ISSN 1392-0588
DARBAI ir DIENOS
2000.24

Andrius UTKA

Kalbinė įranga ir jos galimybės

Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre kuriamas Lietuvių kalbos tekstynas dabar jau peržengė 60 milijonų žodžių apimtį (Marcinkevičienė 1997). Šioje didžiulėje tekstinėje medžiagoje yra daug vertingos kalbinės informacijos leksikografams, tekstynų lingvistams ir kitiems lietuvių kalba besidomintiems asmenims. Daugelis su tekstynu pirmą kartą susidūrusių žmonių būna sugluminti pavyzdžių ir galimybių gausos – juk vietoj dešimties rašytiniuose šaltiniuose per vargus aptiktų vartosenos atvejų tekstyne per kelias minutes galima nesunkiai rasti dešimtis tūkstančių. Tai didžiulis kokybinis ir kiekybinis kalbotyrinių tyrimų šuolis, kuris verčia tyrėjus keisti tradicinius kalbos analizės metodus šiuolaikiškesniais. Dažnai nežinoma, ką su šia gausia kalbine medžiaga daryti, kaip ją apibendrinti, kaip klasifikuoti, kokias padaryti išvadas. Daugelio tokių problemų sprendimui padėtų geresnės žinios apie kalbinę programinę įrangą ir jos galimybes. Anot garsaus leksikografo:

> Kompiuterinės sistemos ir įrankiai, kurie yra vis labiau prieinami ir mokslininkui ir leksikografui praktikui, ir paprastam vartotojui, atskleidžia nesuskaičiuojamas galimybes leksinei informacijai pateikti ir panaudoti (Swanepoel 1994; čia ir kitur versta mano. – A. U.).

Akivaizdu, kad tekstynų lingvistikos pažanga yra tiesiogiai susijusi su informatikos mokslo raida ir su efektyvesniu informatikų bei lingvistų bendradarbiavimu, – juk norint geriau aprėpti ir išanalizuoti didėjantį informacijos kiekį reikia naudoti vis pažangesnes lingvistines kompiuterines priemones.

Šiame straipsnyje bus mėginama apžvelgti pagrindines kalbinės programinės įrangos rūšis ir jų funkcijas. Kelerių pastarųjų metų darbo patirtis Kompiuterinės lingvistikos centre parodė, kad lietuvių kalbos tekstynas dažniausiai buvo naudojamas dviejų rūšių informacijai gauti – žodžių dažninėms charakteristikoms bei konkordansams, todėl straipsnyje bus detaliau gilinamasi į dažninių sąrašų generatorių ir konkordavimo programas.

Straipsnis iliustruojamas pavyzdžiais, gautais dirbant su anglų lingvisto Mike'o Scott'o 1986 m. sukurtu kalbinės programinės įrangos paketu „WordSmith Tools". Jis vykdo svarbiausias funkcijas, leidžiančias kurti ir analizuoti žodžių dažnines charakteristikas bei konkordansus ir dėl to gerai žinomas tekstynų lingvistikos specialistams.

KALBINĖS PROGRAMINĖS ĮRANGOS RŪŠYS

Dėl didelės gausos ir įvairovės išsamiai apžvelgti visą kalbinę programinę įrangą (angl. *lingware*) būtų neįmanoma, todėl šiame straipsnyje paminėsime tik reikšmingiausias jų rūšis. Šias programas galima skirstyti pagal įvairius jų bruožus: atliekamas funkcijas, pritaikymo konkrečioms kalboms galimybes, patikimumą, tinkamumą tam tikroms operacinėms sistemoms, naudojamą metodiką, be to, dar į komercines ir nemokamas, visiškai užbaigtas ir eksperimentines, statistines ir paremtas taisyklėmis. Siekdami labiau apibendrinti šią kompiuterinių priemonių gausą, bandysime jas sugrupuoti pagal atliekamas funkcijas ir paskirtį.

Konverteriai. Esant didžiulei įvairių tekstinių kodavimų ir formatų įvairovei, reikalinga tam tikra įranga, kuri leistų konvertuoti tekstą į norimą formatą. Tai pirmiausia įvairūs konverteriai, kurie keičia lietuviškų raidžių kodavimus (KBL, Baltic Rim, HTML ir SGML kodai ir t.t.); kiti konverteriai keičia failo tipą, pvz., HTML ar SGML failus konvertuoja į tekstinius arba atvirkščiai. Daugelis tokių programėlių sukurtos tik vienam kuriam darbui. Nors jos gali būti laikomos nereikšmingomis kaip labai siauros paskirties simbolių keitikliai, vis dėlto tekstynų lingvistas, intensyviai dirbantis su elektroniniais

tekstais, be šių programų pagalbos neišsivers. Be to, tokie konverteriai parodo, su kokia įvairia tekstine medžiaga tenka susidurti lingvistui.

Dažninių sąrašų generatoriai. Šios kompiuterinės priemonės skaičiuoja dažnines žodžių charakteristikas, t.y. generuoja dažninius sąrašus (*frequency lists*). Tuose sąrašuose nurodomas kiekvienos žodžio formos pasirodymo dažnis tiriamame tekste (*žr. 1 lentelę*). Dažniniai sąrašai gali būti pateikti abėcėline arba dažnio tvarka. Dažniniai sąrašai suteikia vertingos informacijos apie santykinį žodžių dažnį ir teksto leksemų įvairovę.

Konkordavimo programos. Konkordavimo programa yra vienas svarbiausių tekstynų lingvisto ar leksikografo įrankių. Iš esmės ši programa yra teksto arba tekstyno paieškos sistema, leidžianti lingvistui rasti reikalingą informaciją apie atskirus teksto elementus (žodžius ar žodžių junginius) ir pateikti ją patogia forma – konkordansu. Konkordansas yra sąrašas eilučių, kuriose buvo rastas tiriamas žodis ar žodžių junginys, paimtas iš teksto ar tekstyno. Plačiau apie dažninius sąrašus bei konkordavimo programų galimybes bus kalbama kituose šio straipsnio skyriuose.

Sintaksinės analizės programos. Šios programos (*parsers)* yra sudėtingesnė programų rūšis, leidžianti analizuoti sakinių sintaksę. Paprastai jos identifikuoja žodžius sakinyje, nustato jų sintaksinį priklausymą, sugrupuoja į aukštesnio lygio vienetus (žodžių junginius ir prijungiamuosius sakinius) ir atitinkamai juos pavadina (McEnery and Wilson 1996:129). Sintaksinės analizės būna pagrįstos tikimybiniais skaičiavimais (*probabilistic/ stochastic parser*) arba iš anksto sudarytomis taisyklėmis (*rule-based parser)*. Lietuvių kalbai sintaksinės analizės programa dar nesukurta.

Klaidų tikrintuvai. Tai bene geriausiai žinoma kalbinės programinės įrangos rūšis, kurios svarbiausia funkcija – surasti klaidas elektroniniame tekste ir padėti jas ištaisyti. Pagrindinis klaidų tikrintuvo komponentas yra vidinis elektroninis žodynas, nuo kurio priklauso programos efektyvumas: programa su išsamesniu žodynu atpažįsta daugiau žodžių ir tokiu būdu tekstą patikrina greičiau. Vartotojui taip pat yra svarbi galimybė šį elektroninį žodyną pildyti savarankiškai. Komercinė kompanija UAB „Fotonija" lietuvių kalbai sukūrė ir toliau tobulina klaidų tikrintuvą „Juodos avys", kuris paremtas „Dabartinės lietuvių kalbos žodynu" (1972), „Tarptautinių žodžių žodynu" (1982), o žodžių kaitybinės formos atpažįstamos remiantis gramatikos taisyklėmis, esančiomis daugiatomėje gramatikoje „Lietuvių kalbos gramatika. Fonetika ir morfologija" (1965, 1971).

Skiemenuokliai. Šios kompiuterinės programos automatiškai perkelia žodžius į kitą eilutę pagal žodžių kėlimo taisykles. Skiemenuokliai leidžia geriau išnaudoti teksto eilutę, kai teksto ilgis yra ribojamas. Dėl to šios programos yra ypač svarbios žurnalų ir laikraščių leidėjams. Lietuvių kalbai skiemenuoklį „Skie-muo" sukūrė jau minėta UAB „Fotonija".

Lemuokliai. Jie įvairias vieno žodžio formas sujungia į vieną antraštinę – lemą. Pvz., žodžio *knyga* lemą sudaro šie žodžiai: *knyga, knygos, knygai, knygą, knygoje, knygos, knygų, knygoms, knygas, knygose, knygomis.* Sulemuotas tekstynas ar žodžių sąrašas leidžia lingvistui šias skirtingas gramatines žodžių formas traktuoti kaip vieną. Lietuvių kalbai lemuoklį sukūrė Vytautas Zinkevičius (žr. jo straipsnį šiame „Darbų ir Dienų" tome).

Morfologiniai analizatoriai. Šios rūšies kompiuterinės programos automatiškai nustato analizuojamo žodžio gramatines charakteristikas. Paprastai morfologiniai analizatoriai kiekvieną žodį analizuoja nepriklausomai nuo konteksto. Dėl to morfologinė analizė susiduria su sudėtinga problema – įvairių rūšių homonimija arba daugiaprasmiškumu (*ambiguity*). Pvz., žodis *knyga* gali turėti net tris gramatines interpretacijas: vienaskaitos vardininkas, vienaskaitos įnagininkas ir vienaskaitos šauksmininkas. Tobulesni morfologiniai analizatoriai gali panaikinti šį daugiaprasmiškumą (*disambiguation*), analizuodami kontekstinę žodžio aplinką.

Anotatoriai. Tai specialios kompiuterinės programos, kurios prie žodžių ar kitų tekstinių vienetų prirašo tam tikras pažymas (*tags*), aiškinančias jų bruožus. Šiuo žymėjimu elektroninis tekstas praturtinamas lingvistinės arba struktūrinės informacijos. Siekiant standartizuoti šią informaciją, pažymos sutvarkomos pagal griežtą hierarchinę kodavimo

sistemą, kuri sutinka su tam tikru elektroninio teksto žymėjimo standartu. Paprastai kalbinė programinė įranga bei duomenys suderinamos su SGML standartu (*Standard Generalized Markup Language* – apibendrinta standartine žymėjimo kalba) ir su „TEI rekomendacijomis" – *Text Encoding Initiative Guidelines* (Ide and Veronis 1995; Sperberg-McQueen and Burnard 1994) (*žr. 1 priedą*).

Atskira anotatorių rūšis yra kalbos dalių žymekliai, kurie nustato kiekvieno teksto žodžio kalbos dalį. Jų veikimas – trijų žingsnių: suskirsto tekstus į žodžius, kiekvienam žodžiui pateikia potencialią morfosintaksinę interpretaciją ir panaikina daugiaprasmiškumą (Chanod 1997). Savo funkcijomis jie panašūs į morfologinius analizatorius, skiriasi tik tuo, kad prioritetinė morfologinio analizatoriaus funkcija yra pateikti gramatines žodžio charakteristikas, o anotatoriaus – vienareikšmiškai sužymėti tekstą pagal tam tikrą standartą. Galima sakyti, kad anotatoriai yra patobulinti morfologiniai analizatoriai.

Paralelinimo programos. Jos skirtos vertimo ir originalo kalbų tekstams lygiagretinti. Tarkim, turime du tekstus – A ir B. A yra originalus tekstas, o B yra A teksto vertimas (*žr. 2 priedą*). Daugumos paralelinimo programų pagrindinis uždavinys yra automatiškai surasti, kuris sakinys ar sakiniai tekste A atitinka sakiniui ar sakiniams tekste B (McEnery and Oakes 2000:2). Kitas, sudėtingesnis ir ambicingesnis žingsnis yra automatiškas dviejų kalbų paralelinimas žodžių lygmenyje. Reikia pasakyti, kad paralelinimo programos efektyviai veikia tik su tokiais tekstais, kuriuose yra pažymėtos paragrafų ir sakinių ribos. Todėl prieš paralelinant tekstus jie žymimi pagal paralelinimo programos standartus. Suparalelinti tekstai gali duoti naudingos informacijos besimokantiems užsienio kalbų, be to, lingvistams ir vertėjams. Kai kurie lingvistai teigia, kad paralelinimo programos yra vertingiausia, kas iki šiol buvo nuveikta mašininio vertimo kryptimi.

Kompiuterinės lingvistikos centras dalyvavo projektuose, kuriuose George'o Orwell 'o „1984-ieji" lietuviškas vertimas iš anglų kalbos suparalelintas su 11, o Platono „Respublika" – su 17 Europos šalių kalbų vertimais (Erjavec, Lawson, and Romary 1997). Tiesa, šie ekspermentai parodė, kad tuose projektuose naudota paralelinimo programa „Vanilla" (Danielsson and Ridings 1997) nedavė visiškai tikslių rezultatų, ir dėl to rezultatai turėjo būti koreguojami rankiniu būdu.

Leksinės duomenų bazės – labai svarbus įrankis, leidžiantis analizuoti didžiulį leksinės informacijos kiekį. Duomenų bazėse formalizuota leksinė informacija išdėstyta keliais lygiais (Gellerstam 1995:61). Speciali duomenų bazių paieškos sistema suteikia lingvistui galimybę greitai gauti reikalingą leksinę informaciją pagal pageidaujamus parametrus.

Mašininio vertimo sistemos skirtos automatiškai versti tekstams iš vienos kalbos į kitą. Jau egzistuoja automatinio vertimo sistemos, kurios gerokai pagreitina ir palengvina tekstų vertimus, tačiau šios sistemos dažniausiai verčia siauros paskirties tekstus, pvz., tam tikros technikos instrukcijas arba teisinės kalbos dokumentus. Anot vokiečių lingvisto Wolfgango Teuberto (1997:148), „niekada nebus sukurta tokia mašininio vertimo sistema, kuri galės pateikti teisingus ir galutinius vertimus 'atviriems tekstams', priklausantiems tam tikram kultūriniam ar socialiniam diskursui". Taigi didžiausias šiuolaikinių mašininių vertimo sistemų trūkumas yra tas, kad jos analizuoja ir verčia izoliuotus nuo platesnio kalbinio, kultūrinio ir socialinio konteksto sakinius (Ramm 1994:7).

DAŽNINIAI ŽODŽIŲ SĄRAŠAI

Kaip jau minėta, dažniniai žodžių generatoriai sudaro dažninius sąrašus (*frequency lists*), kuriuose nurodomas kiekvienos atskiros žodžio formos pasirodymo dažnis tiriamame tekste (*žr. 1 lentelę*). Dažninis sąrašas kuriamas paverčiant sudėtingą ir daugiaprasmę tekstinę informaciją paprastesne, vienareikšmiška. Kitaip sakant, dėl žodžių dažnių atsisakoma visos kitos informacijos, kaip žodžių konteksto, išsidėstymo tvarkos, skyrybos ir kt. Pirmieji dažniniai lietuvių kalbos žodynai buvo parengti Lietuvių kalbos institutui bendradarbiaujant su Matematikos ir informatikos institutu. Šie dažniniai žodynai paremti 1,2 mln. žodžių tekstyno pamatu. Kol kas pasirodė du „Dažninio dabartinės rašomosios lietuvių kalbos žodyno" variantai:

viename antraštiniai žodžiai surūšiuoti abėcėlės tvarka (Grumadienė ir Žilinskienė 1998), kitame – mažėjančio dažnio tvarka (Grumadienė ir Žilinskienė 1997).

Kompiuterinės lingvistikos centre lietuvių kalbos tekstyno žodžių formų dažniniai sąrašai kasmet atnaujinami. *1-je lentelėje* 50 dažniausių žodžių yra paimti iš žodžių dažninio sąrašo, kuris sukurtas 60 mln. tekstyno pamatu.

Kam reikalingi dažniniai sąrašai? Pirmiausia tam, kad atsakytume į paprastą klausimą: kurie žodžiai dažnesni, o kurie retesni. Galbūt ir galima tą nujausti, bet vis dėlto žmogaus intuicija yra linkusi klysti, o jos teiginius sunku paremti kuo nors kitu nei pačia intuicija. Tuo tarpu dažninį sąrašą sudaro objektyvūs, patikrinami duomenys, ir tai leidžia daryti pagrįstas išvadas apie žodžių santykinių dažnių tendencijas tekste.

Didelio tekstyno dažninis sąrašas turėtų būti nepakeičiamas leksikografo įrankis. Anglų lingvistė Della Summers (1996:261–262) teigia, kad žodžių dažnių informacija lemia leksikografiją, o labiausiai dažninė informacija turėtų prisidėti prie žodžių reikšmių rikiavimo žodynuose laikantis principo: žodyniniuose straipsniuose dažniau vartojamos žodžių reikšmės turėtų būti pirmesnės. Dažniniai žodžių sąrašai taip pat atsako į klausimus: koks teksto žodingumas, katras iš dviejų vienodos reikšmės žodžių dažnesnis, kokio ilgio tiriamos žodžio formos konkordansas, kokios rūšies leksika vyrauja tiriamame tekste. Be to, dažniniai sąrašai naudojami įvairiuose statistiniuose skaičiavimuose, pvz., aptinkant statistiškai reikšmingas kolokacijas.

Dažniniai sąrašai gali būti surūšiuoti pagal dažnį arba abėcėlę, taip pat didėjančia arba mažėjančia tvarka. Pagal dažnį surūšiuotuose sąrašuose išryškėja dažniausi žodžiai ir jų formos. *1-oje lentelėje* pateikiamas 50 dažniausių žodžių formų sąrašas tekstyne. Iš šio sąrašo akivaizdu, kad absoliučiai dažniausias žodis tekstyne yra jungtukas *ir*, dažniausias prielinksnis – *į*, dažniausias prieveiksmis – *tik*, tarp veiksmažodinių formų dažniausia yra žodžio *būti* būtojo laiko III asmens forma *buvo*, tarp daiktavardinių formų – žodis *Lietuva* kilmininko linksniu.

1 lentelė

50 dažniausių nekaitomų žodžių ir kaitomų žodžių formų 60 mln. žodžių lietuvių kalbos tekstyne

Žodis	Dažnis	Žodis	Dažnis
1. ir	1909469	26. dar	138558
2. kad	485436	27. po	130015
3. į	455142	28. už	125449
4. iš	384463	29. per	124967
5. su	334613	30. dėl	122274
6. o	307064	31. bei	120441
7. buvo	**307006**	32. tačiau	117195
8. tai	296544	33. kas	113630
9. kaip	290564	34. jos	112155
10. yra	272266	35. a	108893
11. tik	233566	36. to	102668
12. ar	225493	37. metų	99506
13. ne	216942	38. labai	97922
14. Lietuvos	**216773**	39. gali	94044
15. savo	211425	40. mūsų	94024
16. bet	207853	41. būti	93557
17. jis	178280	42. turi	92320
18. apie	178234	43. d	91832
19. m	172902	44. arba	90331
20. nuo	163268	45. jie	90032
21. taip	162723	46. prie	89469
22. jo	151041	47. iki	89367
23. kai	149361	48. ji	89170
24. jų	147543	49. pat	88633
25. jau	142905	50. nors	83821

Pagal abėcėlę išrūšiuotose sąrašuose labai patogu sugrupuoti visus bendrašaknius ir giminiškus žodžius. Šiuose sąrašuose taip pat lengviau surasti konkretų žodį (ypač jeigu šis sąrašas išspausdintas). *2-oje lentelėje* matome žodžio *filologija* dažnių pasiskirstymą pagal linksnius abėcėlės tvarka išrikiuotame dažniniame sąraše.

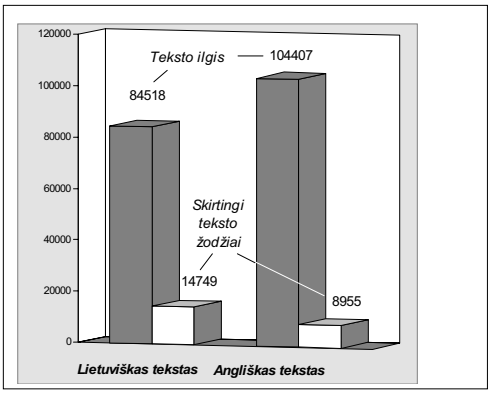
2 lentelė

Žodis *filologija* pagal abėcėlę surūšiuotame dažniniame sąraše

Žodis	Dažnis
filologes	1
filologės	10
filologija	26
filologiją	43
filologijai	8
filologijas	1

filologijoje	10
filologijomis	1
filologijos	367
filologine	12
filologinę	12
filologinė	15
filologinei	2
filologinėje	1

Dėl platesnės lietuvių kalbos daiktavardžių ir veiksmažodžių paradigmos lietuvių kalbos žodžių formų dažniniai sąrašai gerokai ilgesni nei analitinių kalbų (pvz., anglų, vokiečių, prancūzų). Pailiustruosime tai palygindami tos pačios knygos lietuviško vertimo ir angliško originalo žodžių formų dažninius sąrašus. *1 paveiksle* matyti, kad nors lietuviškos knygos teksto ilgis matuojant žodžiais (84 518 žodžių) yra trumpesnis nei originalo (104 407), jos skirtingų žodžių sąrašas (arba dažninis sąrašas) yra 1,6 karto ilgesnis. Tai rodo lietuviškojo teksto žodžių ir žodžių formų įvairovę.



1 paveikslas. Teksto ilgis ir skirtingų žodžių kiekis Džordžo Orvelo knygos „1984-ieji" originale ir vertime

Žodžių formų sąrašo trūkumas yra ne tik jo ilgis, bet ir žodžio dažnio pasiskirstymas per žodžio morfologines formas. Juk iš tiesų tikrasis antraštinio žodžio *filologija* dažnis yra visų to žodžio linksnių dažnių suma (*žr. 2 lentelę*, 26 + 43 + 8 + 1 + 10 + 1 + 367 = 456). Šią problemą padeda spręsti lemuokliai, kurie suveda visas gramatines žodžio formas į vieną antraštinę. Lemuotas žodžių sąrašas sutrumpėja, be to, gali pasikeisti žodžių išsidėstymo tvarka surūšiavus jį pagal dažnį, tačiau lemuojant žodžių sąrašą susiduriama su homonimijos problema. Juk, pvz., žodžio forma *laužo*, kuriai dažniniame sąraše yra priskiriamas tik vienas dažnis, iš tiesų skyla į daiktavaržio *laužas* kilmininką ir veiksmažodžio *laužyti* esamojo laiko formą.

Lemavimas buvo taikytas ir jau minėtame „Dažniniame dabartinės rašomosios lietuvių kalbos žodyne", kur pateikti vien antraštinių žodžių dažniai, tik šiuo atveju lemuojamas buvo pats tekstynas, o ne žodžių sąrašas. Tiesa, reikia neužmiršti, kad antraštinių žodžių sąraše prarandama informacija apie atskirų linksnių arba veiksmažodžio formų dažnius.

KONKORDANSAI

Kaip jau minėta, konkordavimo programa yra teksto arba tekstyno paieškos sistema, leidžianti lingvistui rasti reikalingą informaciją apie atskirus teksto elementus (žodžius ar žodžių junginius) ir pateikti ją patogia forma – konkordansu. *2 paveiksle* matote žodžio *filologija* konkordansą.

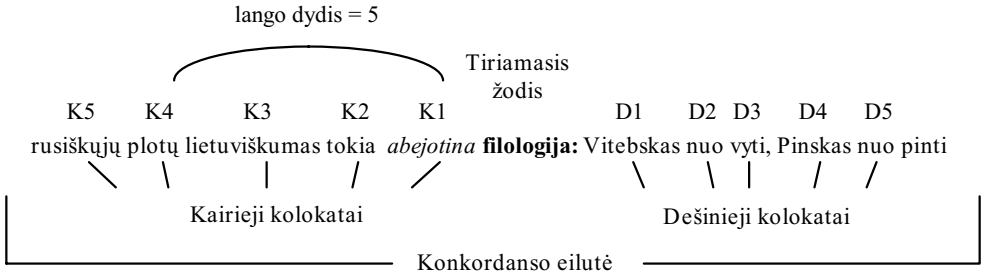
Konkordanso eilutės centre (žr. *3 pav.*) yra tiriamasis žodis (*node*), o jo artimiausiame kontekste iš kairės – kairieji kolokatai, iš dešinės – dešinieji (*left collocates, right collocates*). Kolokacijų tyrimuose dažnai būtina apriboti kontekstą, t.y. priimti domėn tik tam tikrą kolokatų skaičių iš kairės arba iš dešinės, kitaip tariant, nustatyti „langо dydį" (*span*). Dažniausiai nustatomas 4–5 žodžių langо dydis. Kompiuterių programose dėl patogumo kairieji kolokatai žymimi sutrumpintai K1–K5, o dešinieji – D1–D5 (angl. *L1–L5, R1–R5*), kur skaičius nurodo, kurioje pozicijoje yra kolokatas.

Dauguma konkordavimo programų leidžia didinti konkordanso eilutės ilgį: eilutę galima apriboti sakiniu arba tiksliai nurodyti jos ilgį simboliais. Tobulesnės programos, be to, leidžia kurti bendrašaknių arba bendragalūnių žodžių konkordansus. Pvz., jei paieškos langе nurodysime *filolog**, konkordavimo programa mums suras visus žodžius su šia šaknimi, jeigu *gija – visus žodžius su šia galūne.

2 paveikslas. Nerūšiuotas žodžio *filologija* konkordansas

1 amo abstrakcijos ir universalumo lygio struktūras	(filologija	Akcentuoja konkretybės šventumą). Galg
2 tai ne taip akivaizdu, bet tiek pat nuostolinga.	filologija	su saviironišku budrumu žodiniam, sutar
3 hermeneutinio susitikimo būdai taps archeologija.	filologija	Nebepažins mylėtino Logoso. Buvau atkre
4 , – tam tikėjimui nepaliaujamai stiprinti gyvuoja	filologija.	Jos prielaida ta, kad esama tų nedauge
5 knygas rašo arba gali rašyti. Noriu pasakyti, kad	filologija	Remiasi kilniu tikėjimu – tikėjimu tuo,
6 menkų pastangų reikalaujantis autorius, jį domina	filologija	ir filosofija, jis retai kada praleidži
7 ama rusiškųjų plotų lietuviškumas tokia abejotina	filologija:	Vitebskas nuo vyti, Pinskas nuo pinti,
8 tete įkurtas naujas didaktikos skyrius – lietuvių	filologija,	Kuriai iki šiol vadovauja profesorius
9 o ir tyrinėjimų disciplinos pagrindas lieka baltų	filologija,	tai yra lietuvių ir/arba latvių kalba,
10 retacijos perskyra taip pat tradicinė. Ja remiasi	filologija:	Vienas dalykas yra tekstas ir jo prasm
11 net studijuoti tokius nepraktiškus mokslus kaip	filologija.	Baigęs mokslus, B. Brazdžionis dėstė l
12 užgesino" mama. "Vaikeli, anokia čia specialybė –	"filologija"!,	– tikino ji. – Juk nebūtina studiju
13 slo šakomis: antropologija, biologija, zoologija,	filologija	(senais vietovardžiais)104. Nors, anot
14 avus į kitą kraštą, kiti ir universitetai. Taigi	filologija,	Nors to termino aš ir nelabai suprantu
15 vo, perkeitė ir pertvarkė tokios disciplinos kaip	filologija,	Kuri savo ruožtu natūralizavo, modern
16 vadinimas neaiškus (studijų planuose – "lietuvių	filologija",	Mano skaitytojo biliete – "lietuvių k
17 semiotikos šakos, hermeneutika, komparatyvinė	filologija	ir kt.). Šį dvinarį bakalauro studij
18 s jie įteigė inžinieriaus, mediko specialybes, o	filologija	Neturėjo pasisekimo. Taip lietuviška s
19 s Redakcijos svečias Baltistika, arba baltų	filologija,	– mokslas apie baltų kalbas, literatū
20 tus ir filologinių specialybių spektras: lietuvių	filologija,	Anglų filologija, vokiečių filologija,
21 specialybių spektras: lietuvių filologija, anglų	filologija,	Vokiečių filologija, vokiečių ir rusų
22 : lietuvių filologija, anglų filologija, vokiečių	filologija,	Vokiečių ir rusų kalbos, kurias daugia
23 iš medicinos srities atstovų, antroje vietoje –	filologija,	Trečioje – technikos mokslai. Susitik
24 va, – teigė menininkas, – pagimdė ne istorija, o	filologija;	Sunku įsivaizduoti didesnę negu lietu
25 slysta balsės į priebalsius, ir stambaus kalibro	filologija	Lieka bejėgė išreikšti tą meilės ir jau
26 al iš šimto dalykų. Dabar vienas tokių – ir baltų	filologija.	– Pernai studentų buvo tik keturi.
27 as neateitų, be to, būtų per sudėtinga. Baltų	filologija	Kaip atskira specialybė dėstoma tik tri
28 igynė disertaciją "Latvių kalbos sėliškųjų tarmių	filologija".	1980–1990 metais dirbo mokslinį darbą
29 versitete populiariausia nauja specialybė – anglų	filologija.	Norinčių studijuoti ją, kaip ir ekonom

3 paveikslas. Konkordavimo programose vartojamų terminų iliustracija



Jeigu tiriame dešimt ar penkiolika konkordanso eilučių, mums nereikia jokios ypatingos analizavimo metodologijos: mes galime puikiai patys įvertinti tą, ką matome. Jei tiriame 50–100 eilučių konkordansą, tai rekomenduotina jam pritaikyti kokį nors elementarų konkordanso rūšiavimą. O jeigu mūsų medžiaga susideda iš tūkstančio ar daugiau eilučių, mums tiesiog būtina pritaikyti konkordansui specialias rūšiavimo ir skaičiavimo procedūras, kitaip nesugebėsime aprėpti visos jame esančios informacijos.

Rūšiavimo nauda akivaizdi iš šio pavyzdžio. *2 paveiksle* žodžio *filologija* konkordanso eilutės yra nesurūšiuotos. Skaitant jį paaiškėja, kad *filologija* iš kairės dažniausiai sudaro junginius su kalbų pavadinimais (*lietuvių*, *baltų*, *anglų*, *vokiečių*). Vis dėlto net ir

šiame trumpame konkordanse sunku greitai pasakyti, kuris šių junginių yra dažnesnis. Jei surūšiuosime konkordansą pagal kairiuosius kolokatus, iš pradžių pagal pirmąjį iš kairės (K1), po to pagal antrąjį iš kairės (K2), gausime konkordansą (žr. *4 paveikslą*), kuriame žymiai lengviau pamatyti šiuos besikartojančius junginius bei įvertinti jų dažnį. Lygiai taip pat konkordanso eilutes galima rūšiuoti ir pagal dešiniuosius kolokatus.

Besikartojantys junginiai lengviau matomi naudojant vadinamąjį zigzaginį rūšiavimą (*zigzag sorting*) (Altenberg and Eeg-Olofsson 1990; Oakes 1998). Rūšiuojant zigzaginiu būdu, pirmiausia rūšiuojama pagal tiriamąjį žodį, po to pagal pirmąjį iš dešinės; pagal pirmąjį iš kairės; pagal antrąjį iš dešinės; antrąjį iš kairės ir t.t. Tokiu būdu susidaro besikartojančių junginių grandinės, kurios lengvai pastebimos vizualiai.

Jei mūsų konkordansas yra 1000 ir daugiau eilučių, mums gali nepadėti ir kolokatų rūšiavimas. Tada reiktų automatiškai skaičiuoti kolokatus ir nustatyti tikslų kolokatų skaičių tam tikroje pozicijoje. Dauguma konkordavimo programų leidžia atlikti tokius skaičiavimus.

3-oje lentelėje pateiktas kolokatų sąrašas, sudarytas žodžio *filologija* konkordansui. Čia įtraukti tik tie kolokatai, kurie žodžio *filologija* kontekste (lango dydis 5 žodžiai) pasirodo dažniau nei 1 kartą. Šis sąrašas surūšiuotas pagal K1 stulpelį, vadinasi, dažniausi kolokatai K1 pozicijoje yra šio sąrašo viršuje. Iš sąrašo matyti, kad dažniausias *filologijos* kolokatas šioje pozicijoje yra žodis *baltų* (*žr.*

4 paveikslas. Žodžio *filologija* konkordansas, rūšiuotas pagal K1 ir po to pagal K2

1 ama rusiškųjų plotų lietuviškumas tokia *abejotina*	filologija:	Vitebskas nuo vyti, Pinskas nuo pinti,
2 specialybių spektras: lietuvių filologija, *anglų*	filologija,	vokiečių filologija, vokiečių ir rusų
3 versitete populiariausia nauja specialybė – *anglų*	filologija.	Norinčių studijuoti ją, kaip ir ekonom
4 hermeneutinio susitikimo būdai taps *archeologija*.	filologija	nebepažins mylėtino Logoso. Buvau atkre
5 as neateitų, be to, būtų per sudėtinga. *Baltų*	filologija	kaip atskira specialybė dėstoma tik tri
6 s Redakcijos svečias Baltistika, arba *baltų*	filologija,	– mokslas apie baltų kalbas, literatū
7 o ir tyrinėjimų disciplinos pagrindas lieka *baltų*	filologija,	tai yra lietuvių ir/arba latvių kalba,
8 al iš šimto dalykų. Dabar vienas tokių – ir *baltų*	filologija.	– Pernai studentų buvo tik keturi.
9 menkų pastangų reikalaujantis autorius, jį *domina*	filologija	ir filosofija, jis retai kada praleidži
10 , – tam tikėjimui nepaliaujamai stiprinti *gyvuoja*	filologija.	Jos prielaida ta, kad esama tų nedauge
11 knygas rašo arba gali rašyti. Noriu pasakyti, *kad*	filologija	remiasi kilniu tikėjimu – tikėjimu tuo,
12 vo, perkeitė ir pertvarkė tokios disciplinos *kaip*	filologija,	kuri savo ruožtu natūralizavo, modern
13 net studijuoti tokius nepraktiškus mokslus *kaip*	filologija.	Baigęs mokslus, B. Brazdžionis dėstė l
14 slysta balsės į priebalsius, ir stambaus *kalibro*	filologija	lieka bejėgė išreikšti tą meilės ir jau
15 ė, semiotikos šakos, hermeneutika, *komparatyvinė*	filologija	ir kt.). Šį dvinarį bakalauro studij
16 vadinimas neaiškus (studijų planuose – "*lietuvių*	filologija",	mano skaitytojo biliete – "lietuvių k
17 tete įkurtas naujas didaktikos skyrius – *lietuvių*	filologija,	kuriai iki šiol vadovauja profesorius
18 tus ir filologinių specialybių spektras: *lietuvių*	filologija,	anglų filologija, vokiečių filologija,
19 tai ne taip akivaizdu, bet tiek pat *nuostolinga*.	filologija	su saviironišku budrumu žodiniam, sutar
20 s jie įteigė inžinieriaus, mediko specialybes, *o*	filologija	neturėjo pasisekimo. Taip lietuviška s
21 vą, – teigė menininkas, – pagimdė ne istorija, *o*	filologija;	sunku įsivaizduoti didesnę negu lietu
22 retacijos perskyra taip pat tradicinė. Ja *remiasi*	filologija:	vienas dalykas yra tekstas ir jo prasm
23 užgesino" mama. "Vaikeli, anokia čia *specialybė* –	"filologija"!,	– tikino ji. – Juk nebūtina studiju
24 amo abstrakcijos ir universalumo lygio *struktūras*	(filologija	akcentuoja konkretybės šventumą). Galg
25 avus į kitą kraštą, kiti ir universitetai. *Taigi*	filologija,	nors to termino aš ir nelabai suprantu
26 igynė disertaciją "Latvių kalbos sėliškųjų *tarmių*	filologija".	1980–1990 metais dirbo mokslinį darbą
27 iš medicinos srities atstovų, antroje *vietoje* –	filologija,	trečioje – technikos mokslai. Susitik
28 : lietuvių filologija, anglų filologija, *vokiečių*	filologija,	vokiečių ir rusų kalbos, kurias daugia
29 slo šakomis: antropologija, biologija, *zoologija*,	filologija	(senais vietovardžiais) 104. Nors, anot

3 lentelė

Žodžio *filologija* kolokatų (pasikartojančių daugiau nei 1 kartą) sąrašas

Nr.	Kolokatas	Iš viso	K5	K4	K3	K2	K1		D1	D2	D3	D4	D5
1.	BALTŲ	5	0	0	0	0	4	-	0	0	0	1	0
2.	LIETUVIŲ	7	1	0	1	0	3	-	0	0	1	0	1
3.	ANGLŲ	4	0	0	1	0	2	-	1	0	0	0	0
4.	KAIP	4	0	0	0	0	2	-	1	0	0	1	0
5.	O	2	0	0	0	0	2	-	0	0	0	0	0
6.	VOKIEČIŲ	3	0	0	0	0	1	-	1	0	1	0	0
7.	KAD	2	0	0	0	0	1	-	0	0	0	1	0
8.	REMIASI	2	0	0	0	0	1	-	1	0	0	0	0
9.	FILOLOGIJA	6	0	1	0	2	0	-	0	2	0	1	0
10.	IR	12	2	1	2	1	0	-	2	0	0	1	3
11.	SPECIALYBĖ	3	0	0	1	1	0	-	0	0	1	0	0
12.	DISCIPLINOS	2	0	1	0	1	0	-	0	0	0	0	0
13.	LIEKA	2	0	0	0	1	0	-	1	0	0	0	0
14.	MOKSLUS	2	0	0	0	1	0	-	0	1	0	0	0
15.	PAT	2	0	1	0	1	0	-	0	0	0	0	0
16.	SPEKTRAS	2	0	1	0	1	0	-	0	0	0	0	0
17.	ATSTOVŲ	2	0	1	0	0	0	-	0	0	0	0	1
18.	IŠ	2	0	0	0	0	0	-	0	1	0	1	0
19.	LATVIŲ	2	0	1	0	0	0	-	0	0	0	0	1
20.	MEDICINOS	2	0	0	0	0	0	-	0	0	1	0	1
21.	NORS	2	0	0	0	0	0	-	1	0	1	0	0
22.	NUO	2	0	0	0	0	0	-	0	1	0	0	1
23.	SPECIALYBIŲ	2	1	0	1	0	0	-	0	0	0	0	0
24.	SRITIES	2	1	0	0	0	0	-	0	0	0	1	0
25.	STUDIJUOTI	2	1	0	0	0	0	-	0	1	0	0	0
26.	TAIP	2	1	0	0	0	0	-	0	0	1	0	0
27.	TIK	2	0	0	0	0	0	-	0	0	0	0	2
28.	TIKĖJIMU	2	0	0	0	0	0	-	0	0	1	0	1
29.	TO	2	1	0	0	0	0	-	0	1	0	0	0
30.	VIENAS	2	1	0	0	0	0	-	1	0	0	0	0
31.	YRA	2	0	0	0	0	0	-	0	1	1	0	0

pirmąją *3-os lentelės* eilutę): K1 pozicijoje šis žodis pavartotas 4 kartus. Pats dažniausias žodžio *filologija* kolokatas visose pozicijose yra jungtukas *ir*. Dirbant su didesniais nei 1000 eilučių konkordansais, tokios kolokatų išklotinės labai palengvina analizuoti konkordansus. Tiesa, konkordansų sąrašai nepatogūs tuo, kad juose labai ribota kontekstinė informacija. Iš sąrašo akivaizdu, kad žodžių junginys *baltų filologija* konkordanse pavartotas 4 kartus, tačiau neaiškūs šio junginio ryšiai su kitais žodžiais, ir tada vėl tenka grįžti prie konkordanso skaitymo.

Konkordansų kūrimas ir net gera analizavimo metodologija ne visada išsprendžia visas problemas: kai susiduriama su 10 tūkstančių eilučių ir didesniais konkordansais, tai konkordanso skaitymas ir analizavimas gali tapti tokiu pat varginančiu darbu, kaip ir konkrečių vartosenos atvejų ieškojimas knygose ar spaudoje. Todėl lingvistui labai svarbu iš anksto įvertinti savo ir naudojamos programinės įrangos galimybes. Galbūt tada vertėtų pamąstyti apie tiriamos medžiagos kiekio apribojimą arba apie tobulesnius programinius įrankius.

UNIVERSALUMO PROBLEMA

Labai svarbi tema tekstynų lingvistikoje yra kalbinės programinės įrangos bei leksinių duomenų naudojimo universalumas (*reusability*). Buvo pastebėta, kad didelė dalis sukurtos kalbinės programinės įrangos ar surinktų leksinių duomenų pritaikomi tik konkretiems pro-

jektams bei uždaviniams ir nesirūpinama jų panaudojimu kitiems tikslams. Dažnai sunku pasinaudoti panašų darbą dirbančių žmonių įdirbiu dėl skirtingų operacinių sistemų, dokumentacijos stokos, kodavimo skirtumų, kalbos specifiškumo, autorystės teisių ir pan. Dalį šių problemų gali išspręsti jau minėti elektroninių tekstų standartai (CES, SGML), kurie gali suteikti vienodą struktūrą įvairių stilių ir kalbų tekstams. Vis dėlto reikia pripažinti, kad daugelį kompiuterinių priemonių lengviau sukurti iš naujo nei bandyti prisitaikyti svetur padarytas kalbines programas.

Ar gali kalbinė programinė įranga veikti vienodai efektyviai su skirtingų kalbų tekstais arba žodžiais? Tos kompiuterinės programos, kurios žodžius traktuoja kaip abstrakcijas, kaip skirtingų ženklų seką, nesusijusią jokiais lingvistiniais ryšiais, arba tos, kurios skirtos tiktai abstrakcijų paieškai arba skaičiavimui, gali taip pat gerai veikti tiek su anglų, tiek su lietuvių, tiek ir su kitų kalbų tekstais. Tuo tarpu kalbinės programos, kurios pagrįstos konkrečios kalbos žodynais ar gramatinėmis taisyklėmis, t.y. tos, kurios žodžiams suteikia tam tikrą morfologinę, sintaksinę ar semantinę reikšmę, sunkiai pritaikomos kelioms skirtingoms kalboms.

Pvz., dauguma konkordavimo, paralelinimo programų ir dažninių sąrašų generatorių yra nepriklausomi nuo kalbos. Tiesa, vienintelė šių programų problema būna raidžių kodavimas. Jei kompiuterinė programa nebus pritaikyta dirbti su skirtingų kalbų raidžių kodavimais, tai jos panaudojimas kitoms kalboms bus labai ribotas.

Baigiant reiktų pripažinti, kad Lietuvoje dauguma lingvistų dar dirba senais, tradiciniais metodais. XXI amžiuje dirbti pasenusiais metodais – tai būti lėtam, tai neturėti visos informacijos ir galų gale beviltiškai atsilikti nuo mokslo ir technologijų raidos. Savaime suprantama, kad nereiktų besąlygiškai priimti ir naujosios metodologijos nuostatų, nes ir ji turi silpnų vietų, trūkumų, yra ribota. Su kritišku, bet atviru požiūriu naujovėms, išnaudojus naują metodų ir kalbinės įrangos pranašumus, manyčiau, galima bendromis pastangomis ženkliai prisidėti prie kalbotyros mokslo raidos.

LITERATŪRA

Altenberg B. and M. Eeg-Oloffson (1990) 'Phraseology in Spoken English: Presentation of a Project', in Aarts and Meijs 1990.

Aarts J. and Meijs W. (eds) (1990) *Theory and Practice in Corpus Linguistics*. Amsterdam: Rodopi.

Botley S. P. et al (eds) (2000) *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi.

Chanod J. P. (1997) "Current Developments for Central and Eastern European Languages", in *Proceedings of the Second TELRI Seminar on Language Applications for a Multilingual Europe*, Kaunas, Lithuania, 21–34.

Danielsson P. and Ridings, D. (1997) *Aligner "Vanilla":* <URL: http:// nl.ijs.si/ et/ project/ TELRI/ cat/ x125.htm

Erjavec T., Lawson A., and Romary L. (1997) "East meets West: a Compendium of Multilingual Language Resources" in *Proceedings of the Third TELRI Seminar on Translation Equivalence*, Montecatini Terme, Italy, 49–56.

Gellerstam M. (1995) "Lexical Resources and Their Application", in *Proceedings of the First TELRI Seminar on Language Resources for Language Technology*, Tihany, Hungary, 57–64.

Grumadienė L. ir Žilinskienė V. (1997) *Dažninis dabartinės rašomosios lietuvių kalbos žodynas (mažėjančio dažnio tvarka)*. Vilnius: Mokslo aidai.

Grumadienė L. ir Žilinskienė V. (1998) *Dažninis dabartinės rašomosios lietuvių kalbos žodynas (abėcėlės tvarka)*. Vilnius: Mokslo aidai.

Ide N. and Véronis J. (1995) "Corpus Encoding Standard". *Document MUL/EAG CES1*. <URL: http://www.lpl.univ-aix.fr/projects/multext/CES/CES1.html>.

Marcinkevičienė R. (1997) "Tekstynų lingvistika ir lietuvių kalbos tekstynas". *Lituanistica* 1, 58–78.

McEnery T. and Wilson A. (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery T. and Oakes M. (2000) "Bilingual Text Allignment: an Overview" in Botley, S. P. et al. 2000, 1–37.

Oakes M. P. (1998) *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Orwell G. (1949) *1984*. San Diego: Harcourt Brace Jovanovich Book. (First Plume Printing, April, 1983).

Orvelas Dž. (1991) *Gyvulių ūkis; 1984-ieji* (vertė A. Sabonis, V. Čepliejus). Vilnius: Vyturys.

Ramm W. (1994) "Introduction and Overview" in Ram, W. (ed.) *Text and Context in Machine Translation: Aspects of Discourse Representation and Discourse Processing, Vol. 6.* Luxembourg: Office for Official Publications of the European Communities.

Scott M. (1996) *WordSmith Tools Manual*. Oxford: Oxford University Press; <URL: http://www.liv.ac.uk/~ms2928/homepage.html>

Sperberg-McQueen C. M. and Burnard L. (1994) "Guidelines for Electronic Text Encoding and Interchange (P3)". Chicago: Text Encoding Initiative.

Summers D. (1996) "Computer Lexicography: the Importance of Representativeness in relation to frequency" in Thomas, J. and Short M. 1996.

Swanepoel P. (1994) "Problems, Theories and Methodologies in Current Lexicographic Semantic Research" in Martin, W. et al (eds.), *Euralex 1994 Proceedings.* Amsterdam, 11–26.

Teubert W. (1997) "Translation and the Corpus" in *Proceedings of the Second TELRI Seminar on Language Applications for a Multilingual Europe*, Kaunas, Lithuania, 147–164.

Thomas J. and Short M. (eds) (1996) *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. London: Longman.

1 PRIEDAS. PAPRASTO TEKSTINIO FORMATO IR SGML FORMATO TEKSTAS

Dialogo dalyviai:
SOKRATAS, GLAUKONAS, POLEMARCHAS,
TRASIMACHAS, ADIMANTAS, KEFALAS

PIRMOJI KNYGA

I

[Sokratas]. Vakar su Aristono sūnumi Glaukonu buvau nuėjęs į Pirėją. Norėjau pasimelsti deivei, be to, pasižiūrėti, kaip jie švęs tą šventę – juk ji švenčiama pirmą kartą. Vietinių gyventojų eitynės atrodė labai gražiai, bet ne mažiau iškilmingos buvo ir trakų eitynės. Pasimeldę ir pasižiūrėję iškilmių, ėjome atgal į miestą. Iš tolo pamatęs mus einant namo, Kefalo sūnus Polemarchas paliepė tarnui mus pasivyti ir paprašyti, kad jo palauktume. Pribėgęs tarnas truktelėjo mane iš užpakalio už apsiausto ir tarė:

– –

```
<p>Dialogo dalyviai:<lb>
SOKRATAS, GLAUKONAS, POLEMARCHAS,<lb>
TRASIMACHAS, ADIMANTAS, KEFALAS<lb>
</p>
<div type="book">
<head>PIRMOJI KNYGA</head>
<div type="section">
<head>I.</head>
<p><seg>[Sokratas]. Vakar su Aristono s&umacr;numi Glaukonu buvau nu&edot;j&eogon;s
&iogon; Pir&edot;j&aogon;.</seg> <seg>Nor&edot;jau pasimelsti deivei, be to, pasi&zca-
ron;i&umacr;r&edot;ti, kaip jie &scaron;v&eogon;s t&aogon; &scaron;vent&eogon; &mdash; juk
ji &scaron;ven&ccaron;iama pirm&aogon; kart&aogon;.</seg> <seg>Vietini&uogon; gyven-
toj&uogon; eityn&edot;s atrod&edot; labai gra&zcaron;iai, bet ne ma&zcaron;iau i&scaron;kil-
mingos buvo ir trak&uogon; eityn&edot;s.</seg> <seg>Pasimeld&eogon; ir pasi&zca-
ron;i&umacr;r&edot;j&eogon; i&scaron;kilmi&uogon;, &edot;jome atgal &iogon;
miest&aogon;.</seg> <seg>I&scaron; tolo pamat&eogon;s mus einant namo, Kefalo s&umacr;nus
Polemarchas paliep&edot; tarnui mus pasivyti ir papra&scaron;yti, kad jo palauktume.</seg>
<seg>Prib&edot;g&eogon;s tarnas truktel&edot;jo mane i&scaron; u&zcaron;pakalio u&zcaron;
apsiausto ir tar&edot;:</seg></p>
```

2 PRIEDAS. SUPARALELINTI SAKINIAI, PAIMTI IŠ DŽORDŽO ORVELO KNYGOS *1984-IEJI*

<p><s>PIRMA DALIS 1

<p><s>Part 1, Chapter 1

<p><s>Buvo šviesi ir šalta balandžio diena, laikrodžiai mušė tryliktą valandą,

<p><s>It was a bright cold day in April, and the clocks were striking thirteen.

<s>Vinstonas Smitas, įtraukęs smakrą užantin ir gindamasis nuo smarkaus vėjo, greitai šmurkštelėjo pro stiklines Pergalės rūmų duris, bet drauge su juo vidun vis dėlto spėjo plūstelėti verpetas aštrių dulkių.

<s>Winston Smith, his chin nuzzled into his breast in an effort to escape the vile wind, slipped quickly through the glass doors of Victory Mansions, though not quickly enough to prevent a swirl of gritty dust from entering along with him.

<p><s>Koridoriuje kvepėjo virtais kopūstais ir senais kilimais.

<p><s>The hallway smelt of boiled cabbage and old rag mats.

<s>Gale ant sienos kabėjo spalvotas, patalpoms pernelyg didelis plakatas.

<s>At one end of it a coloured poster, too large for indoor display, had been tacked to the wall.

Gauta 2000 11 30
Parengta 2000 12 07

Andrius UTKA

LINGUISTIC SOFTWARE IN CORPUS LINGUISTICS

Abstract

The article deals with linguistic software (lingware) which is used in corpus linguistics. The types of the lingware are enumerated as well as their functions, capabilities, and availability for the Lithuanian language research. Most of the software tools such as converters, word list generators, concordance programs, lemmatizers, morphological analysers are extensively used and developed at the Centre of Computational Linguistics at Vytautas Magnus University in Kaunas. The article also touches upon the problem of universality of lingware. A special attention is paid to the central software in corpus linguistics, that is concordancers and word list generators. The article shows that these computer programs, if used properly would enable a linguist to better cope with large amounts of data that is taken from multi-million word electronic corpora.