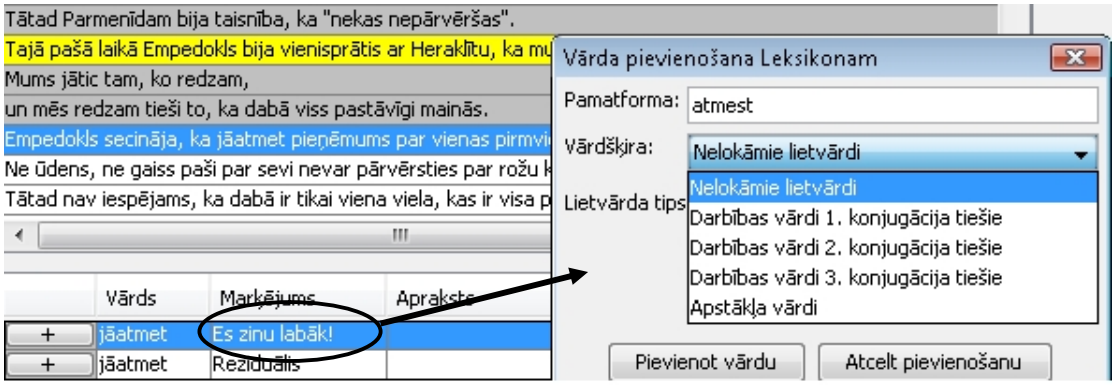


6. att. Vārdformas *mēs* morfoloģiskās analīzes varianti

Opcija *Es zinu labāk!*

Ja neviens no vārdformas raksturojumiem nav pareizs, programmas lietotājs var piedāvāt savu vārda morfoloģisko raksturojumu, uzklikšķinot uz *Es zinu labāk!,* un vienlaicīgi to pievienot jau eksistējošam leksikonam. Arī tad, ja informācijas par kādu vārdu vārdnīcā nav un marķētājs nepiedāvā nevienu analīzes variantu, lietotājs to var pievienot pats, uzklikšķinot uz *Es zinu labāk!*

- Parādīsies logs (sk. 7. att.), kura ailītē *Pamatforma* jāieraksta konkrētā (analizējamā) vārda pamatforma.
- Pēc pamatformas ierakstīšanas, ailē *Vārdšķiras* automātiski tiks piedāvātas iespējamās vārdšķiras, arī lietvārdu deklinācijas, verbu konjugācijas u. tml.
- Tālākā izvēle atkarīga no vārda piederības konkrētai vārdšķirai, t. i., ja analizējamais vārds ir darbības vārds, būs iespējams rakstot 1) verba transitivitāti (pārejošs/nepārejošs), 2) noteikt darbības vārda tipu (*Patstāvīgs darbības vārds/ palīgdarbības vārds/ Modāls/ Fāzes/ Izpausmes veida/ Palīgverbs 'būt'/ Palīgverbi 'tikt' un 'tapt'*). Ja analizējamais vārds ir lietvārds, var raksturot tā tipu, t. i., vai tas ir īpašvārds vai sugas vārds.
- Ja analizētajam vārdam nepieciešamās pazīmes ir pievienotas, tiek piedāvāta iespēja [Pievienot vārdu] vai [Atcelt pievienošanu]. Izvēloties iespēju pievienot vārdu, tas ar attiecīgi izveidoto morfoloģisko marķējuma variantu parādīsies saskarnes lejasdaļā, kur to būs iespējams apstiprināt, nospiežot pluszīmi.



7. att. Vārda pievienošana leksikonam

Ja lietotājs nav apmierināts ar piedāvāto morfoloģisko vārdformas raksturojumu, to var labot manuāli, uzklikšķinot uz atbilstošās ailītes saskarnes labajā pusē slejā *Marķējums.*

Lai redzētu visas jau samarķētās teksta vienības, saskarnes logā (zem vārdformu un pieturzīmju saraksta) jāatzīmē *Rādīt visus vārdus.*

Marķētājs piedāvā iespēju pārmarķēt jau apstrādātu tekstu vai arī turpināt marķēt citā teksta vietā, piemēram, izlaižot vairākus teikumus. Lai to izdarītu, jāiezīmē rinda (teksta fragments), no kuras teksts tiks pārmarķēts, un ar peles labo taustiņu jāuzklikšķina uz tās. Tiks piedāvāta izvēlne, kurā jāizraugās iespēja [Turpināt marķēt no šīs rindas].

Morfoloģiskās pazīmes

Morfoloģiskās analīzes laukā katrai vārdformai tiek noteikta morfoloģisko pazīmju kopa. Pazīmju skaits kopā katrai vārdšķirai ir atšķirīgs, piemēram, darbības vārds tiek raksturots, izmantojot 11 pazīmes (sk. 8. att.), lietvārds — izmantojot 6 pazīmes, savukārt izsauksmes vārda un partikulas raksturošanai izmanto tikai 2 pazīmes.

Pirmā pazīme (kodifikators) morfoloģisko pazīmju kopā norāda vārda piederību konkrētai vārdšķirai, piemēram, darbības vārda kodifikators ir *v*, lietvārda — *n,* īpašības vārda — *a*, vietniekvārda — *p*, apstākļa vārda — *r.*

Vairākas pazīmes dažādu vārdšķiru vārdiem ir kopīgas, piemēram, dzimte, skaitlis, locījums piemīt lietvārdam un īpašības vārdam (3., 4., 5. pazīme), vietniekvārdam (4., 5., 6. pazīme), divdabim (6., 7., 8. pazīme).

Atsevišķos gadījumos morfoloģisko pazīmju kopā iekļautas arī semantiskas pazīmes, piemēram, kur iespējams, apstākļa vārdiem norādīts, kādas nozīmes apstākļa vārdi tie ir — laika, vietas, veida, mēra, cēloņa. Saikļi raksturoti gan pēc sintaktiskās funkcijas, gan pēc uzbūves, piemēram, saiklis *jo* — vienkāršs pakārtojuma saiklis un atkārtots sakārtojuma saiklis. Norādāmajiem vietniekvārdiem atzīmēts, kādas vārdšķiras tie aizstāj.

□[v,m,n,i,p,t,3,3,p,a,n], □liecināt'>

Pazīme	**vērtība**	**atšifrējums**
kodifikators	v	darbības vārds
tips	m	patstāvīgs (*main*)
atgriezeniskums	n	nē
izteiksme	i	īstenības izteiksme
laiks	p	tagadne (*present*)
pabeigtība	i	nepabeigts (*imperfect*)
konjugācija	3	3. konjugācija
persona	3	3. persona
skaitlis	p	daudzskaitlis (*plural*)
kārta	a	darāmā kārta
noliegums	n	nav noliegtais darbības vārds

8. att. Vārdformas *liecina* morfoloģiskais raksturojums

Sintaktiskā analīze

Ja iespējams, sintaktiskā informācija marķētajam teksta fragmentam (konkrēti — vārdformām), tiek pievienota automātisku. Pamatojoties uz vārdu morfoloģiskajām formām, tiek analizētas teikuma sintaktiskās attieksmes (piem., saskaņojums, atkarības), noteikts, piemēram, ka lietvārds akuzatīvā var būt atkarīgs no verba finītās formas, kā arī tiek noteikta vārdformas sintaktiskā loma teikumā. Analizētājs analizē teikumu gramatiski (morfoloģiski un sintaktiski) analīzi — nosaka katra vārda morfoloģisko formu un, balstoties uz tām, parāda iespējamo teikuma sintaktisko struktūru.

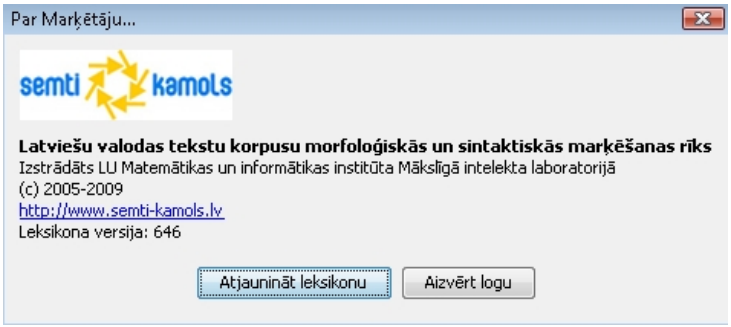
Lai vārdformu analīzes tabulai pievienotu sintaktiskās lomas, atkal jāiezīmē konkrēts teikums/teksta fragments un jāuzklikšķina ar peles labo taustiņu. Parādoties izvēlnei, jāizraugās [Izvēlēties citu analīzes variantu]. Tad jaunā logā tiks piedāvāts teikuma sintaktiskais marķējums, t. s. kastītes (sk. 9. att.).

9. att. Sintaktiskās analīzes piemērs

Ja tiek piedāvāti vairāki varianti, lietotājam jāizraugās visatbilstošākais, uzklikšķinot uz saskarnes labajā malā redzamā cipara. Tad automātiski jaunā informācija par vārdu sintaktiskajām lomām tiks ievietota iepriekš aprakstītās tabulas (sk 1. att., 7. att.) 5. un 6. ailē Ja neviens no piedāvātajiem variantiem nav precīzs / pareizs, jānospiež saskarnes lejā redzamā poga *Neviens no variantiem īsti neatbilst.*

Leksikona atjaunināšana

Marķējot tekstus, ik pa laikam leksikonam tiek pievienots jauns vārds (sk. 7. att.). Tā kā marķētāju vienlaicīgi var izmantot vairāki lietotāji, kas savukārt neatkarīgi cits no cita leksikonam var pievieno vārdus, leksikons ir jāatjaunina. To var izdarīt, ejot uz izvēlni [Palīgs] \ [Par...] un izvēlnes logā, uzklikšķinot uz pogas [Atjaunināt leksikonu] (sk. 10. att.). Pašreizējā versijā lejupielādētais atjauninājums ir pusautomātiski jāsinhronizē ar esošo leksikonu, izmantojot [Rīki] \ [Sinhronizēt leksikonus].



10. att. Leksikona atjaunināšana