LIETUVIŲ KALBOS GRAMATIKOS INFORMACINĖ SISTEMA: I MORFOLOGIJA

Daiva Šveikauskienė
Lietuvių kalbos institutas
P. Vileišio g. 5,
LT-08404 Vilnius, Lietuva
El. paštas: daiva.fmf@gmail.com

1. ĮVADAS

Lietuvių kalbos institute pradėta kurti lietuvių kalbos gramatikos informacinė sistema. Ji apima dvi sritis – morfologiją ir sintaksę. Pirmo etapo metu bus paruošti morfologiniai duomenys. Pagrindinis tikslas – sukaupti išsamią gramatinę informaciją apie visų lietuvių kalbos žodžių visas formas. Vertinant jau atliktus lietuvių kalbos kompiuterizavimo darbus galima pasakyti, kad jie visi turi vieną bruožą – atspindi lietuvių kalbą fragmentiškai. Tolesniuose skyriuose bus pagrįstas šis teiginys.

Daugiausia morfemikos kompiuterizavimo srityje nuveikta Vytauto didžiojo universitete (VDU), kur atliekami darbai remiasi tekstynu. Tačiau ir kitų kalbų lingvistai kaip trūkumą nurodo, kad tokio pobūdžio tyrimai teapima tik tekstyno žodžius ir tegali atspindėti tik juose esančią leksiką. Tai ypač aktualu didelio kaitomumo kalboms, nes „net ir labai didelės apimties tekstynuose gali nebūti rečiau pasitaikančių formų“ (Paikens, Rituma, Pretkalnina 2013, 272). Ne kitokia padėtis ir su lietuvių kalba. VDU paruoštose duomenų bazėse – tiek morfemikos, tiek morfologijos – trūksta kai kurių žodžių formų. Morfemikos duomenų bazėje (1 interneto nuoroda), nėra labai įprastų, gerai visiems žinomų žodžių, pvz., *laikmenai*, *laikmeną*, *laikmenoms*, *laikmenomis*, *laikrodžiui*, *laikrodyje*, *laikrodžių*, *laikrodžiams*, *laikrodžiais*, *laikrodžiuose* ir tai tokie žodžiai, kurių negalima atmesti ir traktuoti juos kaip nevartojamus, t.y. archaizmus ar pan. Trūksta nutrumpėjusių formų, kurios ypač paplitusios šnekamojoje kalboje, pvz., *laikrody*, *laikrodžiuos*, *šnekamojoj*. Bandant gauti informaciją apie žodžio *šnekamojoj* morfemas, sistema nurodo, kad duomenų bazėje tokio žodžio nėra, o 2015 metais sukurta ir viešai internete prieinama Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema (2 interneto nuoroda) teigia netgi klaidinančią informaciją: žodžių junginiui *šnekamojoj kalboj* parašo: „Pateiktas tekstas yra ne lietuvių kalba arba gramatiškai neteisingas“.

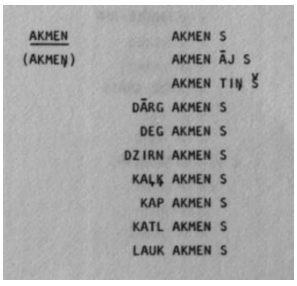
Todėl nuspręsta kurti lietuvių kalbos gramatikos informacinę sistemą, kurios tikslas – pradžioje sukaupti išsamius ir labai aukšto patikimumo duomenis apie visų lietuvių kalbos žodžių gramatinius požymius, o ateityje įtraukti ir sintaksės duomenis.

2. MORFEMINĖ ANALIZĖ

Lietuvių kalbos kompiuterizavimo darbai morfemikos srityje pradėti labai neseniai – pirmasis viešai internete prieinamas morfemikos žodynas pasirodė tik 2011 metais (Rimkutė, Kazlauskienė, Raškinis 2011). Todėl trumpai bus apžvelgtos ir kitų kalbų publikacijos, aprašančios žodžių skaidymą į morfemas.

2.1. Kitų kalbų morfemikos srities darbai

Latvių kalbos žodžių darybos žodynas išleistas 1985 metais (Metuzale-Kangere 1985). Jame morfemos atskiriamos viena nuo kitos tarpais ir šaknis išdėstoma stulpeliu. Šio žodyno pavyzdys pateiktas 1 paveikslėlyje. Naudojant tokį žodžių pavaizdavimą morfemomis turėtų iškilti problemų sudurtinių žodžių atveju, kai šaknys yra dvi ar net trys. Tada nebelieka priemonių kaip jas atskirti nuo priesagų ar galūnės (pvz., *kaipmat*, *tąsyk* ir kt. – antra šaknis užimtų galūnės poziciją).

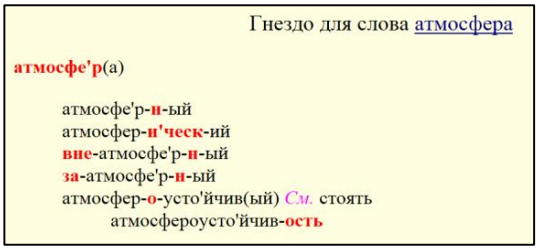


1 pav. Latvių kalbos darybinio žodyno su morfeminiu žodžių išskaidymu pavyzdys (Metuzale-Kangere 1985, 4).

Vėliau pasirodžiusiame čekų kalbos žodžių darybos žodyne šaknis išskiriama pasviraisiais brūkšneliais. 2 paveikslėlyje pateiktas šio žodyno pavyzdys (Sedlaček 2004, 1280). Čia jau nesunku vienareikšmiškai pavaizduoti ir sudurtinius žodžius. Visos morfemos taip pat atskiriamos tarpais.

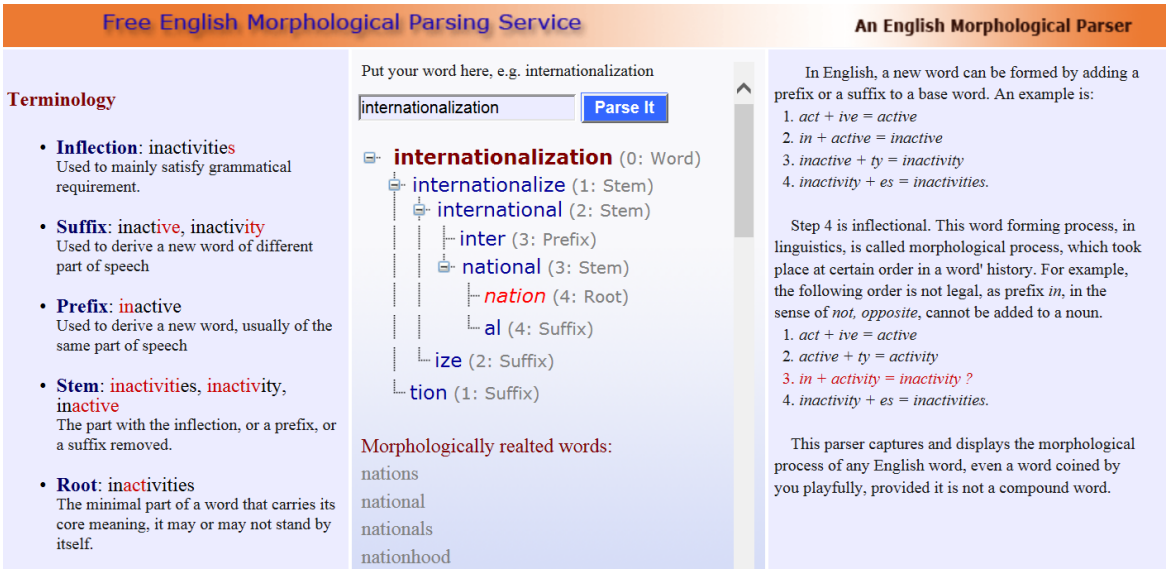
2 pav. Čekų kalbos darybinio žodyno su morfeminiu žodžių išskaidymu pavyzdys (Sedlaček 2004, 1280).

Rusų kalbos morfeminiame žodyne (3 interneto nuoroda) šaknis ir afiksai vaizduojami skirtingomis spalvomis (3 pav.).



3 pav. Rusų kalbos morfeminio žodyno pavyzdys (3 interneto nuoroda).

Internete viešai prieinamas anglų kalbos analizatorius (4 interneto nuoroda), nors ir vadinamas morfologiniu, pateikia morfeminę informaciją apie žodį. 4 paveikslėlyje parodytas anglų kalbos žodžio *internationalization* analizės rezultatas.



4 pav. Anglų kalbos žodžio *internationalization* morfeminė analizė (4 interneto nuoroda).

2.2. Morfemikos srities darbai, atlikti Lietuvoje

Pirmasis stambus lietuvių kalbos morfemikos kompiuterizavimo darbas buvo atliktas Matematikos ir informatikos institute 1992 m. Sukurtoje *Žodžių darybos ir morfemų duomenų bazėje* (ŽDMDB) (Murmulaitytė 2012, 96) sukaupta gana išsami informacija apie morfemas: kiekvienos rūšies morfemos užrašomos skirtingu šriftu. Pavyzdžiui, žodis *tikimybinis* vaizduojamas kaip parodyta 5 pav.: šaknis –

pastorintu šriftu, galūnė – paprastu, priesagos – pasvirusiu. Jei yra kelios priesagos, tarp jų dedami tarpeliai. Darybinei priesagai naudojamos didžiosios raidės. Šalia pateikiamas taip pat ir pamatinis žodis. 6-ame paveikslėlyje parodytas žodžio su priešdėliu – *užjūrinis* – pavyzdys. Priešdėlis užrašomas pasvirusiu pabrauktu šriftu.

bdv.	**tik** *im yb IN* is	tikimybė dkt.

5 pav. Žodžio *tikimybinis* pavaizdavimas *Žodžių darybos ir morfemų duomenų bazėje* (Murmulaitytė 2012, 98).

bdv.	*už* **jūr** *IN* is	užjūris dkt.

6 pav. Žodžio *užjūrinis* pavaizdavimas *Žodžių darybos ir morfemų duomenų bazėje* (Murmulaitytė 2012, 98).

Šioje duomenų bazėje sukaupta tikrai vertinga informacija. Labai blogai, kad ji nėra viešai prieinama, ir ja tegali naudotis patys autoriai. Ir panašu, kad darbai nėra tęsiami.

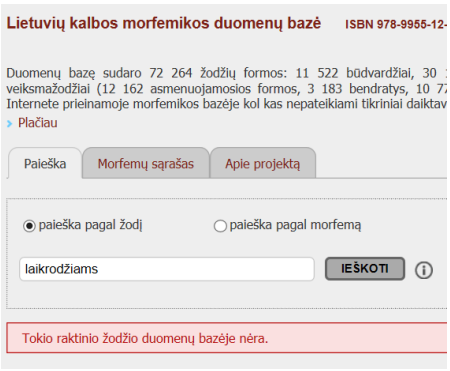
2011 metais pasirodė viešai internete prieinamas morfemikos žodynas, kuris sukurtas Vytauto Didžiojo universitete tekstyno pagrindu ir teapima tik jame esančius žodžius. Sunku suprasti, kodėl buvo pasirinktas toks neinformatyvus morfemų vaizdavimo būdas – jos atskiriamos viena nuo kitos brūkšneliais, visai nepateikiant jokios informacijos apie morfemos tipą ir vienodai vaizduojant skirtingos morfeminės struktūros žodžius – kai gausu gerų pavyzdžių tiek kitų kalbų, tiek lietuvių kalbos žodžių skaidyme į morfemas jau buvo anksčiau. Neaišku, kodėl nepasidomėta ir nepasinaudota tikrai gera ir vertinga patirtimi. Vienintelė priežastis turbūt – ribotos kompiuterių galimybės šioje srityje.

VDU darbai atlikti tyrinėjant žodžių morfeminę struktūrą apima 310 000 žodžių analizę (Rimkutė, Kazlauskienė, Raškinis 2011a, 7). Rezultatai pateikiami trijų tomų žodyne (5, 6 ir 7 interneto nuorodos), kur žodžiai išskaidyti morfemomis, ir jos atskirtos viena nuo kitos brūkšneliais. Taigi, išsamios informacijos jame trūksta. Kaip vieną iš pačių didžiausių trūkumų galima būtų paminėti informacijos apie morfemos tipą nebuvimą. Nors žodyno aprašyme sakoma, kad *-un-* laikoma priesaga žodyje *šunį* (Rimkutė, Kazlauskienė, Raškinis 2011, 7), tačiau žodyne jis pateikiamas tokios pat struktūros, kaip ir žodis *sutemos*: *š-un-s* (Rimkutė, Kazlauskienė, Raškinis 2011a, 686) ir *su-tem-os* (Rimkutė, Kazlauskienė, Raškinis 2011a, 665). Abu šie žodžiai sudaryti iš trijų morfemų, tačiau visai nėra informacijos apie tai, kad žodyje *šuns* pirma morfema yra šaknis, antra – priesaga, o žodyje *sutemos* pirma morfema yra priešdėlis, o antra – šaknis. Patys autoriai įvade nurodo, kad ateityje ketinama parengti daug išsamesnį žodyną. 2013 metais pasirodė viešai prieinama internete *Lietuvių kalbos morfemikos duomenų bazė* (1 Interneto nuoroda), tačiau autorių ketinimai nebuvo įvykdyti. Paruošta tik patogesnė paieška pateikiant morfemikos žodyne esančius duomenis, tačiau informacija nepasidarė nė kiek išsamesnė – tai, kas buvo žodynuose, perkelta į duomenų bazę, bet papildomai neatlikta nieko: žodis į morfemas skaidomas tuo pačiu principu – atskiriant jas brūkšneliais, kaip ir buvo žodyne. Pateikiamų duomenų apimtis taip pat išliko ta pati: žodžio išskaidymas morfemomis, jo lema, dažnumas ir gramatinė informacija (7 pav. ir 8 pav.). Tesiskiria tik informacijos išdėstymas ekrane, bet ne jos turinys. Žodžių kiekis taip pat nepadidėjo: tų žodžių, kurių nebuvo morfemikos žodyne, nėra ir morfemikos duomenų bazėje. Šiuos teiginius gerai pagrindžia pavyzdžiai. Žodyne yra šeši įrašai su žodžio *laikrodis* formomis (Rimkutė, Kazlauskienė, Raškinis 2011a, 332) (7 pav.). Duomenų bazėje taip pat tegalima gauti informaciją tik apie šias šešias žodžio formas. Tų formų, kurių nebuvo žodyne, pvz. *laikrodžiams* (7 pav.), nėra ir duomenų bazėje (1 Interneto nuoroda) (9 pav.). Ir negalima teigti, kad tai retai pasitaikantis žodis: sakinys, pvz., *Manoma, kad laikrodžiams prižiūrėti kasmet reikės iki 3 tūkst. litų* yra labai įprastas ir vartojamas, paimtas iš tekstyno (8 Interneto nuoroda).

laikraštis	10	laik-raš-t-is	laikraštis; dkt. vyr. g. vns. vard.
laikrodėlį	1	laik-rod-ėl-į	laikrodėlis; dkt. vyr. g. vns. gal.
laikrodėlis	1	laik-rod-ėl-is	laikrodėlis; dkt. vyr. g. vns. vard.
laikrodį	9	laik-rod-į	laikrodis; dkt. vyr. g. vns. gal.
laikrodis	9	laik-rod-is	laikrodis; dkt. vyr. g. vns. vard.
laikrodžiai	1	laik-rodž-iai	laikrodis; dkt. vyr. g. dgs. vard.; dkt. vyr. g. dgs. šauksm.
laikrodžio	4	laik-rodž-io	laikrodis; dkt. vyr. g. vns. kilm.
laikrodžiu	2	laik-rodž-iu	laikrodis; dkt. vyr. g. vns. įnag.
laikrodžius	1	laik-rodž-ius	laikrodis; dkt. vyr. g. dgs. gal.
laiku	61	laik-u	laikas; dkt. vyr. g. vns. įnag.
laikų	43	laik-ų	laikas; dkt. vyr. g. dgs. kilm.

7 pav. Morfemikos žodyne pateikiamų žodžio *laikrodis* formų sąrašas bei jų analizė (Rimkutė, Kazlauskienė, Raškinis 2011a, 332).

8 pav. Žodžio *laikrodį* paieškos morfemikos duomenų bazėje rezultatas (1 Interneto nuoroda).



9 pav. Žodžio *laikrodžiams* paieškos morfemikos duomenų bazėje rezultatas (1 Interneto nuoroda).

Paieška pagal morfemą taip pat pateikia tuos pačius duomenis apie žodžio morfeminę struktūrą – informacijos apie morfemos tipą nėra jokios: žodžiai *antakius* ir *antele* pavaizduoti kaip turintys tą pačią morfeminę struktūrą – sudaryti iš trijų morfemų (10 pav.), nors žodyje *antakius* pirma morfema *ant-* yra priešdėlis, o antra – šaknis, antrame gi žodyje – *antele* – pirmoji morfema *ant-* yra šaknis, o antroji – priesaga.

10 pav. Informacijos pateikimas apie žodžius *antakius* ir *antele* morfemikos duomenų bazėje (1 interneto nuoroda).

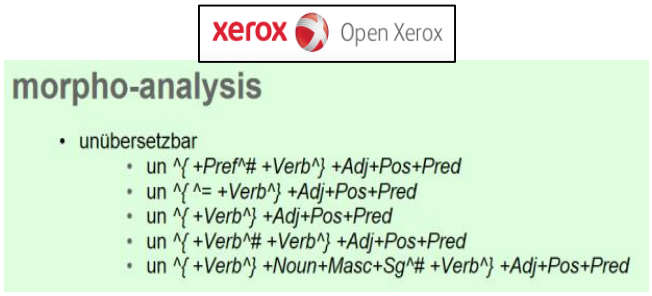
Ir tai yra viena priežasčių, kodėl buvo nuspręsta sukurti lietuvių kalbos gramatikos informacinę sistemą, apimančią išsamius gramatinius duomenis apie lietuvių kalbos žodžius bei sakinius. Ji bus laisvai prieinama internete ir skiriama plačiajam vartotojų ratui, todėl duomenys bus pateikiami populiariai: jais naudotis galės ir neturintys specialaus išsilavinimo žmonės. VDU išleistame morfemikos žodyne ir jo pagrindu paruoštoje morfemikos duomenų bazėje informacija gali būti naudinga tik gilias lituanistines žinias turintiems specialistams, kurie labai gerai žino žodžių skaidymą į morfemas. Tačiau neturintiems specialaus išsilavinimo žmonėms žodžių užrašymas atskiriant tam tikrus raidžių rinkinius brūkšneliais dažniausiai naudingos morfeminės informacijos nesuteikia.

3. MORFOLOGINIAI ANALIZATORIAI

Morfologiniai analizatoriai paprastai būna prieinami internete laisvai ir nurodo gramatinius duomenis apie pageidaujamą žodį. Tačiau daugelio kalbų atveju informacija pateikiama taip, kad ją suprasti gali tik kompiuterinės lingvistikos specialistai. Šiame skyriuje taip pat bus apžvelgiami morfologiniai analizatoriai, sukurti kitoms kalboms ir Lietuvoje atlikti darbai morfologinės analizės tematika. Ir čia jau reikia pasakyti, kad kai kuriais klausimais Lietuvos padėtis šioje srityje yra geresnė, nei kitose šalyse. 2015 metais VDU sukurtoje Lietuvių kalbos sintaksinės ir semantinės analizės informacinėje sistemoje informacija pateikiama aiškiai, populiariai, pilnais žodžiais.

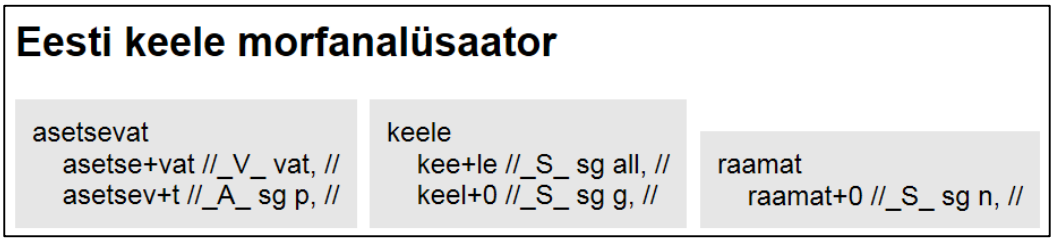
3.1. Kitų kalbų morfologiniai analizatoriai

Vienas iš morfologinio analizatoriaus pateikiamos informacijos pavyzdžių galėtų būti OPEN XEROX išanalizuotas vokiečių kalbos žodis *unübersetzbar* (*neišverčiamas*), kurio rezultatai parodyti 11 paveikslėlyje (9 interneto nuoroda). Blogiausia, kad niekur nėra paaiškinta, ką reiškia žodžio analizės schemoje panaudoti ženklai. Taigi, tokiu pavidalu pateikta informacija, nors ir prieinama viešai, tegali būti naudinga tik nedaugeliui specialistų, susipažinusių su programinės įrangos dokumentacija.



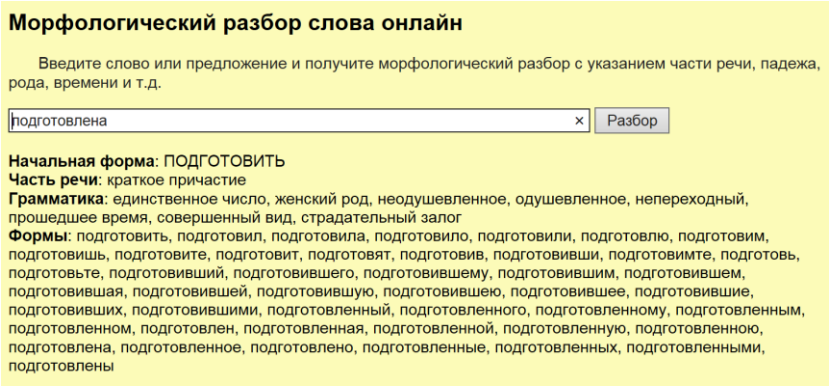
11 pav. Vokiečių kalbos žodžio *unübersetzbar* (*neišverčiamas*) morfologinė analizė atlikta Open Xerox morfologiniu analizatoriumi (9 interneto nuoroda).

Panašiai duomenys pateikiami ir estų kalbos morfologiniame analizatoriuje (10 interneto nuoroda). 12 paveikslėlyje parodyta daugiareikšmio žodžio *asetsevat,* bei žodžių *keele* (*kalba*) ir *raamat* (*knyga*) morfologinė analizė. Čia taip pat nėra paaiškinimų, ką reiškia analizėje panaudoti ženklai.



12 pav. Estų kalbos daugiareikšmio žodžio *asetsevat* bei žodžių *keele* (*kalba*) ir *raamat* (*knyga*) morfologinės analizės pavyzdys (10 interneto nuoroda).

Suprantamiau duomenis pateikia rusų kalbos morfologinis analizatorius (11 interneto nuoroda). Jo analizės pavyzdys parodytas 13 paveikslėlyje. Net ir jokio pasiruošimo neturintis vartotojas be problemų gali suprasti visą pateiktą informaciją. Ta pačia rusų kalba parašytas nurodomos informacijos tipas: pradinė forma, kalbos dalis ir kt.



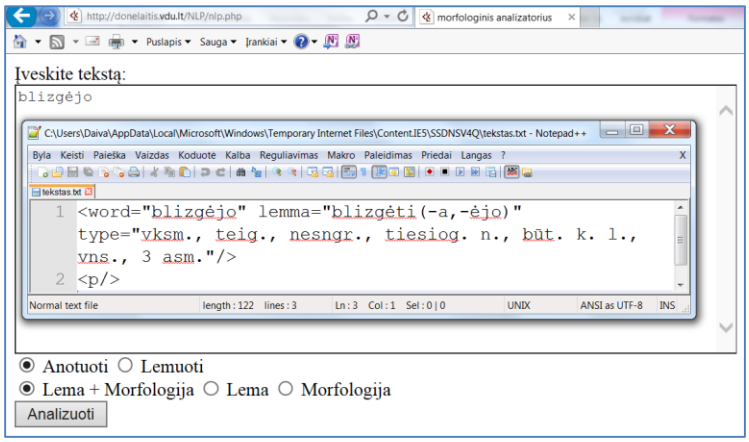
13 pav. Rusų kalbos žodžio *подготовлена* morfologinė analizė (11 interneto nuoroda).

Belieka aptarti Lietuvoje atliktus darbus kompiuterizuojant morfologiją. Pradėti jie buvo Matematikos ir informatikos institute Vilniuje ir vėliau tęsiami Vytauto Didžiojo universitete Kaune.

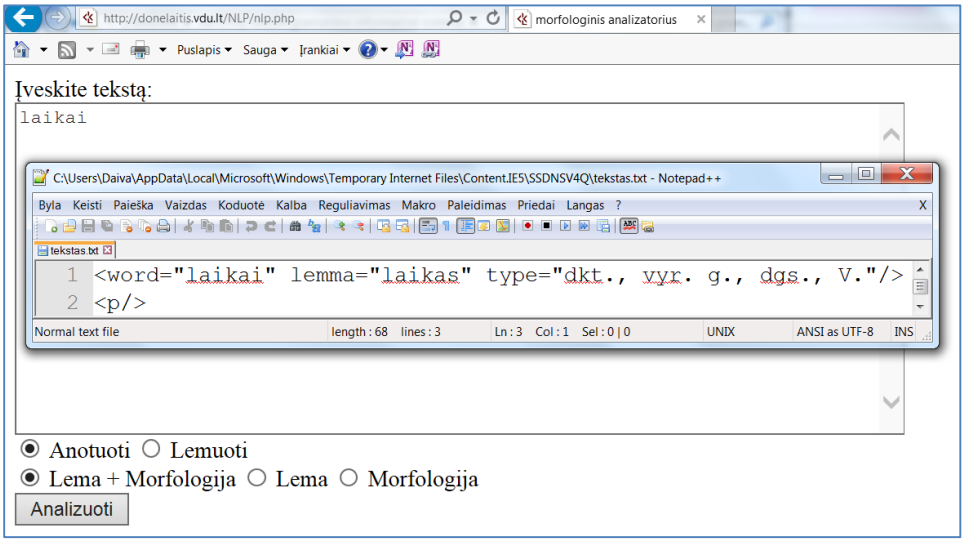
3.2. Lietuvių kalbos morfologinė analizė

Pirmasis lietuvių kalbos morfologinis analizatorius sukurtas Matematikos ir informatikos institute 2000 metais. Tai lietuvių kalbos morfologinės analizės ir sintezės programinė įranga – lemuoklis (Zinkevičius 2000). Didžiausias jo privalumas yra tai, kad jis atpažįsta ir pateikia informaciją apie visus lietuvių kalboje esančius žodžius. Kaip trūkumą galima paminėti pateikiamus perteklinius žodžius, kurių nėra lietuvių kalboje, pvz., žodžiui *blizgėjo* kaip trečias variantas nurodomas daiktavardžio ***blizgėjas* kilmininko linksnis.

Šios programinės įrangos pagrindu Vytauto Didžiojo universitete 2008 m. sukurtas morfologinis analizatorius prieinamas viešai internete (12 interneto nuoroda), tačiau informacija pateikiama naudojant sutrumpinimus ir anglų kalbos žodžius, ir dėl to plačiajai visuomenei toks formatas nėra labai patogus naudotis. Analizatorius turi dvi funkcijas, kurios pavadintos „Anotuoti“ ir „Lemuoti“. Vykdant funkciją „Anotuoti“ pateikiama tik viena reikšmė net ir daugiareikšmių žodžių atveju, pvz., 14 paveikslėlyje pateiktas žodžio *blizgėjo* analizės rezultatas, kuriame nėra daugiskaitos varianto; ir negalima sakyti, kad šio žodžio daugiskaita žymiai rečiau vartojama nei vienaskaita. Žodžiui *laikai* pateikiama taip pat tik daiktavardžio forma, atmetant veiksmažodžio vienaskaitos antro asmens variantą (15 pav.). Taigi, visai ignoruojamas žodžių daugiareikšmiškumas.

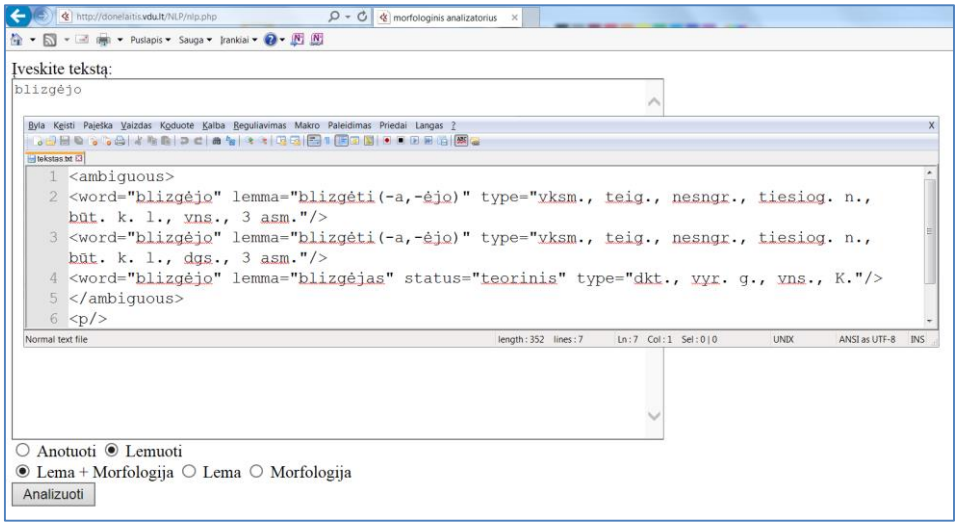


14 pav. Lietuvių kalbos žodžio *blizgėjo* morfologinė analizė vykdant analizatoriaus funkciją „Anotuoti“ (12 interneto nuoroda).



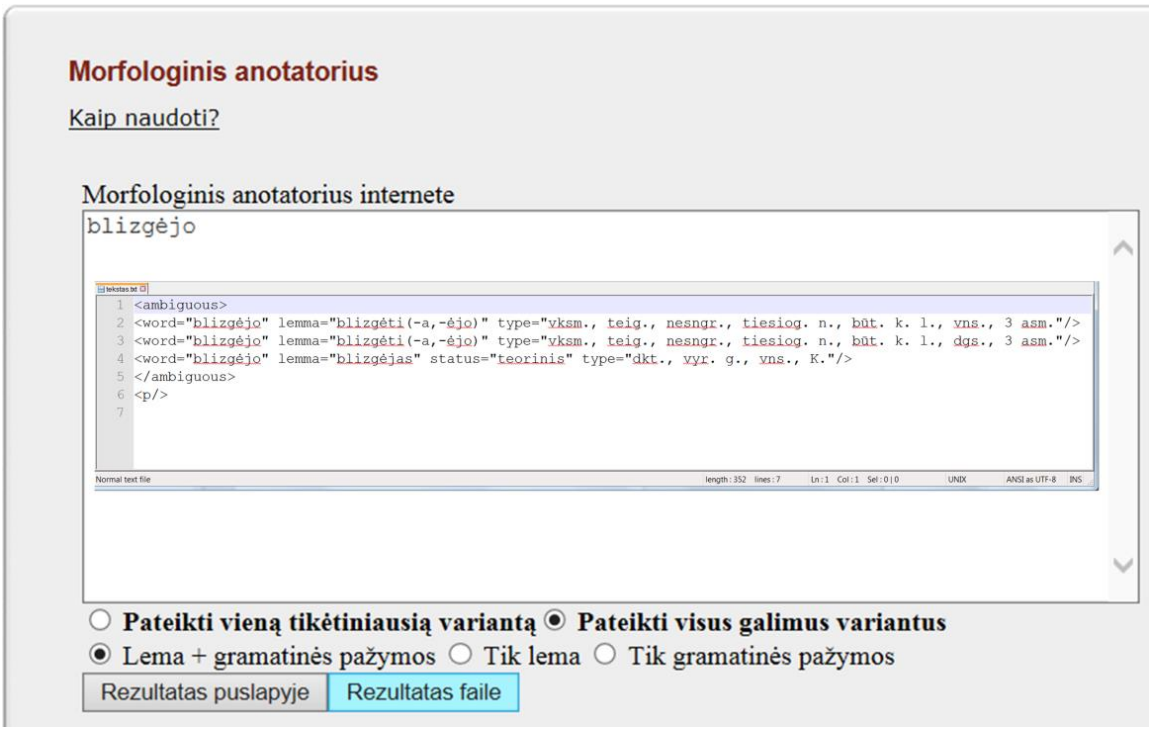
15 pav. Lietuvių kalbos žodžio *laikai* morfologinė analizė vykdant analizatoriaus funkciją „Anotuoti“ (12 interneto nuoroda).

Kita funkcija „Lemuoti“ pateikia visus galimus daugiareikšmių žodžių variantus, bet tuo pačiu ir lietuvių kalboje neegzistuojančius žodžius, tokius kaip, pvz., ***blizgėjas* (16 pav.). Informacijoje apie žodį nurodytas statusas „teorinis“ lietuvių kalbos žodžiu jo nepadaro. Kad tam tikras raidžių rinkinys būtų kokios nors kalbos žodis, jis turi atitikti tris reikalavimus: a) jis turi turėti garsinę struktūrą, b) turi egzistuoti tikrovėje daiktas ar reiškinys, kurį tas žodis pavadina ir c) žmogaus sąmonėje turi būti to daikto ar reiškinio atspindys (Jakaitienė 1980, 16). Raidžių rinkinys „***blizgėjas*“ tenkina tik pirmąjį reikalavimą – turi garsinę struktūrą, kitų dviejų reikalavimų jis neatitinka: nėra nei daikto ar reiškinio tikrovėje, kurį jis pavadintų, nei jo atspindžio žmogaus sąmonėje. Nei tas lietuvis, kuris sako, t.y. ištaria garsų rinkinį „***blizgėjas*“, nei tas, kuris jį girdi, nežino, ką tai reiškia. Vadinasi, tai nėra lietuvių kalbos žodis.



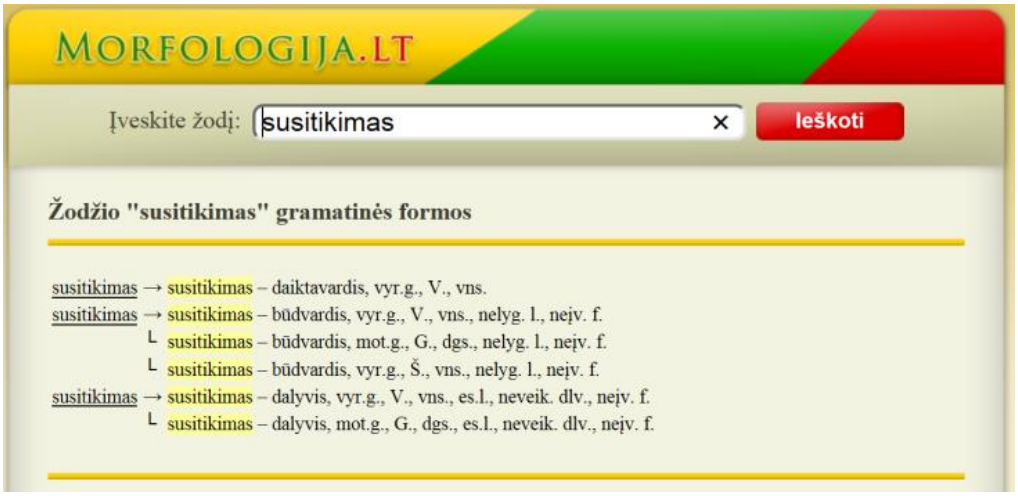
16 pav. Lietuvių kalbos žodžio *blizgėjo* morfologinė analizė vykdant analizatoriaus funkciją „Lemuoti“ (12 interneto nuoroda).

Pateikiamos informacijos kiekis ir pobūdis nepasikeitė ir sukūrus morfologinį anotatorių (13 interneto nuoroda). Pasirinkus funkciją „Pateikti visus galimus variantus“ gaunami tie patys analizės rezultatai (17 pav.) su neegzistuojančiu lietuvių kalboje žodžiu ***blizgėjas*. Šiuo atveju padėtis tokia pat kaip ir su morfemikos žodynu, kai jo pagrindu buvo sukurta duomenų bazė: informacija liko ta pati, pasikeitė tik jos pateikimo forma.



17 pav. Lietuvių kalbos žodžio *blizgėjo* morfologinė analizė vykdant anotatoriaus funkciją „Pateikti visus galimus variantus“ (13 interneto nuoroda).

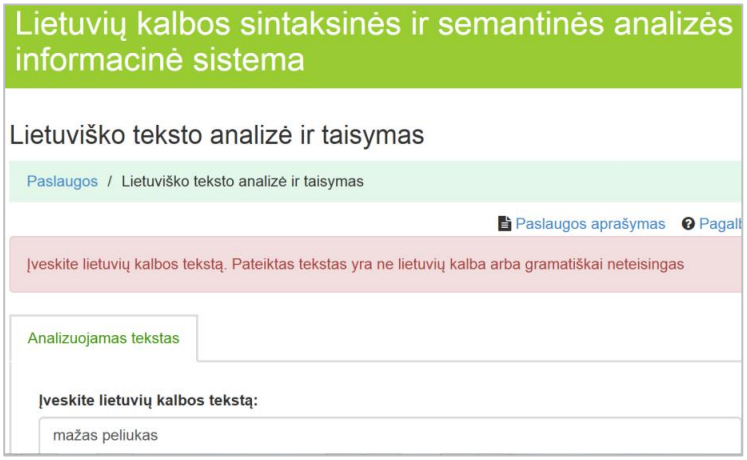
Lygiai taip pat pertekliniai žodžiai pateikiami ir tinklalapyje MORFOLOGIJA.LT (14 interneto nuoroda). Žodžiui *susitikimas* nurodoma ne tik daiktavardžio forma, bet ir būdvardžio bei dalyvio variantai, kurie nėra vartojami lietuvių kalboje. 18 paveikslėlyje parodyta žodžio *susitikimas* analizė. Dabartinės lietuvių kalbos žodynas žodžiui *susitikimas* pateikia tik daiktavardžio variantą (Keinys et al. 1993, 778) ir net neįmanoma įsivaizduoti, kokį žodį galėtų pažymėti toks būdvardis. Akademinis lietuvių kalbos žodynas (Naktinienė at al. 2008) taip pat nepateikia žodžiui *susitikimas* būdvardžio varianto. Net ir pagal lietuvių kalbos gramatikos taisykles būdvardžių vediniai su priesaga *-imas* galimi tik iš būdvardžių, ir tai yra tokie būdvardžiai, kurių pamatiniai žodžiai retai bevartojami, pvz., *artimas, gretimas, svetimas* (Ambrazas et al. 1997, 201). Tritomėje gramatikoje taip pat pateikiami būdvardžių vediniai su priesaga *-imas* tik iš būdvardžių: *artimas, tolimas* ir kt. (Uvydas et al. 1965, 556).



18 pav. Lietuvių kalbos žodžio *susitikimas* morfologinė analizė tinklapyje MORFOLOGIJA.LT (14 interneto nuoroda).

VDU naujai sukurta ir 2015 metais pateikta viešai internete *Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema* (2 interneto nuoroda) pateikia ir kai kuriuos morfologinius duomenis. Džiugu, kad informacija nurodoma populiariai, be sutrumpinimų. Sistema jau nebedaro tų klaidų, kurios dar yra

likusios analizatoriuje ir anotatoriuje. Žodžiui *susitikimas* pateikiama vien daiktavardžio forma ir nebenurodomas būdvardis, žodžiui *blizgėjo* daiktavardžio ***blizgėjas*** formos taip pat nėra. Tačiau, kai sistema pradeda teigti klaidinančią informaciją, pvz., kad *šnekamojoj kalboj* yra ne lietuviškas tekstas arba netgi pateikus analizei žodžių junginį *mažas peliukas* gaunamas pranešimas, kad „Pateiktas tekstas yra ne lietuvių kalba arba gramatiškai neteisingas“ (19 pav.), tai jau pradeda kilti labai didelių abejonių dėl visos sistemos patikimumo, nes tokių „nelietuviškų“ žodžių yra begalė – *toliaregis, apyrankė, nebeatsinešdavau* ir t.t.



19 pav. Lietuvių kalbos žodžių junginio *mažas peliukas* morfologinė analizė (2 interneto nuoroda).

Išvada galėtų būti tokia: panaikinant klaidas dėl perteklinių, lietuvių kalboje neegzistuojančių žodžių pateikimo, tuo pačiu prarandami milžiniški kiekiai ir naudingos informacijos. Sistemoje, kuri nebepateikia lietuvių kalboje nesančių žodžių, tokių kaip ***blizgėjas*** ar žodžio *susitikimas* būdvardžio varianto, labai daug taisyklingų ir dažnai vartojamų lietuvių kalbos žodžių pasidaro „tekstas ne lietuvių kalba“. Taigi lieka klausimas: kiek galima tikėti tokios sistemos duomenimis ir jos teikiama informacija? Apibendrinant galima būtų pasakyti: VDU darbai, tobulinant morfologinio analizatoriaus ir anotatoriaus veikimą, nepasiteisino, sintaksinės ir semantinės analizės informacinė sistema, sukurta 2015 m., daro dar didesnes klaidas nei analizatorius, pateiktas viešam naudojimui 2008 m. Todėl tikslinga kurti iš principo naują lietuvių kalbos gramatikos informacinę sistemą, teikiančią išsamią ir patikimą informaciją, kur būtų numatytos išplėtimo galimybės ir sistemos papildymas bei patobulinimas nesukeltų naujų klaidų, kurių nebuvo ankstesnėse versijose (kaip kad yra VDU morfologinės analizės atveju: 2015 m. pasirodęs patobulintas morfologinės analizės variantas daro klaidas, kurių nebuvo 2008 m. versijoje – labai dideliam kiekiui žodžių ne tik nepateikia gramatinių duomenų, bet netgi teigia klaidinančią informaciją, t.y. nurodo, kad tai ne lietuvių kalbos tekstas, ir tai yra tokie žodžiai, kuriems ankstesnė versija gramatinę informaciją pateikia).

Pasaulio praktikoje yra pasitaikę atvejų, kai sukurtos sistemos buvo atmetamos, nes jų tobulinimas ar išplėtimas pasirodė brangesni, nei naujos sistemos sukūrimas, pvz., automatinio vertimo sistema TAUM-AVIATION (Isabelle, Bourbeau 1985). Ji buvo kuriama Kanadoje 1976–1980 m. ir skirta labai siauros tematikos tekstams – apie lėktuvų hidrauliką. Vertimo metodika rėmėsi tuo, kad šios srities tekstų sintaksė yra ribota, ir su žodynu, apimančiu apie 70 000 įrašų, buvo galima gana neblogai išversti tokios rūšies tekstus. Tačiau, pabandžius sistemą išplėsti, pasirodė, kad tai yra per brangu, ir 1980 m. darbai buvo nutraukti (Schwanke 1991, 49).

Panašus atvejis dabar yra VDU morfologinė analizė: paskutinė (2015 metais pasirodžiusi) morfologinės analizės versija dirba žymiai blogiau nei ankstesnės (2008 metų).

4. GRAMATIKOS INFORMACINĖ SISTEMA

Aptarti jau atlikti lietuvių kalbos gramatikos kompiuterizavimo darbai ir parodyta, kad jų kokybė netenkina vartotojų poreikių gauti tikslią, patikimą ir išsamią informaciją. Todėl ir buvo nuspręsta kurti lietuvių kalbos gramatikos informacinę sistemą bei jos portalą, kuris viename tinklapyje pateiktų išsamią ir įvairiapusę informaciją apie lietuvių kalbos gramatiką bei žodžių gramatinius požymius.

Pagrindinė problema, dėl kurios imtasi kurti lietuvių kalbos gramatikos informacinę sistemą, buvo tai, kad morfemikos žodyne, nors informacija ir pateikta visiems suprantamu formatu, tačiau duomenys kartais būna netgi klaidinantys, kai skirtingą morfeminę struktūrą turintys žodžiai vaizduojami vienodai. Pavyzdžiui, kaip tos pačios struktūros žodžiai pavaizduoti *ant-el-e* ir *ant-ak-ius* (Rimkutė, Kazlauskienė, Raškinis 2011a, 28). Morfemikos duomenų bazėje padėtis liko nepakitusi (10 pav.). Bet juk jų morfeminė sudėtis skiriasi: žodyje *ant-el-e* pirmoji morfema yra šaknis, antroji – priesaga, o žodyje *ant-ak-ius* pirmoji morfema, kuri sudaryta iš to paties raidžių rinkinio *ant-*, yra priešdėlis, o antroji – šaknis. Šios problemos sprendimas, kuriant gramatikos informacinę sistemą, siūlomas toks: kiekvieną morfemos tipą vaizduoti skirtinga spalva: ant-el-e, ir ant-ak-ius. Kitas pavyzdys galėtų būti žodžiai *laikrodis* ir *laikmenoje*. Jie morfemikos duomenų bazėje taip pat pateikiami kaip turintys tą pačią struktūrą: *laik-rod-is* ir *laik-men-oje* (1 interneto nuoroda). Tačiau pirmas žodis *laikrodis* turi dvi šaknis ir antroji morfema jame yra antra šaknis, kai tuo tarpu žodyje *laikmenoje* antra morfema yra priesaga. Šių žodžių pavaizdavimas gramatikos informacinėje sistemoje atrodys taip: laik-rod-is ir laik-men-oje.

Gramatikos informacinė sistema tai tarsi popieriuje spausdintų gramatikų inversija. Paprastai gramatikos vadovėliuose pateikiamos taisyklės, kurios tinka tam tikrai žodžių grupei, bet apsiribojama vien keliais pavyzdžiais ir neišvardijami visi žodžiai, vartojami pagal tą taisyklę. Kuriant gramatikos informacinę sistemą į kalbą bandoma žiūrėti kitu aspektu: ne iš gramatinių kategorijų pozicijos, bet iš žodžio pozicijos, t.y. išeities taškas turi būti ne gramatikos taisyklė ir kaip jos iliustracija pateikti keli žodžiai, kuriems ji tinka, bet pats žodis turi būti pagrindas ir iš gramatikos išrenkami duomenys apie jį pagal visas su juo susijusias taisykles.

Kuriant gramatikos informacinę sistemą siekiama, kad ji būtų patogi plačiajam vartotojų ratui. Stengiamasi sukaupti išsamią informaciją apie lietuvių kabos žodžių gramatinius požymius ir pateikti ją visiems suprantamai, pilnais žodžiais, be sutrumpinimų. Bus nurodytas ne tik morfemos tipas, bet ir specifiniai jos požymiai, nepriklausantys nei nuo žodžio gramatinės formos, nei nuo jo reikšmės, pvz., priesaga: kaitybinė, darybinė, mažybinė; priešdėlis: dalelytinės kilmės, prielinksninės kilmės, tarptautinis; galūnė: įvardžiuotinė, nutrumpėjusi ir pan.

Sukurtą lietuvių kalbos gramatikos informacinę sistemą planuojama ateityje sujungti su Raštija.lt skaitmeninių išteklių žodynais. Tai galės būti vokiečių portalo CANOONET – Deutsche Wörterbücher und Grammatik (15 interneto nuoroda), kuriame pateikiami kartu žodynai ir gramatika, analogas.

Ruošiant informacinę sistemą reikia išspręsti du pagrindinius uždavinius: sukurti gramatikos duomenų bazę ir patogų vartotojui informacijos pateikimo būdą. Apie tai plačiau rašoma tolesniuose skyriuose.

4.1. DUOMENŲ BAZĖ

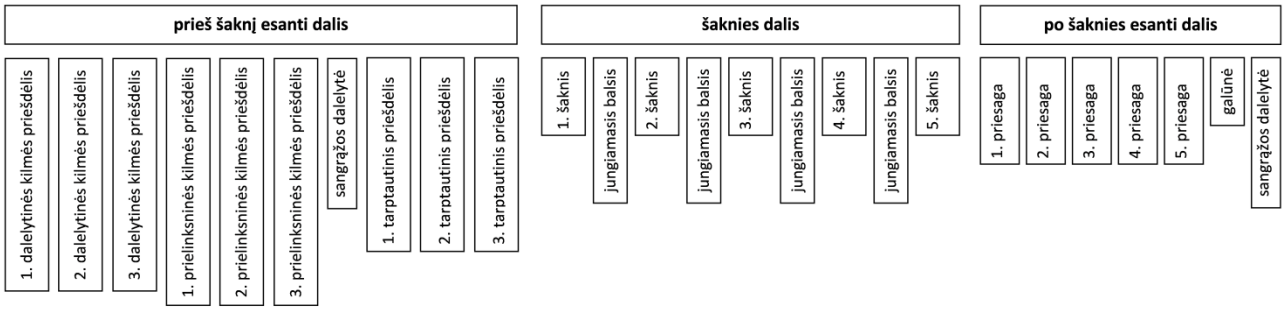
Informaciją apie lietuvių kalbos žodžius bei jų gramatinius požymius patogiausia kaupti duomenų bazėje. Iš jos paimti duomenys galės būti panaudoti įvairiai: XML formatas yra patogus apdorojant kalbą kompiuteriu; plačiajai visuomenei bus sukurta sąsaja su vartotoju, kuri labai populiariai ir suprantamai pateiks visą informaciją internete.

Lietuvių kalbos žodžio formatas

Kad būtų lengviau struktūriškai aprašyti morfologinius duomenis apie žodį, buvo sudarytas apibendrintas lietuvių kalbos žodžio formatas, apimantis visus galimus lietuvių kalbos žodžių struktūros variantus. Kiekvienas į duomenų bazę įtraukiamas žodis talpinamas į apibendrintą formatą.

Pirmoje pakopoje lietuvių kalbos žodis skaidomas į prieš šaknį esančią dalį, šaknies dalį ir po šaknies esančią dalį. Prieš šaknį esančiai daliai pavadinti buvo atsisakyta žodžio „priešdėliai“, nes daugelis kalbininkų dalelytės *si* nelaiko priešdėliu, kai ji stovi prieš šaknį. Šaknies dalis apima vieną ar kelias (sudurtinių žodžių atveju) šaknis su jungiamaisiais balsiais. Apibendrintame lietuvių kalbos žodžio formate šaknims skirtos 5 pozicijos atsižvelgiant į tarptautinių žodžių vartojimą. Tyrinėjant sudurtinius lietuvių kalbos žodžius didžiausias pastebėtas šaknų kiekis buvo trys šaknys – *sienlaikraštis*. Po šaknies esanti dalis apima priesagas, galūnę ir sangrąžos dalelytę *si*. Galūnė į atskirą žodžio dalį nebuvo išskirta, nes ji visada žodyje būna tik viena, ko negalima pasakyti apie priesagas.

Išsamūs lietuvių kalbos priešdėlių tyrimai atlikti Lietuvių kalbos institute. Nustatyta, kad dalelytinės kilmės priešdėliai visada išsidėsto žodžio pradžioje ir niekada nebūna įsiterpę tarp prielinksninės kilmės priešdėlių (Šveikauskienė 2015, 196). Lietuvių kalbos gramatika teigia, kad „sangrąžos formantas *si* visada eina tarp priešdėlio ir šaknies“ (Ambrazas et al. 1997, 283). Tačiau tarptautiniai priešdėliai eina po dalelytės *si* (*nebesusikondensavo, nesikoncentruoja*). Dalelytinės kilmės priešdėlių žodyje negali būti daugiau kaip trys. Taip pat nebuvo pastebėta žodžių, turinčių daugiau nei tris prielinksninės kilmės priešdėlius. Todėl sudarant apibendrintą lietuvių kalbos žodžio formatą prieš šaknį esančiai daliai skirta 10 pozicijų: po tris dalelytinės kilmės, prielinksninės kilmės ir tarptautiniams priešdėliams bei viena pozicija – sangrąžos dalelytei *si*, kuri talpinama tarp lietuviškų ir tarptautinių priešdėlių. Apibendrintas lietuvių kalbos žodžio formatas pateikiamas 20 paveikslėlyje.



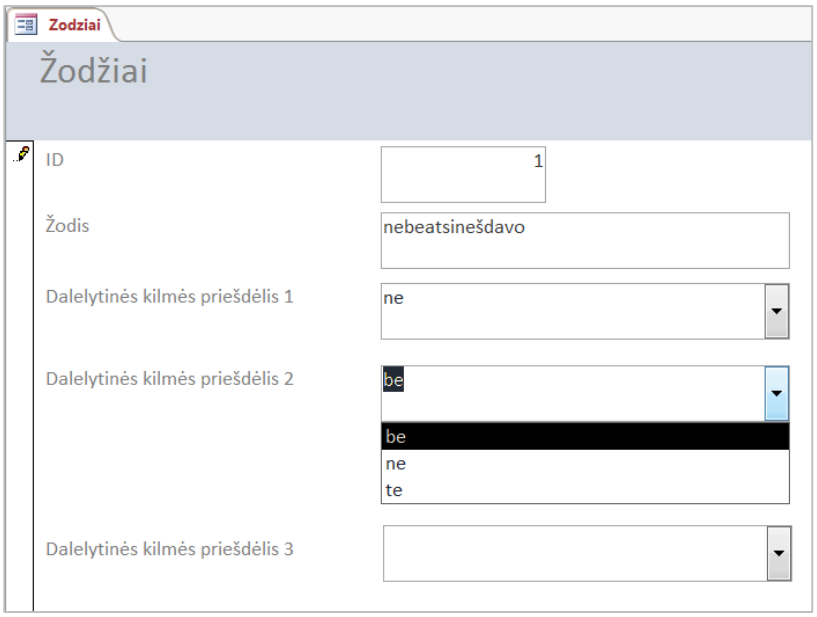
20 pav. Apibendrintas lietuvių kalbos žodžio formatas.

Apie kiekvieną morfemą bus pateikiama papildoma informacija, ne tik jos pavadinimas, pvz., jei šaknis turi infiksą ar balsių kaitą, tai atsispindės ir duomenų bazėje. Taip pat bus galima atlikti paiešką žodžių, turinčių vieną ar kitą gramatinį požymį.

4.1.1. Informacijos suvedimas į duomenų bazę

Kuriant duomenų bazę siekiama kuo didesnio tikslumo ir duomenų patikimumo. Patikimumui užtikrinti naudojama daug žmogaus darbo. Duomenys apie pradinę žodžio formą (lemą) bus suvedami rankomis. Visos likusios formos generuojamos automatiškai panaudojant lietuvių kalbos žodžių morfologinės sintezės programinę įrangą, kuri 2000 metais buvo sukurta Matematikos ir informatikos institute Vilniuje (Zinkevičius 2000). Šios programinės įrangos darbe nebuvo pastebėta klaidų sintezuojant žodžio kaitybines formas, kai duota lema ir nurodyta, kokia gramatinė forma turi būti sugeneruota. Todėl visos likusios aprašomo žodžio gramatinės formos ir duomenys apie jas bus suvedami į duomenų bazę automatiškai.

Siekiant duomenų tikslumo įvertinama ir tai, kad labai didelė ir svarbiausia bei atsakingiausia darbo dalis bus atliekama rankomis, o žmonės daro „žioplas“ klaidas. Kad būtų jų išvengta, naudojamos apsaugos priemonės, neleidžiančios žmogui suklysti. Pavyzdžiui, suvedant žodį jis bus iš karto skaidomas į morfemas, kaip parodyta apibendrinto žodžio formato paveikslėlyje (20 pav.) ir, kad būtų išvengta korektūros klaidų, morfema turės būti pasirenkama iš sąrašo. Dalelytinės kilmės priešdėlių lietuvių kalboje yra tik trys, vadinasi, suvedant informaciją apie žodį į pirmąsias tris pozicijas (20 pav.) tebus galima įrašyti vieną iš trijų morfemų *te-*, *be-*, *ne-*. Kaip atrodo darbo laukas užpildant duomenų bazę, parodyta 21 paveikslėlyje, kuriame pateikiamas dalelytinės kilmės priešdėlių suvedimo fragmentas.



21 pav. Darbo lauko fragmentas dalelytinės kilmės priešdėliams suvesti.

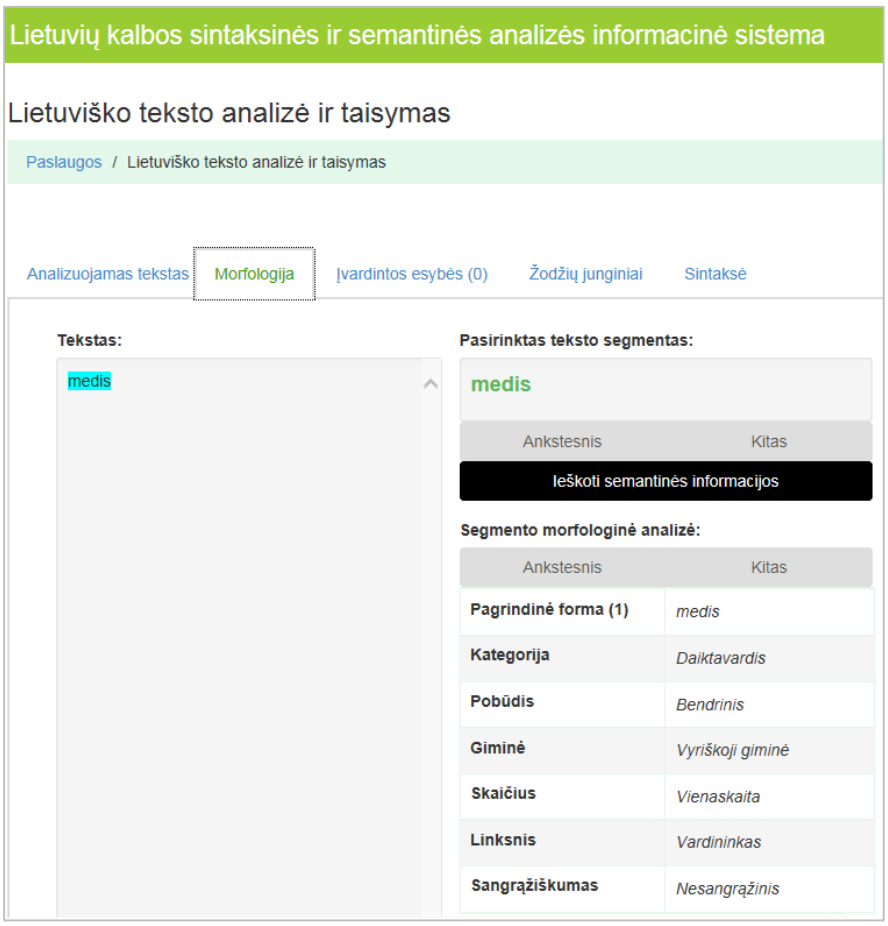
Prielinksninės kilmės priešdėlių sąrašas bus ilgesnis – pagrindinių priešdėlių yra 17 ir dar keli jų variantai, pvz., *ap-*, *api-*, *apy-* ir kt. Iš sąrašo į duomenų bazę pasirenkant bus suvedamos priesagos ir galūnės. Tik šaknis paliekama suvesti iš klaviatūros rankomis. Taip pat iš sąrašo bus pasirenkama informacija ir apie morfemų savybes, pvz., šaknies balsių kaita, infiksas, galūnė: įvardžiuotinė, nutrumpėjusi ir pan.

Ateityje, kuriant sintaksinę informacinės sistemos dalį, prie morfologinių duomenų bus pridedami tam tikri leksinės semantikos požymiai, kurie gali turėti įtakos nustatant žodžio sintaksinę funkciją, pvz., laiko požymis: *skaityti knygą* ir *skaityti naktį* – tik semantinis laiko požymis, kuris bus priskirtas žodžiui *naktis* ir kurio neturės žodis *knyga*, leis vienareikšmiškai nustatyti kompiuteriu papildinio ir aplinkybės funkciją sakinių *Visą naktį ji skaitė tą knygą* ir *Visą knygą ji perskaitė tą naktį* galininkams. Kadangi kuriama informacinė sistema kaupia duomenis, susijusius su lietuvių kalbos gramatika, todėl labai išsamios semantinės informacijos nebus pateikta, nes semantika nėra gramatikos dalis. Semantikos duomenys bus įtraukiami tik tie, kurie tarnauja sintaksės reikmėms, t.y. padeda vienareikšmiškai nustatyti kompiuteriu sakinio dalį.

4.2. Informacijos pateikimas vartotojui

Kadangi informacinė sistema skirta plačiajai visuomenei, todėl pateikiant duomenis nebus naudojami sutrumpinimai – visa gramatinė informacija bus pateikiama pilnais žodžiais. Siekiant kuo didesnio vaizdumo atskiriems morfemų tipams naudojamos skirtingos spalvos. Įvertinant VDU *Lietuvių kalbos sintaksinės ir semantinės analizės informacinės sistemos* duomenų pateikimo būdą, reikia pasakyti, kad jis nėra optimalus. Kad vardininkas yra linksnis, o vienaskaita yra skaičius, reikia pasakyti kompiuteriui, bet

ne žmogui. Žmogus šią informaciją ir taip žino. Todėl gramatikos informacinėje sistemoje pasirenkamas lakoniškesnis žodžių gramatinių požymių pateikimo būdas: jie surašomi ištisiniu tekstu, visi viename laukelyje ir svarbiausia – nenurodant požymių, kurių žodis neturi, pvz., netikslinga daiktavardžiui *medis* nurodyti, kad jis yra nesangrąžinis (22 pav.). Tai nereikalingas balastas. Lygiai taip pat netikslinga kiekvienam žodžiui nurodyti, kad jo šaknyje nėra infikso ar nevyksta balsių kaita. Tokio tipo požymiai gramatikos informacinėje sistemoje bus pateikiami tik prie tų žodžių, kurie juos turi, pvz., žodžiui *nebeatsinešdavau* bus nurodyta, kad jis yra sangrąžinis. Pateikiant duomenis apie žodį bus nurodomi tik tie gramatiniai požymiai, kurie jam būdingi, o apie kitus, kurių analizuojamas žodis neturi, net neužsimenama.



22 pav. Žodžio *medis* analizės rezultatas (2 interneto nuoroda).

Taigi toks morfologinių duomenų pateikimas, koks yra VDU sukurtoje *Lietuvių kalbos sintaksinės ir semantinės analizės informacinėje sistemoje*, plačiajam vartotojų ratui nėra optimalus. Todėl gramatikos informacinėje sistemoje siūlomas labiau vartotojui priimtinas morfologinių ir morfeminių duomenų apie žodį atvaizdavimo būdas.

4.2.1. Dviejų tipų informacija apie žodį

Duomenų bazėje kaupiama ir vartotojui populiariai pateikiama dviejų tipų informacija apie žodį: morfologinė ir morfeminė. Morfologinėje dalyje nurodomi duomenys apie visą žodį – kalbos dalis ir su ja susijusių morfologinių kategorijų gramatiniai požymiai: daiktavardžiui – linksnis, skaičius, giminė ir t.t., veiksmažodžiui – laikas, asmuo, skaičius, nuosaka ir kt. Taip pat nurodoma žodžio pradinė forma, o vediniams bei dūriniams – dar ir pamatiniai žodžiai.

Morfeminėje dalyje vaizdžiai parodoma žodžio struktūra pateikiant ne tik jo suskaidymą į morfemas, bet ir nurodant išsamią informaciją apie kiekvieną morfemą. Skirtingiems morfemų tipams

naudojamos atitinkamos spalvos, papildomai pateikiant tikslesnes pačios morfemos charakteristikas, pvz., priesaga: darybinė, kaitybinė; galūnė: įvardžiuotinė, nutrumpėjusi ir pan.

Informacinėje sistemoje pateikiamų duomenų apimtis atitinka maždaug vienatomėje (Ambrazas et al. 1997) ir tritomėje (Ulvydas et al. 1965, 1971, 1976) gramatikose išnagrinėtus klausimus.

4.2.2. Duomenų išdėstymas ekrane

Ekrane informacija pateikiama suskirstant langą į keturias sritis. Pirmoji – skirta vartotojui įvesti žodį, apie kurį jis pageidauja gauti išsamius gramatinius duomenis. Antroje srityje pateikiama pati bendriausia informacija apie įvestą žodį – jo pradinė forma ir pamatiniai žodžiai sudurtinių bei išvestinių žodžių atveju. Trečia sritis apima morfologinius duomenis. Jie pateikiami laukelyje nurodant ištisiniu tekstu kalbos dalį bei jos morfologinių kategorijų gramatinius požymius. Ketvirtoje srityje talpinamas paveikslėlis, vaizduojantis žodžio morfeminę struktūrą. Kiekviena morfema įrašoma į spalvotą rėmelį pagal jos tipą. Morfemos pavadinimas pateikiamas pilnu žodžiu. Jei nagrinėjamo žodžio morfemos turi tam tikrų požymių, jie parašomi taip pat pilnais žodžiais. 23 paveikslėlyje pateiktas žodžio *nebeatsinešdavau* analizės pavyzdys.

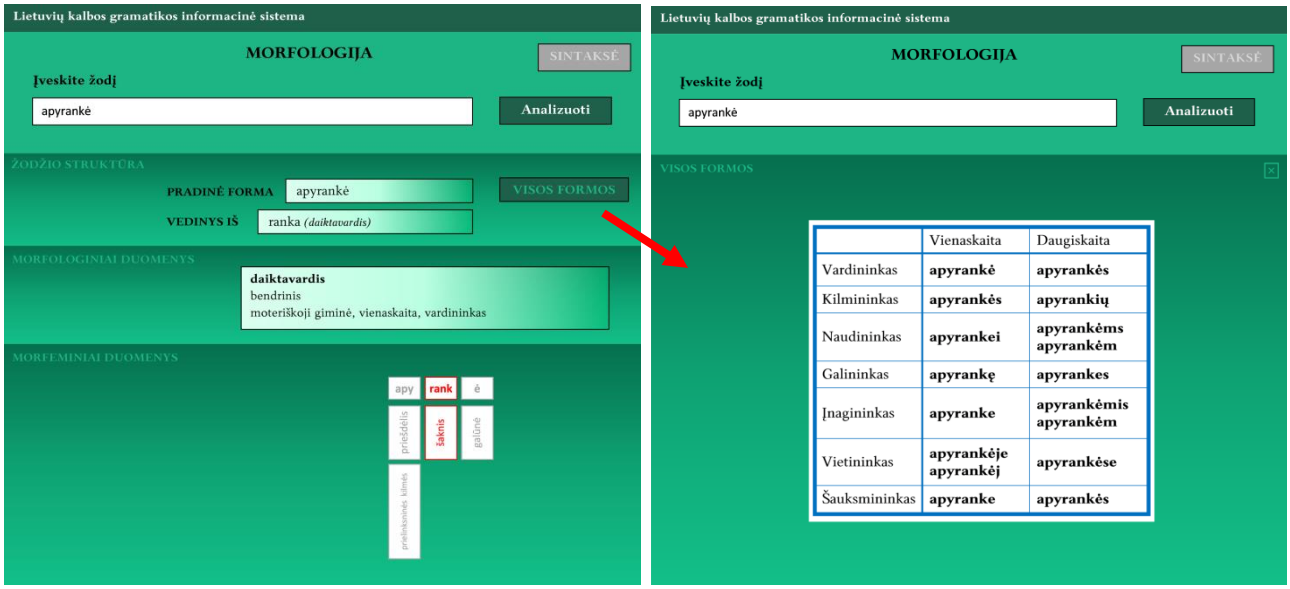
23 pav. Žodžio *nebeatsinešdavau* analizės pavyzdys.

Kokia spalva bus naudojama kiekvienam morfemos tipui žymėti, parodyta 1-oje lentelėje.

MORFEMA	SPALVA
Priešdėlis užrašomas mėlynai.	
Šaknies spalva yra raudona.	
Priesagai naudojama žalia spalva.	
Galūnei skirta juoda spalva.	
Sangrąžos dalelytė *si* žymima ruda spalva.	
Sudurtinių žodžių **jungiamieji balsiai** žymimi skirtingai nei pati šaknis – violetine spalva	
Cirkumfiksas traktuojamas kaip atskiras morfemos tipas, todėl jam numatytas atskiras žymėjimas: ir priešdėlis, ir galūnė vaizduojami tos pačios spalvos – pilkos. Tačiau morfemų pavadinimai lieka įprasti: priešdėlis ir galūnė. Cirkumfikso tipą (kad priešdėlis kartu keičia ir galūnę, pvz., *apy**rankė***) šiuo atveju rodo tik spalva. Reikia pasakyti kad cirkumfikso reiškiniai stebimi ir kitose kalbose, pvz., vokiečių dalyvio forma sudaroma būtent cirkumfikso pagalba: ***fragen*** – *ge**fragt***.	

1 lentelė. Morfemų ir spalvų atitikimas duomenų bazėje.

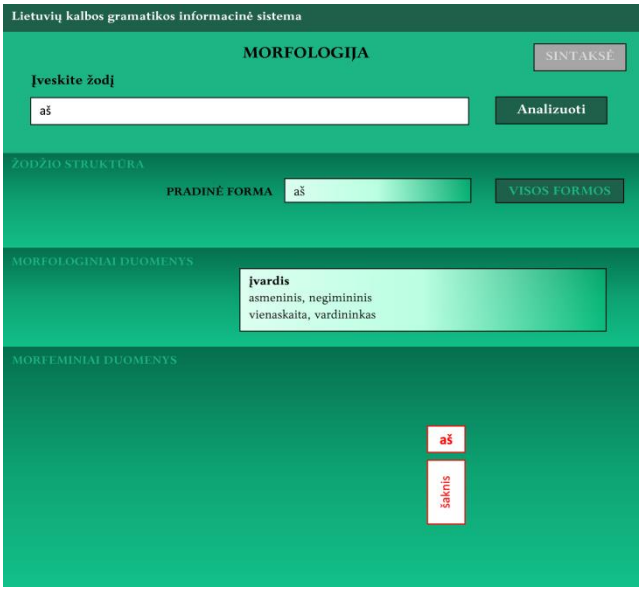
Rusų kalbos morfologinė analizė pateikia visas nagrinėjamo žodžio formas (7 pav.). Kuriant lietuvių kalbos morfologinę duomenų bazę nuspręsta nepateikti ekrane visų formų, nes tik retais atvejais jos visos gali būti reikalingos, be to tai užima daug vietos. Šis klausimas sprendžiamas kiek kitu būdu: šalia pradinės formos talpinamas mygtukas „VISOS FORMOS“, kuriuo atidaromas langas turintis visų nagrinėjamo žodžio formų sąrašą. Žodžio *apyrankė* visų formų lango pavyzdys pateiktas 24 paveikslėlyje.



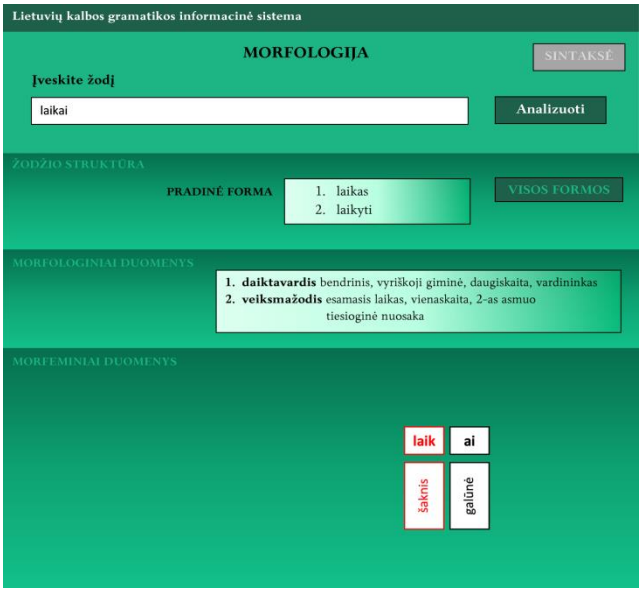
24 pav. Žodžio *apyrankė* analizės pavyzdys su mygtuko „VISOS FORMOS“ paspaudimu atvertu langu.

Šiuo metu yra paruoštas *Lietuvių kalbos gramatikos informacinės sistemos* bandomasis pavyzdys iš dvylikos žodžių (16 interneto nuoroda). Buvo stengtasi apimti kuo skirtingesnius atvejus, todėl parinkti šie žodžiai:

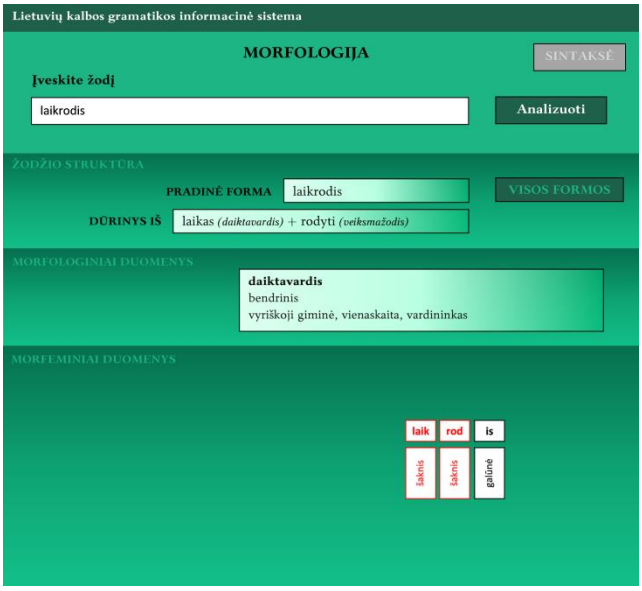
žodis, susidedantis vien iš šaknies, t.y. neturintis galūnės *aš* (25 pav.); daugiareikšmis žodis *laikai* (26 pav.); sudurtinis žodis *laikrodis* (27 pav.); žodis su nutrumpėjusia galūne *laikmenoj* (28 pav.). Esant nutrumpėjusiai galūnei ne tik nurodomas pats nutrumpėjimo faktas, bet kartu pateikiama ir pilna galūnė, t.y. kaip ji turėtų atrodyti, jeigu nebūtų nutrumpėjusi.



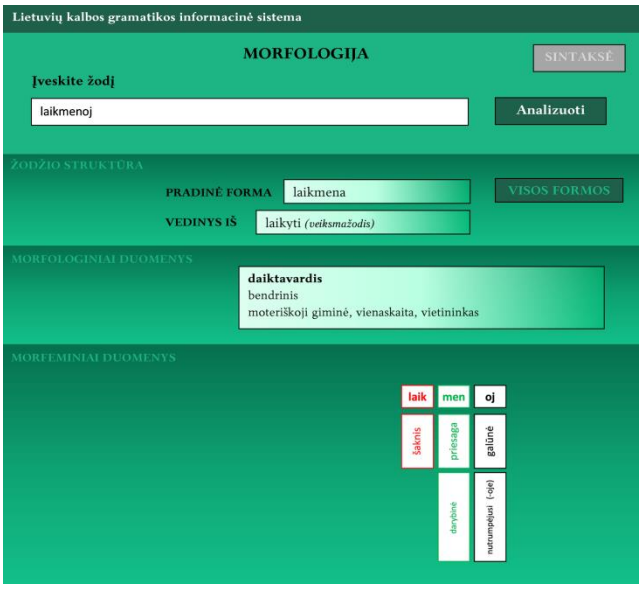
25 pav. Žodžio *aš* analizės pavyzdys.



26 pav. Žodžio *laikai* analizės pavyzdys.

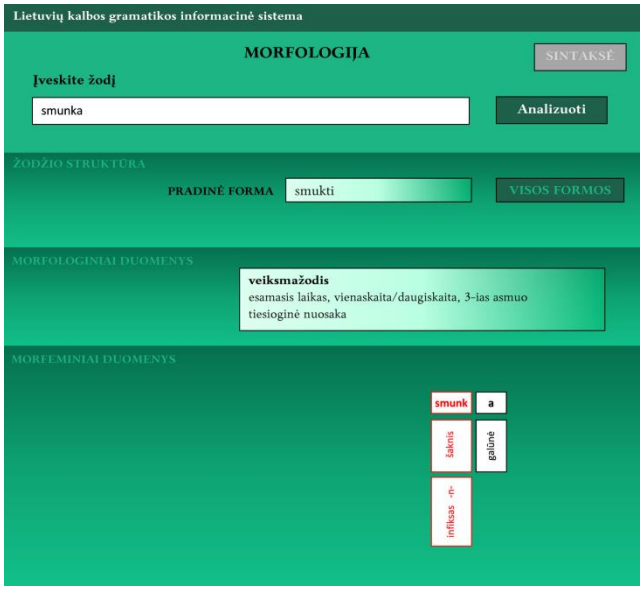


27 pav. Žodžio *laikrodis* analizės pavyzdys.

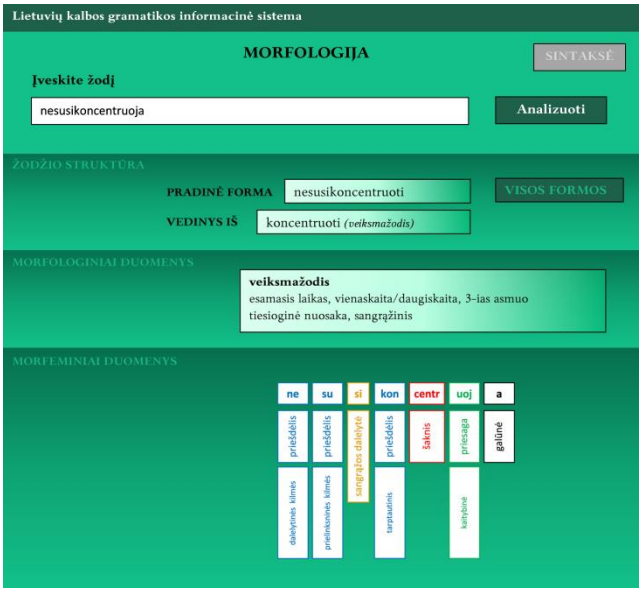


28 pav. Žodžio *laikmenoj* analizės pavyzdys.

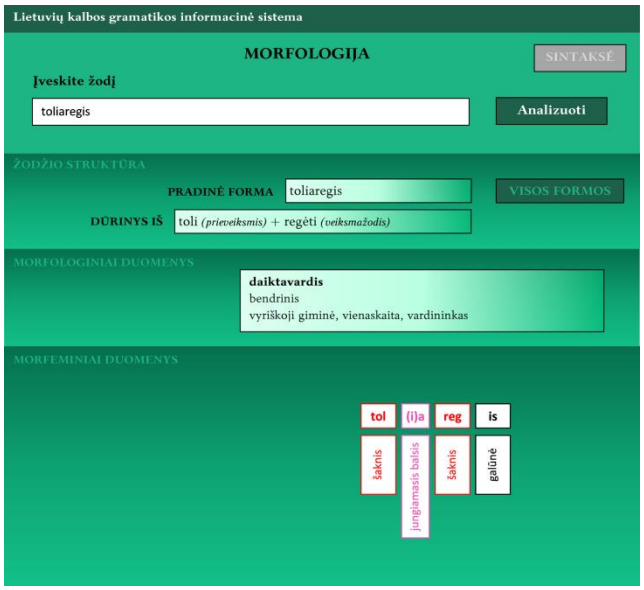
Į bandomojo pavyzdžio žodžių sąrašą įtrauktas taip pat žodis su infiksu šaknyje *smunka* (29 pav.); žodis su tarptautiniu priešdėliu *nesusikoncentruoja* (30 pav.), kuriame gerai matyti, kad dalelytė *si* gali būti įsiterpusi tarp dviejų priešdėlių – lietuviško ir tarptautinio; sudurtinis žodis su jungiamuoju balsiu *toliaregis* (31 pav.); žodis *nesu*, kuriame yra įvykusi kontrakcija (balsių e susiliejimas iš „ne+esu“). Išnykęs susiliejimo metu balsis parašomas skliausteliuose ir mažesniu šriftu (32 pav.). Žodžio su įvardžiuotine galūne pavyzdys yra *jaunesniesiems* (33 pav.); žodis su dalelyte *si* po galūnės – *mokosi* (34 pav.).



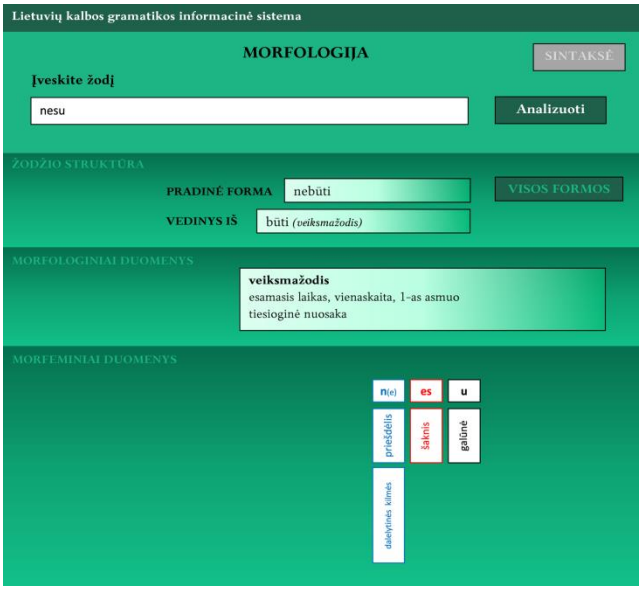
29 pav. Žodžio *smunka* analizės pavyzdys.



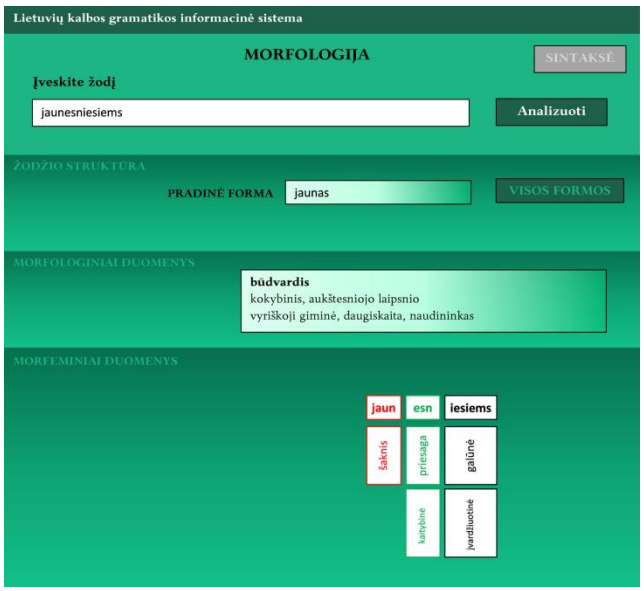
30 pav. Žodžio *nesusikoncentruoja* analizės pavyzdys.



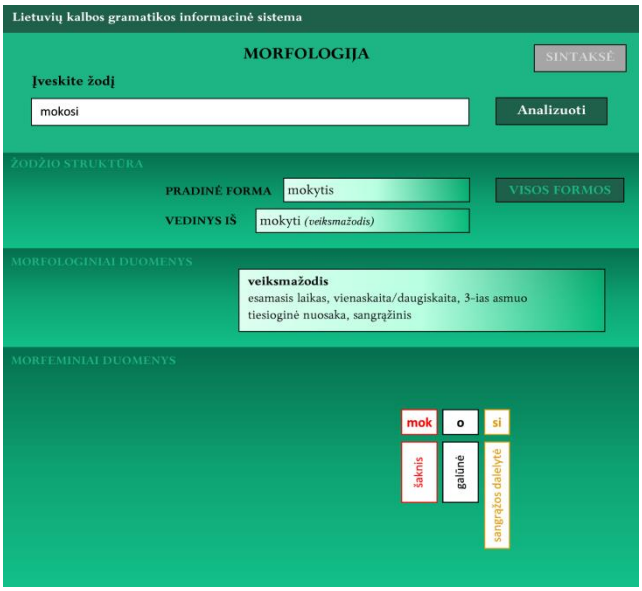
31 pav. Žodžio *toliaregis* analizės pavyzdys.



32 pav. Žodžio *nesu* analizės pavyzdys.



33 pav. Žodžio *jaunesniesiems* analizės pavyzdys.



34 pav. Žodžio *mokosi* analizės pavyzdys.

5. IŠVADOS

- Šiuo metu lietuvių kalbos morfemikos tyrimų srityje padėtis nėra labai gera. Viešai prieinami šaltiniai, pvz., Vytauto Didžiojo universitete sudarytas morfemikos žodynas bei jo pagrindu sukurta morfemikos duomenų bazė nepateikia išsamios informacijos apie morfemos tipą, todėl gali būti naudingi tik kalbininkams, turintiems specialiąsias žinias žodžių skaidymo į morfemas srityje. Nespecialistams žodis padalintas į raidžių grupes atskiriant jas brūkšneliais dažniausiai naudingos morfeminės informacijos nesuteikia, o kartais teigia netgi klaidinančią informaciją, kai vienodai pavaizduojami skirtingą morfeminę struktūrą turintys žodžiai. Morfemikos duomenų bazėje, sukurtoje Matematikos ir informatikos institute, sukaupti išsamesni duomenys apie morfemos tipą – kiekvienam jų skirtas vis kitoks šriftas. Tačiau šios duomenų bazės trūkumas yra tai, kad ji viešai neprieinama.

- Kuriant *Lietuvių kalbos gramatikos informacinę sistemą* stengiamasi užpildyti abi šias spragas – pateikti viešam vartojimui išsamią informaciją apie lietuvių kalbos morfologiją ir morfemiką.

- Gramatikos informacinėje sistemoje kaupiama ir vartotojui populiariai pateikiama dviejų tipų informacija apie žodį – morfologinė ir morfeminė. Morfologinėje dalyje nurodoma žodžio pradinė forma, kalbos dalis bei jos morfologinių kategorijų gramatiniai požymiai. Morfeminėje dalyje pateikiama išsami informacija apie žodį sudarančias morfemas. Kiekvienam morfemos tipui pavaizduoti naudojama kitokia spalva. Taip pat nurodoma informacija apie pačios morfemos požymius, pvz., priesaga: darybinė, kaitybinė ir pan.

- Kuriama lietuvių kalbos gramatikos informacinė sistema pagerins lietuvių kalbos morfologijos ir morfemikos tyrimų bei jų pateikimo rezultatus plačiajai visuomenei.

- Siūlomas iš principo naujas informacijos pateikimo lygis.

Šaltiniai

1 interneto nuoroda: Lietuvių kalbos morfemikos duomenų bazė
http://tekstynas.vdu.lt/page.xhtml?id=morfema-db [žiūrėta 2016-01-22]
2 interneto nuoroda: Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema
http://www.semantika.lt/SyntaticAndSemanticAnalysis/Analysis [žiūrėta 2016-01-22]
3 interneto nuoroda: К Р. Галиуллин. *Словообразовательно-морфемный словарь русского языка.*
http://old.kpfu.ru/infres/slovar1/begall.htm [žiūrėta 2016-01-22]
4 interneto nuoroda: NlpDotNet
http://nlpdotnet.com/services/Morphparser.aspx [žiūrėta 2016-01-22]
5 interneto nuoroda: http://donelaitis.vdu.lt/lkk/pdf/AbcI.pdf [žiūrėta 2016-01-22]

6 interneto nuoroda: http://donelaitis.vdu.lt/lkk/pdf/AbcII.pdf [žiūrėta 2016-01-22]

7 interneto nuoroda: http://donelaitis.vdu.lt/lkk/pdf/DazIII.pdf [žiūrėta 2016-01-22]

8 interneto nuoroda: http://tekstynas.vdu.lt/tekstynas/menu?page=advanced [žiūrėta 2016-01-22]

9 interneto nuoroda: https://open.xerox.com/Services/fst-nlp-tools/Consume/Morphological%20Analysis-176 [žiūrėta 2016-01-22]

10 interneto nuoroda: http://www.filosoft.ee/html_morf_et/ [žiūrėta 2016-01-22]

11 interneto nuoroda: Морфологический разбор слова онлайн http://goldlit.ru/component/slog?words=%D0%BF%D0%BE%D0%B4%D0%B3%D0%BE%D1%82%D0%BE%D0%B2%D0%BB%D0%B5%D0%BD%D0%B0 [žiūrėta 2016-01-22]

12 interneto nuoroda: http://donelaitis.vdu.lt/NLP/nlp.php [žiūrėta 2016-01-22]

13 interneto nuoroda: http://tekstynas.vdu.lt/page.xhtml;?id=morphological-annotator [žiūrėta 2016-01-22]

14 interneto nuoroda: http://www.morfologija.lt/[žiūrėta 2016-01-22]

15 interneto nuoroda: http://www.canoo.net/ [žiūrėta 2016-01-22]

16 interneto nuoroda: http://ligis.lki.lt/index.html [žiūrėta 2016-01-22]

Literatūra

Ambrazas Vytautas (red.) 1997, *Dabartinės lietuvių kalbos gramatika*, Vilnius: Mokslo ir enciklopedijų leidykla.

Isabelle Pierre, Laurent Boubeau 1985, TAUM-AVIATION: Its Technical Features and Some Experimental Results. *Computational lingusitics*, Vol. 11, Nr. 1, 18–27.

Jakaitienė Evalda 1980, *Lietuvių kalbos leksikologija*, Vilnius: Mokslas.

Keinys Stasys (red.) 1993, *Dabartinės lietuvių kalbos žodynas*, Vilnius: Mokslo ir enciklopedijų leidykla.

Metuzale-Kangere Baiba 1985, *A Derivational Dictionary of Latvian/Latviesu Valodas Atvasinajumu Vardnica*, Hamburg: John Benjamins Pub Co.

Murmulaitytė Daiva 2012, Lietuvių kalbos morfemikos ir žodžių darybos tyrimų perspektyvos, *Žmogus ir žodis*, Nr.1 (14), 96–102.

Naktinienė Gertrūda (red.) 2008, *Lietuvių kalbos žodynas* (t. I–XX, 1941–2002) elektroninis variantas. www.lkz.lt

Rimkutė Erika, Asta Kazlauskienė, Gailius Raškinis 2011a, *Abėcėlinis lietuvių kalbos morfemikos žodynas*, I dalis, VDU: Kaunas.

Rimkutė Erika, Asta Kazlauskienė, Gailius Raškinis 2011b, *Abėcėlinis lietuvių kalbos morfemikos žodynas*, II dalis, VDU: Kaunas.

Rimkutė Erika, Asta Kazlauskienė, Gailius Raškinis 2011c, *Dažninis lietuvių kalbos morfemikos žodynas*, III dalis, VDU: Kaunas.

Schwanke Martina 1991, *Maschinelle Übersetzung: Ein Überblick über Theorie und Praxis*, Berlin: Springer-Verlag.

Sedlaček Radek 2004, The Core of the Czech Derivational Dictionary, *LREC 2004*, 1279–1282, http://www.lrec-conf.org/proceedings/lrec2004/pdf/696.pdf

Šveikauskienė Daiva 2015, Morphemic structure of the Lithuanian prefixes, *Language: Meaning and form – Language System and Language Use*, Riga: Latvijas universitate, 189–197.

Ulvydas Kazys (red.) 1965, *Lietuvių kalbos gramatika – fonetika ir morfologija*, t. I, Vilnius: Mintis.

Ulvydas Kazys (red.) 1971, *Lietuvių kalbos gramatika – morfologija*, t. II, Vilnius: Mintis.

Ulvydas Kazys (red.) 1976, *Lietuvių kalbos gramatika – sintaksė*, t. III, Vilnius: Mintis.

Zinkevičius Vytautas 2000, Lemuoklis – morfologinei analizei, *Darbai ir dienos* 24, 245–273.

Įteikta 2016-02-14
Priimta 2016-03-04