LIETUVIŲ KALBOS SKAITMENINĖ GRAMATIKA

Daiva Šveikauskienė

Lietuvių kalbos institutas
P. Vileišio g. 5,
LT-10308 Vilnius, Lietuva
El. p. daiva.fmf@gmail.com

Vytautas Šveikauskas

Lietuvių kalbos institutas
P. Vileišio g. 5,
LT-10308 Vilnius, Lietuva
El. p. vytautas.sveikauskas@lki.lt

Įvadas

Straipsnio antraštėje pavartotas terminas *skaitmeninė gramatika* yra vienas iš tų naujų terminų, kurie yra paplitę ir, galima sakyti, „madingi“, bet jų reikšmė įvairiuose tekstuose yra labai skirtinga. Todėl straipsnyje daug vietos skiriama pačiai sąvokai *skaitmeninė gramatika* aiškinti. Nurodomi keli jos pavartojimo atvejai, pateikiami iliustraciniai pavyzdžiai apie mėginimus sukurti skaitmenines gramatikas kitoms kalboms, aptariant jų teigiamus ir neigiamus bruožus bei panaudojimą. Taip pat analizuojamos problemos, su kuriomis susiduria skaitmeninių gramatikų kūrėjai. Pabaigoje aprašomi pirmieji bandymai sukurti lietuvių kalbos skaitmeninę gramatiką.

1. LIETUVIŲ KALBOS GRAMATIKOS KOMPIUTERIZAVIMAS

Skaitmeninė gramatika lietuvių kalbai kol kas dar nėra sukurta. Galima paminėti tiktai kelis darbus, atliktus kompiuterizuojant lietuvių kalbos gramatiką. Vytauto didžiojo universitete (VDU) sudarytas ir elektronine forma internete laisvai prieinamas morfemikos žodynas (Rimkutė et al. 2011). Jis paruoštas remiantis tekstynu, apdorota apie 310 tūkst. žodžių. Šio žodyno pagrindu parengta morfemikos duomenų bazė (1 interneto nuoroda). Pačių autorių nurodytas trūkumas: negalima paieška pagal morfemos tipą. Reikia pasakyti, kad ir morfemikos žodyne informacijos apie morfemos tipą nėra. Kitas VDU darbas – Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema (2 interneto nuoroda). Joje pateikiama išsami kiekvieno žodžio morfologinė informacija bei nurodomi žodžių junginiai sakinyje, tačiau morfeminių duomenų apie žodį nėra.

Lietuvių kalbos institute kuriama Lietuvių kalbos gramatikos informacinė sistema, apimanti tiek morfeminę, tiek morfologinę žodžio informaciją (Šveikauskienė 2016). Šio darbo pagrindinis minusas yra kol kas dar labai nedidelė apdorojamų žodžių apimtis.

2. TERMINO *SKAITMENINĖ GRAMATIKA* VARTOJIMAS

Šiuo metu pasaulyje terminu *skaitmeninė gramatika* kartais vadinami gana skirtingi dalykai. Kompiuterinėje lingvistikoje tai yra formalus gramatikos taisyklių aprašas, ir jis turbūt geriausiai atspindi esmę. Kiti pavartojimo atvejai būtų – elektronine forma išleisti gramatikos vadovėliai.

Reikia pasakyti, kad pradėtas vartoti šis terminas buvo 40 metų prieš atsirandant kompiuteriams, t. y. daugiau nei prieš šimtmetį. 1903 m. Vokietijoje pasirodė Vilhelmo Rygerio (Wilhelm Rieger) skaitmeninė gramatika[1], kurios pavadinime buvo nurodyta ir jos paskirtis: mechaniniam vertimui iš vienos kalbos į visas kitas. Taigi, kaip galima spręsti iš pavadinimo, ir pirmosios skaitmeninės gramatikos paskirtis buvo vertimas, nors ir nekompiuterinis. Pirmoji

[1] *Zifferngrammatik, welche mit Hilfe der Wörterbücher ein mechanisches Übersetzen aus einer Sprache in alle anderen ermöglicht*

skaitmeninė gramatika – tai popieriuje spausdinta knyga, kur skaičiais buvo užkoduotos ne tik morfologinės kategorijos (linksnis, giminė, laikas ir pan.), bet ir sintaksinės, pvz., veiksmažodžių tranzityvumas (Rieger 1903, 107).

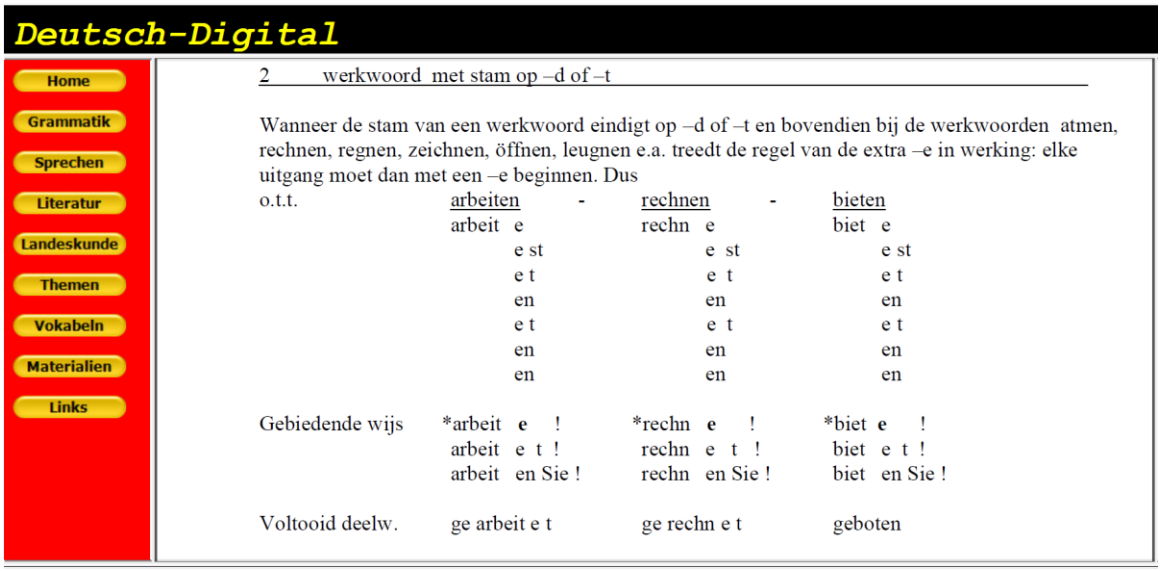
2.1. Plačiajai visuomenei skirti elektroniniai gramatikos leidiniai

Šiais laikais skaitmeninės gramatikos pavadinimu kartais leidžiami elektroniniai gramatikos vadovėliai: tiek mokantis užsienio kalbos (pvz., vokiečių) tiek gimtosios (pvz., švedų).

Leidykla CHRISTOS KARABATOS (3 interneto nuoroda) išleido kompaktinę plokštelę graikams, besimokantiems vokiečių kalbos „Meine Grammatik – Digital“ (4 interneto nuoroda) (1 pav.).

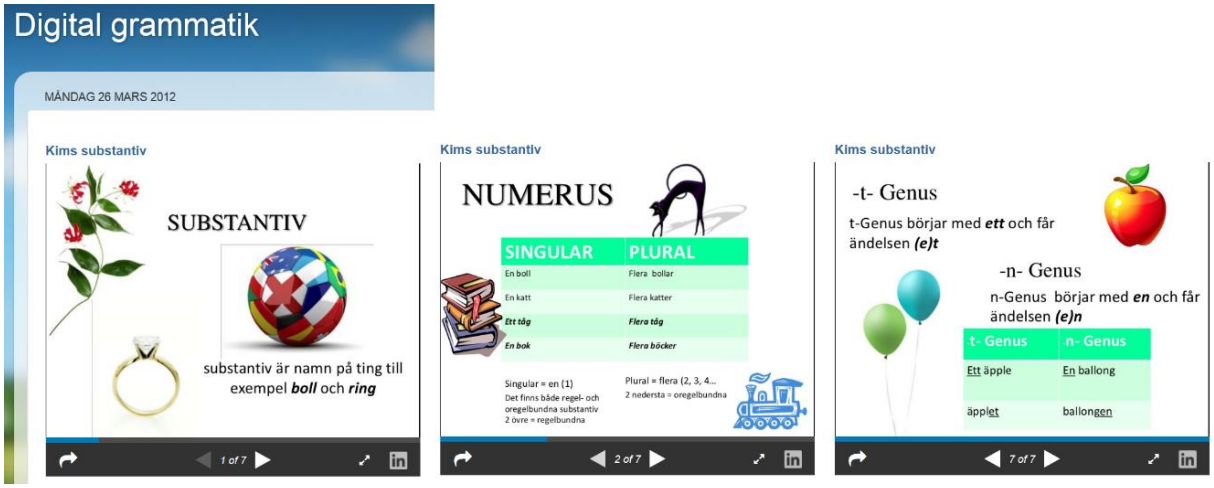
Buvo dar vienas skaitmeninės gramatikos, skirtos užsieniečiams, besimokantiems vokiečių kalbos, pavyzdys: internetinis puslapis 'Deutsch-Digital' skirtas olandams (5 interneto nuoroda). Veiksmažodžių asmenavimo pavyzdys pateiktas 2 pav.

1 pav. Vokiečių kalbos vadovėlis graikams (4 interneto nuoroda)



2 pav. Olandams skirtos vokiečių kalbos skaitmeninės gramatikos pavyzdys (5 interneto nuoroda)

Švedijoje *Skaitmeninės gramatikos* pavadinimu paruoštos skaidrės, padedančios mokytis gimtosios kalbos (6 Interneto nuoroda). Jose labai populiariai išdėstomos gramatikos taisyklės, paveikslėliais iliustruoti gramatinių kategorijų (skaičiaus, linksnio, giminės ir kt.) aiškinimai. 3 pav. pateikiama keletas tokių skaidrių pavyzdžių.



3 pav. Švedų kalbos skaitmeninės gramatikos skaidrių pavyzdžiai (6 interneto nuoroda)

2.2. Formalus gramatikos taisyklių aprašas

Skaitmeninės gramatikos, skirtos kalbos kompiuterizavimui, buvo pradėtos kurti, kai paaiškėjo, kokie sunkumai iškyla apdorojant kalbas statistiniais metodais: dėl didelės duomenų apimties darosi sunku numatyti proceso baigtį ir beveik neįmanoma jo kontroliuoti. Statistiniais metodais naudojantis greitai sukuriamos galingos sistemos, sugebančios apdoroti nepaprastai dideles žodžių apimtis, tačiau tik su sąlyga, kad bus toleruojamas tam tikras kiekis klaidų. Skaitmeninės gramatikos – tai kruopščiai sudaromi ir patikrinti kalbos ištekliai, turintys labai aukšto tikslumo informaciją (7 interneto nuoroda).

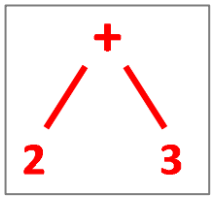
2.2.1. Gramatinė struktūra

Skaitmeninėms gramatikoms rašyti buvo sukurta speciali programinė įranga *Gramatinė struktūra* (angl. Grammatical Framework). Šią programinę įrangą paruošė Geteborgo universiteto profesorius Aarne Ranta ir pirmą kartą ji buvo įdiegta 1998 m. Grenoblyje, Xerox tyrimų centre (7 interneto nuoroda). Gramatinės struktūros paskirtis – formalizuoti pasaulio kalbų gramatikas, kad jas galima būtų panaudoti apdorojant kalbas kompiuteriais. Tai yra atviro kodo programinė įranga, ir visi ja gali naudotis nemokamai (Ranta 2015). Šiuo metu naudojantis Gramatine struktūra kuriamos daugiau nei 30-ties kalbų skaitmeninės gramatikos, tačiau išsamiausi duomenys surinkti apie anglų kalbą. Reikia atkreipti dėmesį, kad ir pats aprašas remiasi anglų kalbos savybėmis, nors stengiamasi atsižvelgti ir į kitų kalbų specifinius bruožus, pvz., didelį kaitomumą. Plačiau šis klausimas bus aprašytas 2.2.2. skyriuje.

Viena svarbiausių Gramatinės struktūros savybių yra tai, kad sintaksė čia suskaidyta į konkrečią ir abstrakčią. Abstrakčioje sintaksėje sukaupti nuo kalbos nepriklausomi duomenys, ji remiasi semantika, o konkreti sintaksė aprašo kiekvienai atskirai kalbai būdingą sintaksinę ir leksinę abstrakčios sintaksės realizaciją (Grūzītis, Dannélls 2017, 42).

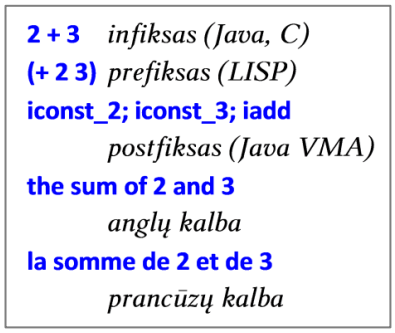
Abstrakti sintaksė yra rinkinys priklausomybių gramatikos medžių (Ranta 2009, 5). Sakinys pateikiamas apibendrintu semantiniu formatu – sąvokų struktūros medžiu. Šis atvaizdavimas

bendras visoms kalboms ir nusako sakinio prasmę. Konkreti sintaksė pateikia abstrakčios sintaksės reikšmę kiekvienai atskirai kalbai būdingu pavidalu. Sistemos universalumą rodo tai, kad kalbos gali būti tiek tautų, tiek formalios (Hallgren et al. 2015, 42), pvz., programavimo. Abstrakčiai ir konkrečiai sintaksei pailiustruoti pateikiamas matematinės operacijos pavyzdys: *Sudėti du ir tris*. Priklausomybių gramatikos medžio viršūnėje talpinamas tarinys. Abstrakčioje sintaksėje tai bus tiesiog ženklas + (4 pav.). Tarinį išplečiantys žodžiai taip pat vaizduojami simboliais – skaičiais. Taigi, šis abstrakčios sintaksės medis apima tris sąvokas – sudėties operaciją ir du skaičius: du ir trys.



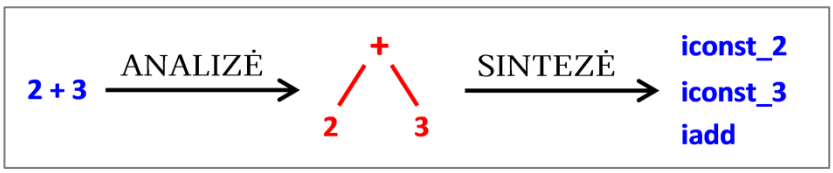
4 pav. Abstrakti sintaksė: sąvokų medis sakiniui *Sudėti du ir tris* (parengta pagal Ranta 2011, 29)

5 paveikslėlyje pateikiamas konkrečios sintaksės pavyzdys: keliomis skirtingomis kalbomis, todėl ir skirtingu pavidalu, parašyta tai, kas yra užkoduota abstrakčios sintaksės medyje (4 pav.). 5 paveikslėlyje aiškiai matyti, kad kalbos gali būti labai įvairios, t. y. tiek programavimo, tiek tautų kalbos. Java ir C programavimo kalbose sudėties aritmetinė operacija užrašoma pliuso ženklu ir nurodoma kaip infiksas, t. y. tarp skaičių arba kintamųjų (kurie gali būti užrašyti raidžių ir/ar skaičių rinkiniu). Programavimo kalboje LISP aritmetinė operacija parašoma prefiksu, t. y. prieš skaičius ar kintamuosius, su kuriais ji turi būti atlikta. Java VMA kalboje operacija nukeliama į postfiksą, t. y. ji užrašoma po skaičių ar kintamųjų. Pats užrašas šioje programavimo kalboje pateikiamas labiau įprasta žmogui forma, t. y. žodžiais, tiksliau, jų sutrumpinimais: *iadd* reiškia *integer add*, *iconst_2* – *integer constant 2* ir t. t. Objektai, su kuriais turi būti atliekama operacija, ir pati operacija atskiriami kabliataškiais. Abstraktaus medžio reikšmė (4 pav.) gali būti perduodama ir tautų kalbomis: anglų, prancūzų ir kitomis. Visi šie užrašai labai skiriasi, bet jie turi tą pačią prasmę – apima tris sąvokas: sudėties operaciją ir du skaičius. Visos šios sąvokos atsispindi kiekvienoje iš konkrečių kalbų. Ir joms visoms vienodai gerai tinka skaitmeninės gramatikos aprašas.



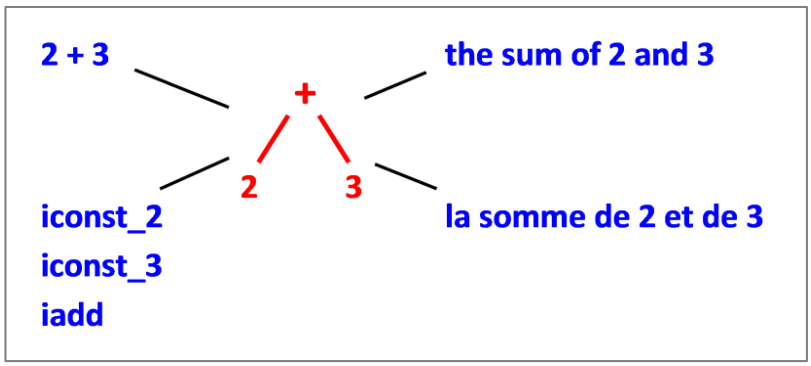
5 pav. Konkreti sintaksė: sąvokų medžio sakiniui *Sudėti du ir tris* realizacija įvairiomis kalbomis (parengta pagal Ranta 2011, 29)

Ryšiai tarp sintaksės konkrečios ir abstrakčios dalies yra dvikrypčiai. Iš bet kurios konkrečios kalbos galima gauti bendrą semantinę sakinio struktūrą, t. y. atlikti sakinio analizę; ir iš abstraktaus pavidalo (medžio) galima sugeneruoti bet kurios kalbos sakinį, turintį tą pačią prasmę, kaip ir abstrakčios sintaksės medis (6 pav.).



6 pav. Dvikrypčiai ryšiai tarp konkrečios ir abstrakčios sintaksės

Pagrindinis Gramatinės struktūros bruožas yra daugiakalbiškumas. Tai reiškia, kad vieną abstraktų sakinio aprašą atitinka daug skirtingų konkrečių to paties sakinio pavidalų (7 pav.).



7 pav. Daugiakalbiškumas: vienas abstraktus atvaizdavimas ir daug konkrečių atitikmenų

2.2.2. Gramateka – skaitmeninių gramatikų biblioteka

2018 metais naudojant Gramatinę struktūrą buvo sukurtos 34-ių kalbų skaitmeninės gramatikos, tarp jų ir estų bei latvių, tačiau lietuvių kalbos šiame sąraše dar nėra (8 interneto nuoroda). Jos visos kaupiamos gramatikos išteklių bibliotekoje – **Gramatekoje** (angl. Resource Grammar Library – RGL). 2001 metais ją tesudarė tik trys kalbos: anglų, švedų ir rusų. 2011 metais ji jau apėmė 20-ties kalbų gramatikas (Ranta 2011, 48).

Aarne Ranta (2013) populiariai aprašo pačią idėją. Visą Gramateką galima įsivaizduoti kaip sudarytą iš dviejų didelių sričių: žodžių ir sintaksės. Kiekviena šių sričių turi po du skyrius: bendrąjį ir specifinį. Remiamasi tuo, kad įvairių kalbų gramatikose didelė dalis yra bendra, nepaisant pastebimų skirtumų. Pavyzdžiui, visos šiuo metu Gramatekoje esančios kalbos turi daiktavardžio kategoriją, bet apibrėžimas, kas tiksliai yra daiktavardis, įvairiose kalbose skiriasi: anglų kalbos daiktavardis turi keturias formas (vienaskaitos ir daugiskaitos vardininką ir kilmininką); prancūzų kalboje yra tik dvi formos (vienaskaita ir daugiskaita), bet čia daiktavardžiai turi giminę, ko nėra anglų kalboje; vokiečių kalbos daiktavardžių yra 8 formos ir 3 giminės; kinų kalba teturi tik vieną vienintelę daiktavardžio formą, o suomių kalboje jų net dvidešimt šešios ir t. t. Plačiau šis klausimas aptariamas 3.1.2. skyriuje.

2.2.2.1. Bendrasis žodžių srities skyrius

Bendrajame žodžių srities skyriuje aprašomi visi žodžiai, nurodant jų charakteristikas. Ši Gramatekos dalis yra vienoda visoms kalboms. Visos joje esančios kalbos turi bendrą žodžių klasifikaciją, skiriasi tik jų morfologija. Pradžioje žodžiai skaidomi į savarankiškus (angl. content words) ir nesavarankiškus (angl. structural words). 8 pav. pateikta lentelė aprašanti savarankiškus žodžius. Pirmame stulpelyje nurodomas sutrumpintas kategorijos pavadinimas *GF* (Grammatical

Framework – Gramatinė struktūra) *name*, kuris naudojamas programinės įrangos kode. Antrame stulpelyje parašyta ta pati kategorija, kaip ji atrodo žmonėms skirtame tekste. Trečiame stulpelyje pateikiami žodžių pavyzdžiai. Ketvirtame stulpelyje surašyti kaitybiniai žodžio požymiai, o penktas stulpelis apima nekaitomus jo požymius. Iš lentelės matyti, pvz., kad daiktavardžiai kaitomi skaičiais ir linksniais, o giminė yra pastovus požymis ir pan.

GF name	text name	example	inflectional features	inherent features
N	noun	*house*	number, case	gender, classifier
PN	proper name	*Paris*	case	gender
A	adjective	*blue*	gender, number, case, degree	position
V	verb	*sleep*	number, person, tense, aspect, mood	subject case
Adv	adverb	*here*	(none)	adverb type (place, time, manner)
AdA	adadjective	*very*	(none)	(none)

8 pav. Savarankiškų žodžių charakteristikos (Ranta 2013)

Šioje lentelėje yra apibendrinti visų Gramatekoje esančių kalbų duomenys ir tokia lentelė užpildoma kiekvienai kalbai. Galima teigti, kad, jei kalba turi giminės kategoriją, tai daiktavardžiai yra nekaitomi giminėmis, t. y. giminė yra pastovus daiktavardžio požymis, o būdvardžiai kaitomi giminėmis. Ir tai lemia ne atskiros kalbos savybės, o sintaksė: būdvardžiai pažymi daiktavardžius ir turi pažymėti juos visus, nepriklausomai nuo giminės, todėl turi įgyti visas giminės formas (Ranta 2009, 8).

Valentingumui įvertinti Gramatekoje naudojama pozicijų sąvoka: nurodomi dvipozicinai, tripoziciniai ir pan. daiktavardžiai, būdvardžiai, veiksmažodžiai. 9 pav. ir 10 pav. pateikti fragmentai iš lentelių, aprašančių atitinkamai veiksmažodžių ir daiktavardžių bei būdvardžių valentingumą.

GF name	text name	example	inherent complement features
V2	two-place verb	*love* (*someone*)	case or preposition
V3	three-place verb	*give* (*something to someone*)	two cases or prepositions
VV	verb-complement verb	*try* (*to do something*)	infinitive form
VS	sentence-complement verb	*know* (*that something happens*)	sentence mood

9 pav. Veiksmažodžių valentingumo išraiška pozicijomis (Ranta 2013)

GF name	text name	example	inherent complement features
N2	two-place noun	*brother* (*of someone*)	case or preposition
N3	three-place noun	*distance* (*from some place to some place*)	case or preposition
A2	two-place adjective	*similar* (*to something*)	case or preposition

10 pav. Daiktavardžių ir būdvardžių valentingumo išraiška pozicijomis (Ranta 2013)

2.2.2.2. Specifinis žodžių srities skyrius

Specifiniame žodžių srities skyriuje nurodomi kiekvienai atskirai kalbai būdingų morfologinių kategorijų parametrai. 11 pav. pateikta lentelė iš anglų kalbos žodžių specifinio skyriaus, aprašanti, kokias reikšmes gali įgyti kiekviena morfologinė kategorija, kuri sutinkama anglų kalboje, pvz., linksnio gali būti tik dvi reikšmės: vardininkas ir kilmininkas ir pan.

GF name	text name	values
`Number`	number	singular, plural
`Person`	person	first, second, third
`Case`	case	nominative, genitive
`Degree`	degree	positive, comparative, superlative
`AForm`	adjective form	degrees, adverbial
`VForm`	verb form	infinitive, present, past, past participle, present participle
`VVType`	infinitive form (for a VV)	bare infinitive, *to* infinitive, *ing* form

11 pav. Anglų kalbos morfologinės kategorijos, nurodytos žodžių srities specifiniame skyriuje (Ranta 2013)

Daugumai Gramatekos kalbų kaitybinės formos užima labai didelę aprašo dalį. Kiekviena kaitybos paradigma pateikiama lentelėje. Anglų kalbos daiktavardžių linksniavimo paradigma parodyta 12 pav.

form	singular	plural
nominative	*dog*	*dogs*
genitive	*dog's*	*dogs'*

12 pav. Anglų kalbos daiktavardžių kaitybos paradigma (Ranta 2013)

Programinė įranga pagal paradigmas sudaro visas galimas pateikto žodžio formas. Tam naudojama programavimo kalbos funkcija, kuri užrašoma specialiu pavidalu. Pvz., **mkV : Str → V** reiškia funkciją *mkV* (make Verb – sudaryti veiksmažodį), t. y. iš eilutės *Str* (string) padaryti veiksmažodį *V* (Verb). Į eilutės (Str) vietą įrašius konkretų anglų kalbos žodį, pvz., *walk*, t. y. **mkV : walk → V** bus gaunama visa paradigma: *walk, walks, walked, walking*. Kita funkcija **mkCl : NP → V → Cl** reiškia *sudaryti sakinį* (mkCl – make Clause), t. y. iš daiktavardinės frazės *NP* (Noun Phrase) ir veiksmažodžio *V* (Verb) padaryti sakinį *Cl* (Clause). Ši funkcija rūpinasi ir galūnių suderinimu: *she walks* ar *they walk* (Ranta 2009, 2).

3. DAUGIAKALBIŠKUMO IR PROGRAMAVIMO PROBLEMOS

Skaitmeninės gramatikos priklauso tarpdisciplininei sričiai, todėl ir problemos čia iškyla dviejų tipų: vienos susijusios su kalbiniais, t. y. lingvistikos dalykais, kitos – su kompiuterine realizacija, t. y. programinės įrangos kūrimu.

3.1. Daugelio kalbų apibendrinimas

Bet kokia programinė įranga, apdorojanti vienu metu daugelio tautų kalbas, turi būti iš karto kuriama kaip labai universali sistema, apimanti visų planuojamų įtraukti kalbų savybes, nes

„skirtumai tarp įvairių kalbų gali būti labai dideli ir esminiai“[2] (Dąbrowska 2015, 1). Reikia pasakyti, kad gramatinės kategorijos sukuriamos atskirai konkrečiai kalbai tada, kai jų tai kalbai prireikia, ir visiškai nesirūpinama, ar jos tinka kitų kalbų gramatikoms[3] (Newmeyer 2008, 51 iš Dąbrowska 2015, 2). Ir atliekant šiuolaikinius lingvistinius tyrimus, atidžiau panagrinėjus neseniai aptiktas naujas kalbas, beveik kiekvienoje jų atsiskleidžia netikėti nauji požymiai[4] (Evans, Levinson 2009, 432).

3.1.1. Universalios gramatikos idėja

Pagrindinė (nors ir ne vienintelė) skaitmeninių gramatikų pritaikymo sritis yra automatinis vertimas. Kuriant daugiakalbes automatinio vertimo sistemas būtų labai patogu turėti gramatiką, kuri apimtų visų pasaulio tautų kalbas.

Tokios universalios gramatikos idėjų ištakas galima įžvelgti dar 13-tame amžiuje išsakytame Rogerio Beikono (Roger Bacon) teiginyje, kad visos kalbos remiasi bendra gramatika[5] (Nordquist 2018). 17-tame amžiuje Port Rojalio (Port Royal) gramatikos mokyklos atstovų požiūris į kalbą buvo pagrįstas idėja, kad žmonės civilizuotame pasaulyje turi bendrą minčių struktūrą[6]. 1830 m. Vilhelmas fon Humboltas (Wilhelm von Humboldt) rašė, kad universali gramatika yra smulkiausių nedalomų gramatinių kategorijų bei jų santykių rinkinys ir visa tai tarnauja kaip statybiniai blokai, kuriems uždedamos sintaksinės struktūros bei jų apribojimai. Taip sudaromos visų kalbų konkrečios gramatikos. Universali gramatika iškelia prielaidą, kad visos kalbos turi tą patį gramatinių kategorijų ir sintaksinių ryšių rinkinį, ir žmonės, turėdami baigtinį šių priemonių skaičių, sudaro begalinį kiekį jų pavartojimo atvejų (Barsky 2017). Pastaruoju metu pasirodė daug publikacijų, bandančių paneigti šiuos teiginius, bet apie tai plačiau bus aprašyta 3.1.2. skyriuje.

Universalios gramatikos idėją 20-tame amžiuje išpopuliarino Noamas Chomskis (Noam Chomsky). Tačiau jis pateikė kiek susiaurintą jos sąvoką, kuri apima tik baigtinį skaičių konkrečių kalbų[7] (Chomsky 1993, 13). Reikia pabrėžti, kad N. Chomskis padėjo pagrindus šiuo metu labai paplitusiam sakinių sintaksinės struktūros pavaizdavimo būdui – frazių gramatikai. Sakinį *S* (sentence) jis vaizdavo kaip susidedantį iš daiktavardinės frazės *NP* (noun phrase) ir veiksmažodinės frazės *VP* (verb phrase): S → NP INFL VP (Chomsky 1993, 25). Aiškindamas universalios gramatikos principus, kaip vieną jų jis nurodė veiksnio būtiną ar nebūtiną buvimą sakinyje. Universali gramatika atskiroms kalboms palieka galimybę pasirinkti, kuris atvejis atitinka jos poreikius. Anglų ir prancūzų kalbose veiksnys būtinas, tačiau kitose kalbose nėra tokio reikalavimo, tai, pavyzdžiui, semitų kalbos, todėl universalioje gramatikoje šis parametras žymimas skliaustuose: S → (NP) INFL VP (Chomsky 1993, 27), t. y. parametro *NP* paėmimas į skliaustus rodo, kad jis gali būti sakinyje, arba jo gali nebūti.

Sukurta universalios gramatikos teorija iš principo yra tik pasakojimas apie visus galimus kalbos garsus, leksinius konceptus, lingvistines reikšmes. Ji turi apimti visus galimus fonologinius

[2] “Languages differ from each other in profound ways” (Dąbrowska 2015)
[3] “...categories are proposed for a particular language when they appear to be needed for that language, with little thought as to their applicability to the grammar of other languages” (Newmeyer 2008)
[4] “...at this stage of linguistic inquiry almost every new language that comes under the microscope reveals unanticipated new features” (Evans, Levinson 2009)
[5] “...all languages are built upon a common grammar” (Nordquist 2018)
[6] “...the 17th century Port Royal grammarians, whose rationalist approach to language and language universals was based on the idea that humans in the “civilized world” share a common thought structure” (Barsky 2017)
[7] “...UG [Universal Grammar] does in fact permit only a finite number of core [particular] grammars” (Chomsky 1993)

ir semantinius požymius ir visas taisykles bei apribojimus, kad būtų galima juos sujungti į žodžius, o žodžius sujungti į begalinį skaičių frazių ir sakinių. Žinoma, tokia sudėtinga teorija niekada negalės būti išbaigta, bet šiuo požiūriu lingvistikos padėtis nėra blogesnė už chemijos, fizikos ar kitų mokslo sričių. Jose darbai taip pat nėra užbaigti[8] (McGilvray 2018).

Šiuo metu pasaulyje priskaičiuojama nuo 5000 iki 8000 kalbų ir tik mažiau nei 10% iš jų, t. y. apie 500 yra tinkamai dokumentuotos, t. y. turi išsamiai aprašytas gramatikas ir žodynus. Ir tai yra viskas, iš ko galima daryti apibendrinimus. Kai kurie mokslininkai teigia, kad istorijos pradžioje kalbų galėjo būti apie pusę milijono, taigi galima manyti, kad mes teturim tik apie 2% visos kalbų įvairovės (Evans, Levinson 2009, 432).

3.1.2. Morfologinių kategorijų ribos

Pastaruoju metu universali gramatika susilaukia nemažai kritikos. Pasigirsta netgi labai kategoriškų teiginių – kad universali gramatika iš viso neegzistuoja[9] (McCrum 2012). Kritikuojama plačiai paplitusi prielaida, kad visos kalbos yra panašios į anglų kalbą, tik skiriasi savo garsų sistema ir žodynu. Teigiama, kad jos skiriasi iš esmės visuose savo struktūros lygmenyse, ir esama nykstamai mažai universalijų, kurias turėtų visos kalbos (Evans, Levinson 2009, 429).

Pagrindinis ir neginčijamas faktas apie kalbas yra jų įvairovė, pvz., kalbos gali turėti mažiau nei 12 skirtingų garsų ir gali turėti jų 12×12, o gestų kalbos iš viso nenaudoja garsų. Kalbos gali neturėti morfologinės darybos, o jų semantika gali skaidyti pasaulį labai skirtingais pjūviais. Sintaksinė kalbų struktūra gali skirtis tiek, kiek 13 pav. ir tiek, kiek 14 pav.

This woman caught that huge butterfly

That$_{\text{object}}$ **this**$_{\text{subject}}$ **huge**$_{\text{object}}$ **caught woman**$_{\text{subject}}$ **butterfly**$_{\text{object}}$

13 pav. Kalbų sintaksės skirtumai (parengta pagal Evans, Levinson 2009, 431)

I cooked the wrong meat for them again.

abanyawoihwarrgahmarneganjginjeng

14 pav. Sintaksės skirtumai polisintetinėse ir analitinėse kalbose (parengta pagal Evans, Levinson 2009, 432)

Izoliacinės kalbos neturi kaitybinių asmens, skaičiaus, laiko afiksų, jos naudoja tik šaknis, o daugiskaitą ar būtąjį laiką gauna arba iš konteksto, arba iš kitų nepriklausomų žodžių. Polisintetinės kalbos sudeda visą anglišką sakinį į vieną žodį (14 pav.). Lao kalba neturi būdvardžių ir ypatybę išreiškia kaip veiksmažodžio potipį. Kai kuriose kalbose nėra skirtumo tarp veiksmažodžio ir daiktavardžio, pvz., jos turi tokius žodžius kaip „bėgti“, „būti žmogumi“, „būti dideliam“ ir pan. (Evans, Levinson 2009, 434). Kai kurios kalbos neturi priemonių išreikšti

[8] “Of course, such a complete theory may never be fully achieved, but in this respect linguistics is no worse off than physics, chemistry, or any other science. They too are incomplete.” (McGilvray 2018)
[9] Interview *Daniel Everett: ‘There is no such thing as universal grammar’* (McCrum 2012)

spalvoms, skaičiams, pvz., piraha kalba. Įdomu tai, kad šios tautos atstovai nėra sukūrę mitų, piešinių ir kolektyvinė atmintis tesiekia tik dvi paskutines kartas, tačiau iki šių dienų tauta yra išlikusi vienakalbė (Everett 2005, 621).

Esant tokiai kalbų įvairovei, logiška, kad lingvistai ima ieškoti kitų būdų, kaip parašyti įvairias gramatikas, anglų kalbos nelaikant pagrindu visiems tyrimams, t.y. nesuabsoliutinant teiginio, kad daiktavardis, veiksmažodis ir būdvardis yra tarpkalbinės kategorijos, nes visos kalbos juos turi[10] (Baker 2003, 298). Haspelmatas (Haspelmath) kritikuoja kalbininkus, bandančius savo darbuose kokią nors kalbą, pvz., čamorų, aprašyti naudojantis anglų kalbos kategorijomis, kurios laikomos universaliomis (Chung 2012, 12), ir teigia, kad tai tas pat, kas bandyti aprašyti anglų kalbą naudojant čamorų gramatikos kategorijas: *klasė I*, kurią sudaro žodžiai turintys objektus ir *klasė II*, kuriai priklauso visi likę žodžiai, t.y. neturintys objektų (Haspelmath 2012, 121). Jis pateikia vaizdų palyginimą šia tema ir sako, kad negalima klausti: ar kalba X skiria daiktavardį ir veiksmažodį; ar kalba X skiria veiksmažodį ir būdvardį; ar kalba X skiria daiktavardį ir būdvardį. Tarpkalbiniam lygmenyje negali būti klausiama: ar visos kalbos skiria daiktavardį ir veiksmažodį; ar visos kalbos skiria veiksmažodį ir būdvardį; ar visos kalbos skiria daiktavardį ir būdvardį. Tokius klausimus jis palygina su klausimu, kiek valstijų sudaro Prancūziją. Šito galima klausti apie JAV, bet ne apie Prancūziją (Haspelmath 2012, 109–114). Taigi, daiktavardis, būdvardis ir veiksmažodis yra apibendrinimai, tinkantys tik atskiroms kalboms.

Reikia rasti universalius požymius, kurie nebūtų specifiniai kurios nors atskiros kalbos bruožai. Floidas (Floyd), nurodydamas skirtumus tarp daiktavardžio ir būdvardžio kečujų kalboje, sako, kad būdvardžiai sakinyje eina prieš daiktavardžius, bet ne atvirkščiai (Floyd 2011, 53 iš Haspelmath 2012, 117). Toks požymis gali būti bendras su anglų kalba, bet jis tinka ne visoms kalboms, pvz., to negalima teigti kad ir apie lietuvių kalbą. Universalesnis teiginys atrodytų, kad būdvardžiai gali eiti daiktavardžio pažyminiu, bet daiktavardžiai negali atributiškai pažymėti būdvardžių[11] (Haspelmath 2012, 117). Tačiau pasigilinus ir šiam teiginiui galima rasti jį paneigiančių pavyzdžių: sakinyje *Kietas lauke ir skystas kambaryje gali būti tik vanduo žiemos metu*. Daiktavardžiai *lauke* ir *kambaryje* yra atitinkamai būdvardžių *kietas* ir *skystas* aplinkybiniai pažyminiai, t. y. pažymi juos atributiškai.

Metodologija nėra tiksli, jei lyginamos įvairių kalbų kategorijos, kurios atskirose kalbose yra nustatytos remiantis skirtingais kriterijais. Ta pati kategorija gali būti apibrėžiama labai nevienodai: pvz., senovės graikų Dionisijaus Trakiečio gramatikoje daiktavardis nusakomas kaip linksniais kaitoma kalbos dalis, reiškianti daiktą arba veiksmą; anglų kalboje teigiama, kad daiktavardis – tai žodis, kuris gali eiti po apibrėžiklių *the, this, that*; kinų pareigūnų kalboje (angl. mandarin Chinese) daiktavardis apibrėžiamas kaip žodis, kuris eina po klasifikatoriaus.

Kalbas reikia lyginti naudojant rinkinį specialių konceptų. Vienas tokių sprendimų buvo – lyginti kalbas remiantis semantinėmis sąvokomis: *daiktų šaknys*, kurios reiškia fizinius objektus; *veiksmų šaknys*, kurios reiškia valingą veiksmą; ir *požymių šaknys*, kurios reiškia savybę. Bet kurioje kalboje galima lengvai nustatyti šaknis (o ne žodžius!) ir taip pat lengvai galima nustatyti daiktus, veiksmus ir savybes (Haspelmath 2012, 122–123).

Kalbų palyginimas gali remtis šaknų ryšiu su pasakymais. Anglų kalbos daiktavardžio šaknies *water*, veiksmažodžio šaknies *run* ir būdvardžio šaknies *wet* pavartojimą trijuose pagrindiniuose pasakymo tipuose – paminėjimas, predikacija ir savybė – galima pailiustruoti lentele (1 lentelė). Paprastai kalbose galima pastebėti tokias tendencijas: jei daikto šaknis

[10] "... all languages have at least a few nouns, verbs and adjectives" (Baker 2003)
[11] "... nouns cannot attributively modify adjectives" (Haspelmath 2012)

paminima, ji neturi jokio funkciją nurodančio kodavimo, t. y. nominalizavimo; jei veiksmo šaknis pateikiama kaip predikatas, ji neturi specialaus funkcijos kodavimo, t. y. jungties; jei požymio šaknis naudojama kaip pažyminys, ji neturi specialaus funkciją nurodančio kodavimo, pvz., posesyvinio linksnio ir pan. Tačiau kečujų kalboje daiktų šaknys pažymėjimo atveju naudojamos taip pat kaip ir požymių šaknys, o tagalų kalboje visų trijų rūšių šaknys naudojamos vienodai visuose trijuose pavartojimo tipuose. Taigi, absoliučiai universalių kriterijų visoms kalboms palyginti kol kas nepasiūlyta.

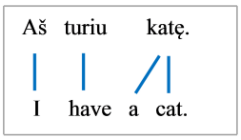
1 lentelė. Šaknų ryšiai su pasakymais (parengta pagal Haspelmath 2012, 124)

	Paminėjimas	**Predikacija**	**Pažymėjimas**
Daiktų šaknys	WATER	(that) **is** water	(colour) **of** water
Veiksmų šaknys	**the** runn-**ing**	(it) RUN (s)	runn-**ing** (water)
Požymių šaknys	the wet-**ness**	water **is** wet	WET (water)

Tačiau, jeigu tarp kalbų nebūtų nieko bendra, nebūtų įmanomas joks vertimas; reikia tik gebėti parinkti lygiagrečias formas visuose teksto sluoksniuose (Mockus 2018).

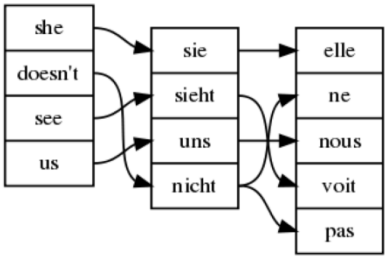
3.2. Neuroniniai tinklai ir statistiniai metodai

Kita sritis, kur skaitmeninių gramatikų kūrėjai susiduria su problemomis – tai programinė įranga. Kaip jau buvo minėta 3.1.1. skyriuje, pagrindinė skaitmeninių gramatikų pritaikymo sritis yra automatinis vertimas. Pirmi pasiūlymai vertimui naudoti kompiuterius pasirodė praeito amžiaus viduryje. Tada buvo minimi ir statistiniai metodai, bet greitai jų atsisakyta dėl nedidelio tuometinių kompiuterių pajėgumo ir nepakankamo kiekio tekstų, sukauptų elektronine forma. Tuo metu buvo plačiai pradėtas vystyti taisyklėmis pagrįstas automatinis vertimas. Tačiau praeito amžiaus pabaigoje, kai buvo surinkti tekstynai, vėl grįžtama prie tikimybių teorija paremto automatinio vertimo (Brown et al. 1990, 79). Taikant statistinius metodus galima versti nesinaudojant nei gramatikos taisyklėmis, nei žodynu (Daudaravičius 2006, 13). Žodžių atitikimas tarp kalbų vertimo metu, nustatomas iš dvikalbių išlygiagretintų tekstynų (Nießen, Ney 2000, 1081). Labai paprastas išlygiagretinto teksto pavyzdys pateiktas 15 pav.



15 pav. Išlygiagretintas lietuvių-anglų kalbų sakinys

Išlygiagretinimo metu reikia įvertinti du pagrindinius faktus: pirma, vienas žodis iš vienos kalbos gali būti verčiamas į kitą kalbą dviem ar daugiau žodžių ir antra, žodžių išsidėstymas sakinyje gali skirtis įvairiose kalbose. 16 pav. parodytas sakinys, išlygiagretintas trims kalboms (Ranta 2011, 35).

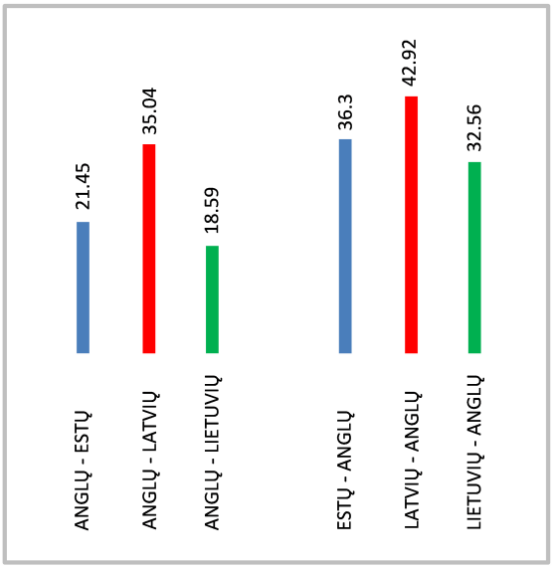


16 pav. Išlygiagretinto trims kalboms sakinio pavyzdys (Ranta 2011, 35)

Naudodami išlygiagretintus tekstynus kompiuteriai verčia tekstus iš vienos kalbos į kitą spėjimo būdu[12] (Geitgey 2016). Kaip atliekamas statistinis vertimas kompiuteryje parodyta 17 pav. Iš daugelio tekstynuose sukauptų vertimo variantų ieškoma labiausiai tikėtino. Tačiau reikia pasakyti, kad labiausiai tikėtinas variantas ne visada žmogui yra pats geriausias.

17 pav. Vertimo, naudojant išlygiagretintus tekstynus, pavyzdys (Ranta 2017, 16)

Didžiulės problemos iškyla, kai verčiami didelio kaitomumo kalbų tekstai, nes tekstynuose sunku apimti visas galimas kiekvieno žodžio formas. Tai ypač aktualu ir lietuvių kalbai. Matyt, neatsitiktinai Tildės 2017 m. duomenimis (18 pav.) lietuvių kalbos vertimai Google statistinio vertimo sistemoje buvo patys prasčiausi iš visų trijų Baltijos šalių (Skadiņš 2017, 22). Analizė pateikta pagal BLEU (bilingual evaluation understudy) įverčius, t. y. skaičiuojama, kokią dalį sakinių kompiuteris išvertė teisingai palyginti su žmogaus vertimu.

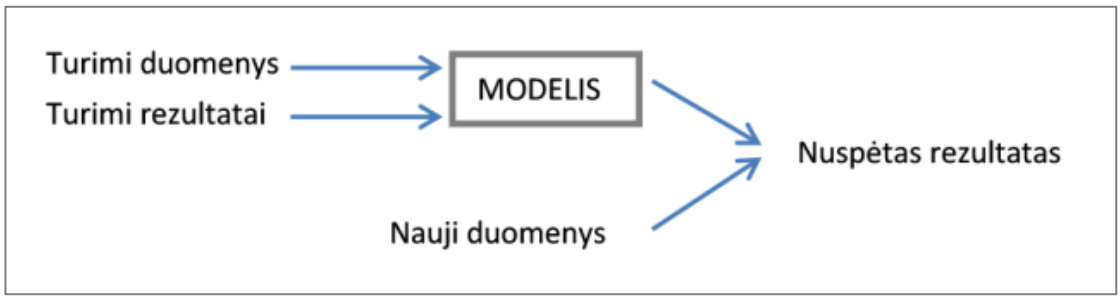


18 pav. Latvių, estų ir lietuvių statistinio Google vertimo kokybė pagal BLEU įverčius (parengta pagal Skadiņš 2017, 22)

Šiuo metu statistinį vertimą išstumia kita taip pat tikimybėmis besiremianti vertimo metodika – neuroniniai tinklai. Abiejų šių metodų pagrindas yra automatinis mokymasis (angl. machine learning), kuris nuo klasikinio programavimo skiriasi tuo, kad čia žmogus kompiuteriui pateikia duomenis ir rezultatus, o kompiuteris turi sudaryti taisykles, kurios vėliau gali būti panaudotos kitiems, naujiems duomenims apdoroti. Klasikiniame programavime kompiuteriui pateikiamos taisyklės ir duomenys, kurie turi būti apdorojami pagal tas taisykles, ir kompiuteris apskaičiuoja tikslų rezultatą (Maršalkaitė 2018). Kitais žodžiais tariant, tradiciniame programavime geresnių rezultatų sulaukiama žmogui tobulinat programinės įrangos kodą, o naudojant

[12] "... computers can use parallel corpora to guess how to convert text from one language to another" (Geitgey 2016)

automatinio mokymosi metodą programos parašomos taip, kad jos pačios pakoreguoja savo atliekamą darbą naudodamos vis daugiau gaunamų duomenų. Tačiau joms ne visada pavyksta sėkmingai save patobulinti. Automatinio mokymosi struktūrinė schema parodyta 19 pav. (Žaliauskas 2017, 5).



19 pav. Automatinio mokymosi struktūrinė schema (parengta pagal Žaliauskas 2017, 5)

Viena naujausių neuroninių tinklų metodu pagrįstų automatinio vertimo sistemų yra eTranslation. Ji skirta teisiniams tekstams versti ir naudoja Europos Sąjungos dokumentų tekstynus. Gana išsami šios sistemos veikimo analizė buvo pateikta 2019 m. vykusio Antrojo ELRC (European Language Resource Coordination – Europos kalbų išteklių koordinavimas) seminaro Lietuvoje metu. Pripažįstant pagrindinius šios vertimo sistemos privalumus, kad eTranslation „dažnai pateikia gerus ar mažai taisytinus išversto teksto gabalus; beveik taisyklinga gramatika (tinkamos žodžių formos); atsižvelgia į kontekstą“ buvo nurodyti ir kol kas dar pastebimi trūkumai: „prasmės iškraipymas (pvz., *active and healthy ageing* išversta *aktyvus ir sveikas senėjimas;* tokios realijos kaip *aktyvus senėjimas* nėra, žmogaus vertimas būtų: *vyresnio amžiaus žmonių aktyvumas ir sveikata*); kiti minusai: terminijos nenuoseklumas; iš piršto laužti (pačios sistemos susikurti) žodžiai, terminai, gramatinės formos“ ir kaip problema buvo įvardyta: „Gramatiškai sklandus vertimas gali slėpti svarbias prasmės klaidas“ (Zaikauskas 2019, 22–23). Pačios sistemos mokymasis buvo stebimas analizuojant gautų rezultatų pokyčius, kai vertimui pateikiamas tas pats sakinys *Every single day somewhere in the world indigenous peoples are being dispossessed of their ancestral lands, territories and resources*. 2018 m. spalio mėn. pateiktas vertimas buvo: *Kiekvieną dieną pasaulio čiabuvių tautose yra* ***apykažiamos*** *jų protėviai, teritorijos ir ištekliai.* Žodis *apykažiamos* yra pačios vertimo sistemos susikurtas: sistema analizuoja kontekstą ir pagal tam tikrą algoritmą (detalių veiksmų seką) parenka geriausią variantą. 2019 m. sausį verstas tas pats sakinys skamba šiek tiek geriau (bent jau nėra nesančių lietuvių kalboje žodžių): *Kiekvieną dieną visame pasaulyje čiabuvių tautos* ***nepaiso*** *savo protėvių žemių, teritorijų ir išteklių*; tačiau iki tikslaus vertimo čia dar toli. Šis pavyzdys rodo, kad neuroninių tinklų metodu pagrįstos automatinio vertimo sistemos iš tikrųjų gali pačios kai ką išmokti, tačiau tas mokymasis ne visada būna sėkmingas ir tai iliustruoja dar vienas pavyzdys. Tuo pačiu metu, 2018 m. spalio mėn., buvo verčiamas į lietuvių kalbą kitas anglų kalbos sakinys: *In point 1 of Annex I to Regulation (EC) No 1275/2008, the entry 'Dish washing machines' is deleted*. Jo vertimas buvo gautas toks: *Reglamento (EB) Nr. 1275/2008 I priedo 1 punkte įrašas „****Diškių skalbimo mašinos****“ išbraukiamas.* Antro etapo metu, 2019 m. sausį jau gautas kitas vertimo variantas, kuris yra ne ką geresnis už pirmąjį: *Reglamento (EB) Nr. 1275/2008 I priedo 1 punkte išbraukiamas įrašas „****Išk plovimo mašinos****“* (Zaikauskas 2019, 18).

4. LIETUVIŲ KALBOS DALIS GRAMATINĖJE STRUKTŪROJE

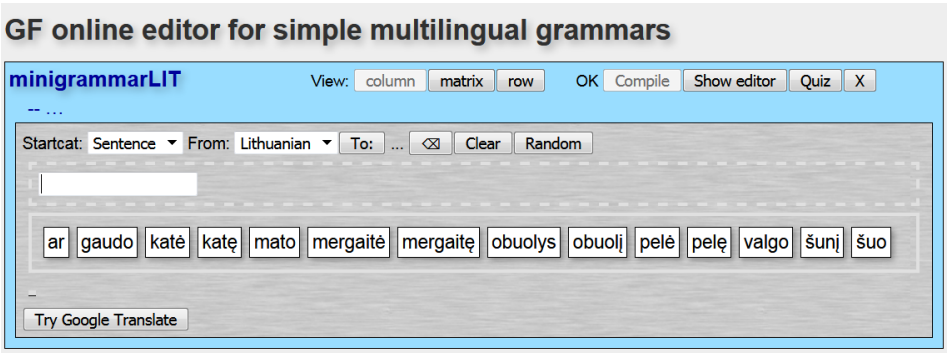
Kadangi net pačios moderniausios statistiniais ir neuroninių tinklų metodais paremtos automatinio vertimo sistemos kol kas dar negali pakeisti žmogaus darbo, todėl dalis mokslininkų vykdo tyrimus ir taisyklėmis grįsto automatinio vertimo srityje. Šiuo metu, nedidelei žodžių apimčiai taisyklinis vertimas duoda geresnius rezultatus už statistinį, tačiau problema čia yra žodžių apimtis: jų kiekiui didėjant vertimo kokybė labai krinta. Šiuo metu ieškoma būdų, kaip išspręsti šią problemą. Atliekant vertimus skaitmeninių gramatikų pagrindu tikrinami du pagrindiniai dalykai:

1. Ar vertimo metu gautas rezultatas yra gramatiškai taisyklingas sakinys (žodžių junginys) toje kalboje, į kurią verčiama.
2. Ar vertimo metu gautas rezultatas turi tą pačią reikšmę, kaip ir pateiktas vertimui sakinys (žodžių junginys).

Paskutiniu metu išpopuliarėjęs neuroninių tinklų metodas naudojamas automatinio vertimo sistemose turi ir privalumų, ir trūkumų palyginus su statistiniu vertimu. Privalumai yra tai, kad pagerėjo vidutinė vertimo kokybė ir sakiniai tapo sklandesni. Trūkumai pasireiškia tuo, kad vertimo sistemos darbą yra daug sunkiau paaiškinti ir dar labiau negalima numatyti vertimo rezultatų. Skaitmeninių gramatikų biblioteka, Gramateka, sukurta taip, kad leistinomis laikytų tik gramatiškai taisyklingas konstrukcijas (Ranta 2014, 3).

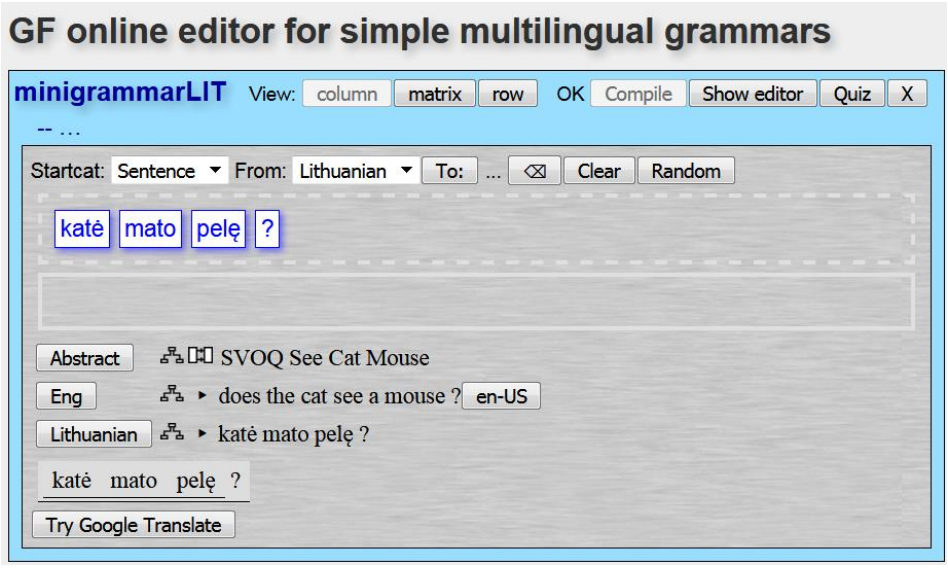
2017 metais vasaros kursų metu buvo sukurtas pirmas bandomasis lietuvių kalbos skaitmeninės gramatikos pavyzdys, kuris prieinamas internete (10 interneto nuoroda). Iš pradžių eksperimentui buvo pasirinkta nuostata atspindėti esminius skirtumus tarp anglų ir lietuvių kalbų: laisva ar griežta žodžių tvarka ir artikelių buvimas ar nebuvimas. Nutarta skirtingą žodžių tvarką lietuvių kalbos sakiniuose pasistengti atspindėti skirtingais artikeliais: Lietuvių kalbos SVO sakiniams anglų kalbos papildinį naudoti su nežymimuoju artikeliu, pvz., *Mergaitė valgo obuolį – The girl eats an apple*, nes tai neutrali žodžių tvarka. Naudojant SOV žodžių tvarką lietuvių kalboje pabrėžiamas papildinys, todėl į anglų kalbą nutarta tokį sakinį versti papildiniui naudojant žymimąjį artikelį: *Mergaitė obuolį valgo – The girl eats the apple* ir pan.

Bandomasis pavyzdys apima visus galimus žodžių tvarkos variantus lietuvių kalbos sakiniuose. Žodyną tesudaro penki daiktavardžiai: *katė, mergaitė, obuolys, pelė* ir *šuo*; trys veiksmažodžiai: *gaudyti, matyti, valgyti*; ir klausiamasis žodelis *ar*. Kaip veikia sukurtas bandomasis pavyzdys galima matyti pasirinkus skirtuko *minibar/show editor* variantą *minibar* Gramatinės struktūros tinklapyje (10 interneto nuoroda), o pasirinkus variantą *show editor* pateikiamas programinės įrangos kodas. Atsidariusiame lange yra trys skirtukai: *Abstract, English, Lithuanian*, nes šiame bandomajame pavyzdyje numatyti vertimai tarp dviejų kalbų. Skirtuke *Abstract* užkoduota abstrakti sintaksė, bendra abiems kalboms, t. y. užkoduotos sąvokos. Visas skaitmeninės gramatikos aprašas sudaromas anglų kalba. Kalbų skirtukuose pateikiami sąvokų atitikmenys konkrečios kalbos žodžiais. Skirtuko *minibar* pradinis langas pateiktas 20 pav. Jame matyti visi žodžiai, kurie įtraukti į bandomąjį pavyzdį. Paspaudus pele norimą žodį, jis perkeliamas į sakinio laukelį. Taip surenkamas visas sakinys. Kai žodžių rinkinį laukelyje, pvz., *katė mato pelę*, sistema atpažįsta kaip galimą sakinį, pateikia jo vertimą, tiksliau, abstrakčios sintaksės sąvokas bei sakinio struktūrą (šiuo atveju *SVO*) ir sakinius abiem kalbomis (21 pav.). Prie šių žodžių pridėjus klaustuką, anglų kalbos sakinys jau bus kitas, kaip ir sakinio struktūra abstrakčiojoje dalyje, kuri dabar jau yra *SOVQ* (22 pav.). Raidė *Q,* papildžiusi prieš tai buvusią struktūrą (SVO), rodo, kad tai yra klausiamasis sakinys.



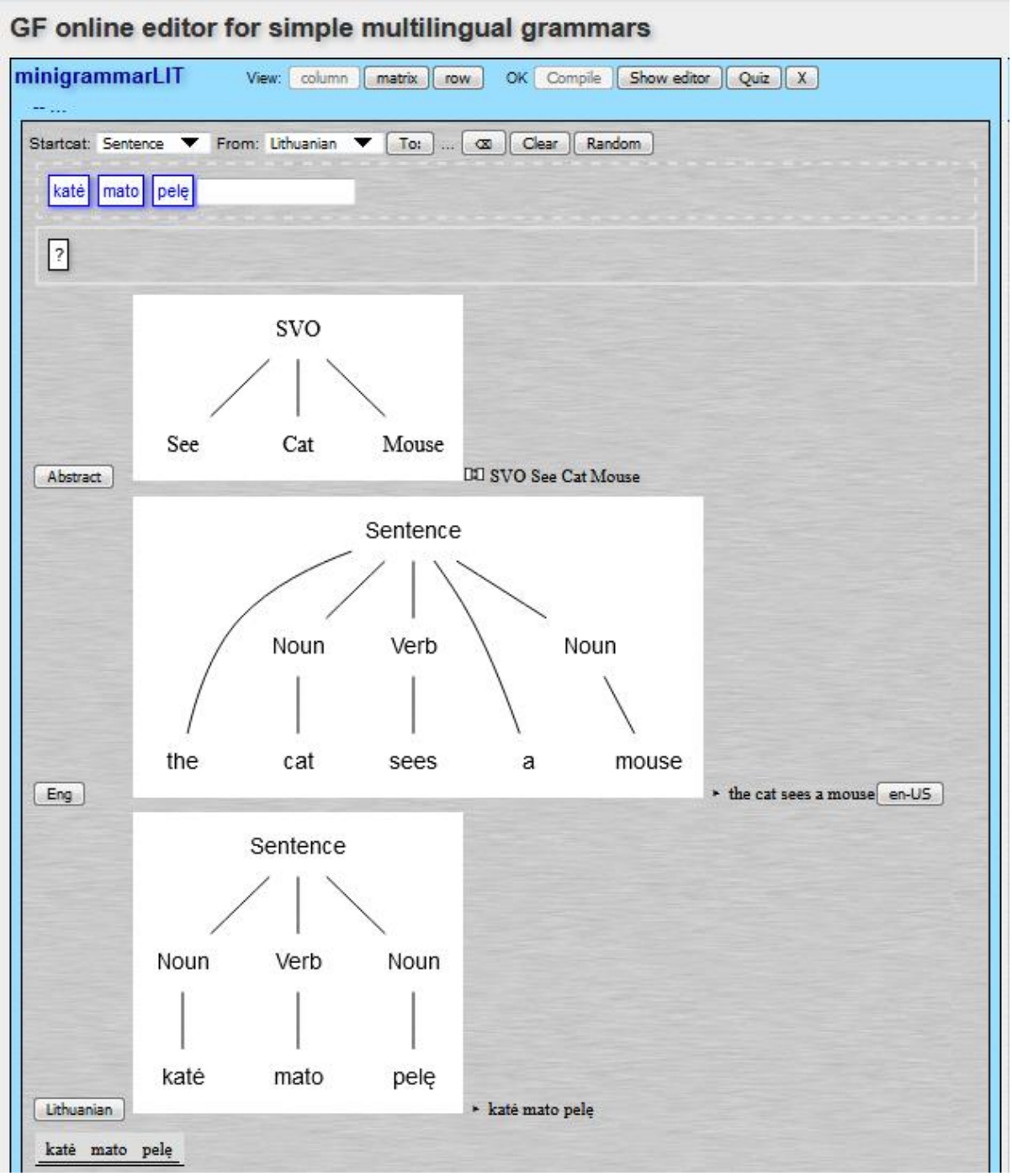
20 pav. Skirtuko *minibar/show editor* pradinis langas pasirinkus variantą *minibar* (10 interneto nuoroda)

21 pav. Sakinys *katė mato pelę* (10 interneto nuoroda)

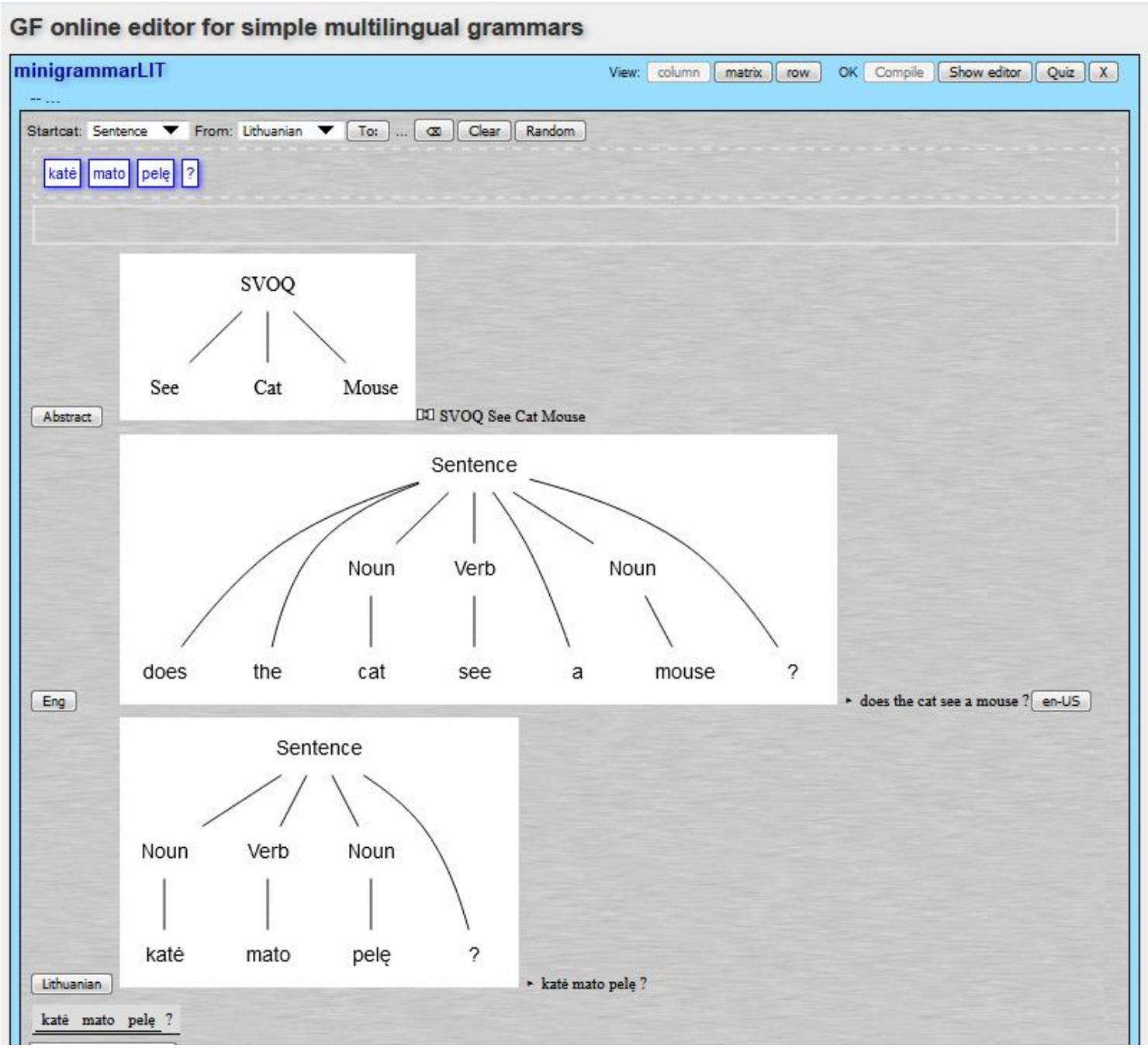


22 pav. Sakinys *katė mato pelę?* (10 interneto nuoroda)

Paspaudus pele ikonas, esančias šalia mygtukų *Abstract, Eng, Lithuanian*, galima pamatyti sintaksinių struktūrų medžius tiek konkrečios, tiek abstrakčios sintaksės. 21 pav. ir 22 pav. pateiktų sakinių struktūros parodytos atitinkamai 23 pav. ir 24 pav.



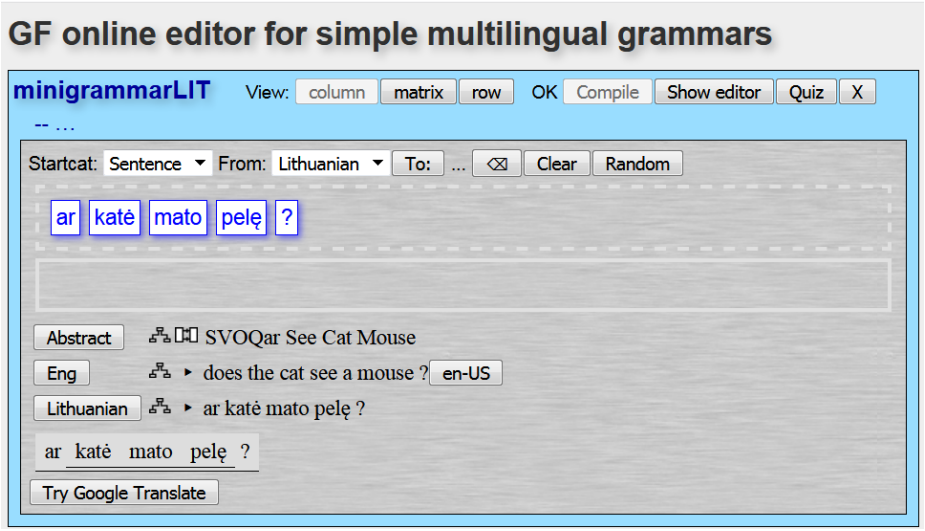
23 pav. Sakinys *katė mato pelę* su sintaksės medžiais (10 interneto nuoroda)



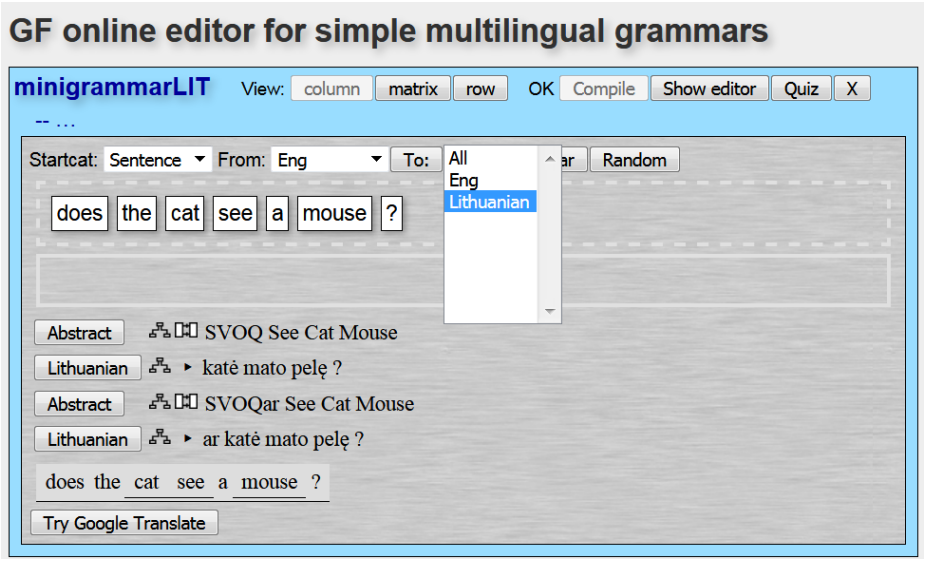
24 pav. Sakinys *katė mato pelę?* su sintaksės medžiais (10 interneto nuoroda)

Ta pati angliško sakinio *katė mato pelę?* (24 pav.) sintaksinė struktūra gaunama ir tuo atveju, kai verčiamas lietuvių kalbos sakinys *ar katė mato pelę?* (25 pav.). Šiame paveikslėlyje gerai matyti abstrakčios sintaksės sąvokų laukelyje (apatinis laukelis kairėje) pabraukti žodžiai, kurie atitinka aprašytas sąvokas. Klaustukas ir žodelis *ar* nepriklauso sąvokoms ir tarnauja kaip tarnybiniai simboliai, todėl jie nepabraukti.

Atliekant vertimus iš anglų kalbos į lietuvių kalbą sakiniui *does the cat see a mouse* pateikiami abu galimi lietuviško sakinio variantai tiek abstrakčios sintaksės lygmenyje (sakinio struktūros SVOQ ir SVOQar), tiek žodinėje sakinio išraiškoje (26 pav.). Čia taip pat labai aiškiai matyti pabrauktos abstrakčios sintaksės sąvokos; joms nepriklauso tarnybiniai žodžiai *does, the, a* ir klaustukas, todėl jie lieka nepabraukti.

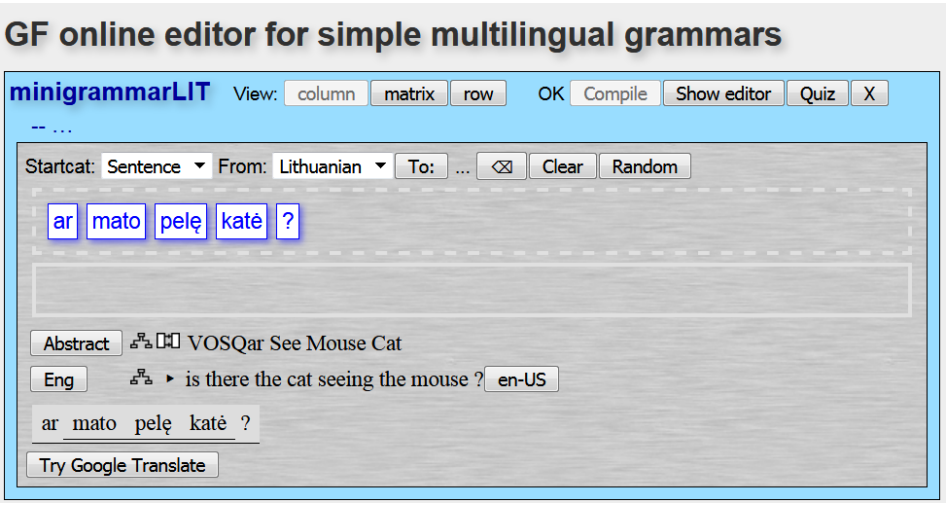


25 pav. Sakinys *ar katė mato pelę?* (10 interneto nuoroda)

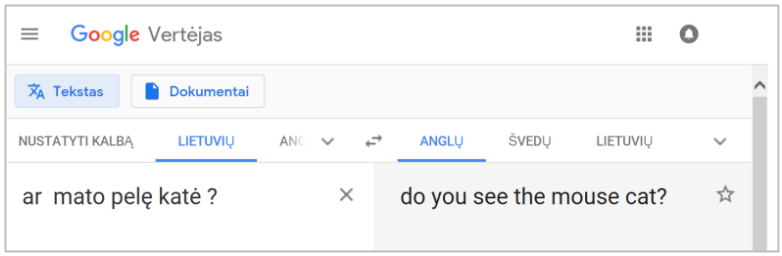


26 pav. Sakinys *does the cat see a mouse?* (10 interneto nuoroda)

Kaip jau minėta anksčiau, kuriant šį bandomąjį pavyzdį buvo stengtasi rasti skirtingus angliškų sakinių atitikmenis kiekvienam galimam lietuviško sakinio žodžių tvarkos variantui. VOS tipo sakinių vertimui pasirinkta anglų kalbos konstrukcija *there is*. Tai nėra galutinis ir nekintamas vertimo tipas. Ateityje, toliau vystant lietuvių kalbos skaitmeninę gramatiką bus tariamasi su vertėjais ir atsižvelgiama į jų pasiūlymus bei pastabas. Tačiau net ir naudojant šią konstrukciją (*there is*) skaitmeninė gramatika leidžia vertimo metu išlaikyti nepakitusias sąvokas, ko negalima pasakyti apie Google vertimus. Taigi, pabaigoje belieka palyginti lietuvių kalbos skaitmeninės gramatikos pagalba atliekamus vertimus su Google vertimais, kuriems naudojami statistiniai metodai. Reikia pasakyti, kad lietuvių kalbos sakinius, kuriuose panaudota anglų kalbos žodžių tvarka, ir Google sistema išverčia visai gerai. Netikslus vertimas prasideda tada, kai lietuviško sakinio žodžių tvarka nebeatitinka anglų kalbos žodžių išsidėstymo sakinyje. 27 pav. ir 28 pav. pavaizduotas sakinys *ar mato pelę katė*? išverstas į anglų kalbą atitinkamai skaitmeninių gramatikų ir Google vertimo sistemos pagalba. Pastaroji nesugeba išlaikyti sąvokų tapatumo, atsiranda nauji prasminiai žodžiai, kurių nebuvo lietuvių kalbos sakinyje, pvz., *you*. Be to, ir kitų žodžių vertimas neatitinka lietuvių kalbos sakinyje pavartotų sąvokų – kas yra *mouse cat* (28 pav.), neaišku.



27 pav. Sakinys *ar mato pelę katė ?* išverstas skaitmeninės gramatikos pagrindu (10 interneto nuoroda)



28 pav. Sakinys *ar mato pelę katė ?* išverstas Google vertimo sistema

Apibendrinant galima pasakyti: kaip matyti iš šių pavyzdžių, bent jau nedidelės žodžių apimties sakiniuose skaitmeninių gramatikų vertimo tikslumas yra didesnis ir pasiekiamas pagrindinis jų tikslas – išlaikyti abstrakčios sintaksės medyje ir perduoti į išverstą sakinį visas sąvokas, buvusias pradiniame sakinyje. Bent jau šiuose pavyzdžiuose sąvokų iškraipymo nebuvo, nauji atsirasdavo tik tarnybiniai žodžiai, kurie savarankiškos reikšmės neturi. To negalima pasakyti apie statistiniais metodais veikiančią Google vertimo sistemą. Išvada – neverta pasikliauti vien tikimybių teorija pagrįsta metodika, reikia vystyti ir kitus, statistika nesiremiančius vertimo metodus.

Išvados

Šiuo metu labiausiai paplitęs statistiniais ir neuroninių tinklų metodais pagrįstas automatinis vertimas kol kas žmogaus darbo pakeisti negali. Pagrindinis tokių vertimo sistemų trūkumas, kad negalima proceso kontroliuoti ir sunku numatyti rezultatus.

Skaitmeninių gramatikų pagrindu atliekamas vertimas nedidelės žodžių apimties sakiniuose duoda geresnius rezultatus, negu statistiniais metodais paremtos sistemos, tačiau esant didelei žodžių apimčiai skaitmeninių gramatikų vertimo kokybė prastėja ir nusileidžia statistiniams metodams.

Atliekant vertimus tarp daugelio kalbų naudinga būtų turėti universalią gramatiką, tačiau dėl kalbų įvairovės bandymai ją sukurti kol kas nepavyko.

Sukurtas lietuvių kabos skaitmeninės gramatikos bandomasis pavyzdys nedidelio žodžių skaičiaus sakiniams pateikia tikslesnį vertimą nei Google, ypač tais atvejais, kai lietuvių kalbos sakiniuose naudojama kita žodžių tvarka, nei anglų kalbos žodžių išsidėstymas sakinyje, nes tokiu atveju Google vertimo sistema sakinio prasmės dažniausiai nebeperduoda.

ŠALTINIAI

1 interneto nuoroda http://tekstynas.vdu.lt/page.xhtml?id=morfema-db (žiūrėta 2018-12-17)
2 interneto nuoroda http://www.semantika.lt/SyntaticAndSemanticAnalysis/Analysis (žiūrėta 2018-12-17)
3 interneto nuoroda http://www.karabatos.gr/ (žiūrėta 2018-12-17)
4 interneto nuoroda http://www.karabatos.gr/de/meine-grammatik-digital-cd-rom-f%C3%BCr-interaktive-whiteboards (žiūrėta 2018-12-17)
5 interneto nuoroda http://deutsch-digital.nl/index_grammatica.htm (žiūrėta 2014-05-30)
6 interneto nuoroda http://digitalgrammatik.blogspot.com/ (žiūrėta 2018-12-17)
7 interneto nuoroda https://www.digitalgrammars.com/ (žiūrėta 2018-12-17)
8 interneto nuoroda http://www.grammaticalframework.org/lib/doc/synopsis.html (žiūrėta 2018-12-17)
9 interneto nuoroda http://www.psytechnologijos.lt/kalba/kalbos-pinkles/ (žiūrėta 2018-12-17)
10 interneto nuoroda http://cloud.grammaticalframework.org/gfse/ skirtukas minigrammarLIT

LITERATŪRA

Baker, Mark 2003, *Lexical categories: verbs, nouns and adjectives*, Cambridge: Cambridge University Press.

Brown, Peter: Cocke, John: Della Pietra, Stephen: Della Pietra, Vincent: Jelinek, Frederick: Lafferty, John: Mercer, Robert: Roossin, Paul 1990, A Statistical Approach to Machine Translation*, Computational Linguistics* Volume 16, Number 2, 79-85.

Chomsky, Noam 1993, *Lectures on government and binding*, Berlin: Walter de Gruyter & Co.

Chung, Sandra 2012, Are Lexical categories universal? The view from Chamoro, *Theoretical linguistics* 38 (1-2), 1-56.

Barsky, Robert 2017, Universal Grammar, *Encyclopaedia Britannica*, Encyclopaedia Britannica inc., prieiga per internetą: https://www.britannica.com/topic/universal-grammar

Dąbrowska, Ewa 2015, What exactly is Universal Grammar, and has anyone seen it?, *Frontiers in Psychology*, Vol 6, Art. 852.

Daudaravičius, Vidas 2006, Pradžia į begalybę, *Darbai ir dienos* 45, 7-18.

Evans, Nicholas: Levinson, Stephen 2009, The myth of language universals: Language diversity and its importance for cognitive science, *Behavioural and Brain Sciences* 32, 429-492, doi:10.1017/S0140525X0999094X.

Everet, Daniel 2005, Cultural Constraints on Grammar and Cognition in Piraha, *Current Anthropology* 46(4), 621-646.

Floyd, Simon 2011, Re-discovering the Quechua adjective. *Linguistic Typology* 15, 25-63.

Geitgey, Adam 2016, Language Translation with Deep Learning and the Magic of Sequences, *Machine learning is fun*, Part 5, prieiga per internetą: https://medium.com/@ageitgey/machine-learning-is-fun-part-5-language-translation-with-deep-learning-and-the-magic-of-sequences-2ace0acca0aa

Grūzītis, Normunds: Dannélls, Dana 2017, A Multilingual FrameNet-based Grammar and Lexicon for Controlled Natural Language, *Language Resources and Evaluation* 51(1), 37-66.

Hallgren, Thomas: Enache, Ramona: Ranta, Aarne 2015, A Cloud-Based Editor for multilingual Grammars, Digital Grammar*, The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural language Processing: Proceedings of the Grammar Engineering Across Frameworks (GEAF) Workshop,* Taberg: Taberg Media Group AB, 41-48.

Haspelmath, Martin 2012, How to compare major word classes across the world's languages, *UCLA Working Paper in Linguistics, Theories in Everything*, Volume 17, Article 16, 109-130.

Maršalkaitė, Gabija 2018, Dirbtinis intelektas, mašininis mokymasis ir gilieji tinklai: kas slepiasi už šių sąvokų?, *TECHNO – portalas į ateitį*, prieiga per internetą:http://techo.lt/dirbtinis-intelektas-masininis-mokymasis-ir-gilieji-tinklai-kas-slepiasi-uz-siu-savoku/

McCrum, Robert 2012, Daniel Everett: 'There is no such thing as universal grammar', *The Gardian*, The guardian News and Media Limited, prieiga per internetą: https://www.theguardian.com/technology/2012/mar/25/daniel-everett-human-language-piraha

McGilvray, James 2018, Noam Chomsky, *Encyclopaedia Britannica*, Encyclopaedia Britannica inc., prieiga per internetą: https://www.britannica.com/biography/Noam-Chomsky#ref1033676

Mockus, Darius 2018, Kalbos pinklės, *Psytechnologijos*, prieiga per internetą: http://www.psytechnologijos.lt/kalba/kalbos-pinkles/

Newmeyer, Frederick 2008, Universals in Syntax, *The Linguistic Review,* 25(1-2), 35-82.

Nießen, Sonja: Ney, Hermann 2000, Improving SMT quality with morpho-syntactic analysis, *COLING '00 Proceedings of the 18th conference on computational linguistics*, Vol 2, 1081-1085, Saarbrücken, Germany, prieiga per internetą: http://acl-arc.comp.nus.edu.sg/archives/acl-arc-090501d3/data/pdf/anthology-PDF/C/C00/C00-2162.pdf

Nordquist, Richard 2018, Universal Grammar (UG), *ThoughtCo*, Dec.7, prieiga per internetą: https://www.thoughtco.com/universal-grammar-1692571

Ranta, Aarne 2009, The GF Resource Grammar Library, *Linguistic Issues in Language Technology* 2(2), CSLI Publications, prieiga per internetą: https://journals.linguisticsociety.org/elanguage/lilt/article/download/214/214-501-1-PB.pdf

Ranta, Aarne 2011, *Grammatical Framework: Programming with Multilingual Grammars,* Stanford: CSLI Publications, prieiga per internetą: http://www.grammaticalframework.org/gf-book/gf-book-slides.pdf

Ranta, Aarne 2013, *English: A Digital Grammar*, prieiga per internetą: http://www.grammaticalframework.org/lib/doc/languages/gf-english.html

Ranta Aarne 2014, Embedded controlled languages, *arXiv:1406.4057v1 [cs.CL]* prieiga per internetą: https://arxiv.org/pdf/1406.4057.pdf

Ranta, Aarne 2015, Data-Driven Documentation*: A Technique for Reliable Multilingual Information Access*, prieiga per internetą: http://www.grammaticalframework.org/~aarne/pic-2015-abstract.pdf

Ranta, Aarne 2017, *Explainable Machine Translation*, prieiga per internetą: http://www.grammaticalframework.org/~aarne/xmt-2017.pdf

Rieger, Wilhelm, 1903, *Zifferngrammatik welche mit Hilfe der Wörterbücher ein mechanisches Übersetzen aus einer Sprache in alle anderen ermöglicht,* Graz: Styria.

Rimkutė, Erika: Kazlauskienė, Asta: Raškinis, Gailius 2011, *Abėcėlinis lietuvių kalbos morfemikos žodynas*, 1 dalis, Kaunas: Vytauto Didžiojo universitetas.

Skadiņš, Raivis 2017, *Neural MT and other Language Technologies at TILDE*, prieiga per internetą: http://school.grammaticalframework.org/2017/slides/Raivis-Neural_MT_at_Tilde.pdf

Šveikauskienė, Daiva 2016, Lietuvių kalbos gramatikos informacinė sistema: I Morfologija, *Lietuvių kalba*, 10, prieiga per internetą: http://www.lietuviukalba.lt/index.php/lietuviu-kalba/article/view/183

Zaikauskas, Egidijus 2019, eTranslation viršūnės ir gelmės, *Antrasis ELRC seminaras Lietuvoje,* Vilnius vasario 1 d.

Žaliauskas, Nikodemas 2017, *Kalbėtojo atpažinimas naudojantis dirbtiniais neuroniniais tinklais: baigiamasi bakalauro darbas*, Vilniaus universitetas, prieiga per internetą: http://talpykla.elaba.lt/elaba-fedora/objects/elaba:23159558/datastreams/MAIN/content

Gauta: 2019-02-15
Priimta: 2019-05-06