Tekstynų lingvistika

Tekstyno sąvoka

Tekstynu vadinsime pakankamai didelį tekstų rinkinį, sudarytą nepriklausomai nuo jo panaudojimo tikslų, t. y. jis turi kuo geriau atspindėti kalbą ar jos atmainą.

Kuo tekstynas panašus į tekstą ir kuo skiriasi. Panašumas: tekstynas sudarytas iš tekstų. Skirtumai: Tekstą prasminga analizuoti visą ištisai, jis turi pradžią, vidurį, pabaigą, yra daugiau ar mažiau rišlus ir vientisas. Tekstynas neturi struktūros, tik sandarą. Tekstyno neverta ištisai tiesiogiai skaityti ir taip nuodugniai nagrinėti kaip tekstą, jį verta skaityti tik su programinėmis priemonėmis ar kitais įrankiais. Taigi esminis skirtumas tarp teksto ir tekstyno – tyrimo metodologija.

Tekstynų tipai

Tekstynai gali būti skirstomi pagal įvairius požymius. Jie gali būti:

Dideli ir **maži**. Kai kurie tekstynai yra maži iš prigimties, pvz., senųjų raštų, mirusių kalbų rašto paminklų, atskirų autorių (V. Šekspyro) ar net atskirų kūrinių (Šventojo Rašto), jų jau nepadidinsi. Tekstyno mažumas yra trūkumas, nes nuo dydžio priklauso tekstyno kokybė. Vieno milijono žodžių apimties tekstyne nuo 40 iki 50% žodžio formų pasirodo tik vieną kartą, o polisemiškų žodžių yra pavartota tik pusė reikšmių. Tuo tarpu bet kuriam kalbos vienetui aprašyti reikia mažiausiai dviejų jo pavartojimo atvejų. Dar daugiau pavartojimų reikia tiriant žodžių kontekstą ir pan. Dideliame tekstyne išryškėja dėsningumai, kurių nematyti mažame. Didelių tekstynų trūkumas, kad reikia galingesnių kompiuterių, geresnių programų, sugaištama daugiau laiko (pvz., 1000 eilučių konkordansas yra riba, kurią peržengus tampa sunku su juo dirbti). Problemos sprendimas: kurti įvairių dydžių tekstynus skirtingiems uždaviniams, rūpintis tekstyno reprezentatyvumu, tobulinti naudojamas programas.

Baigtiniai (statiniai) ir **tęstiniai** (dinaminiai). Baigtiniai tekstynai parodo kalbos vaizdą tam tikru metu. Jie pravartūs, kai lyginami su panašiais, tik kitu laiku ar kitai kalbai (dialektui) sukurtais tekstynais. Tęstiniuose tekstynuose galima įtaisyti filtrus, atrenkančius naujus kalbos faktus. Tęstiniai tekstynai paprastai sudaromi iš ištisų tekstų, o ne tekstinių imčių, todėl būna nesubalansuoti, tačiau jų dydis atstoja nesubalansuotumą.

Oportunistiniai ir **balansuoti-reprezentatyvūs**. Nuo pat tekstynų atsiradimo buvo stengiamasi rengti subalansuotus tekstynus, sudarytus iš įvairių šaltinių pagal aiškius kriterijus. Tik vėliau atsirado tekstynai, į kuriuos dedama visi gaunami tekstai (oportunistiniai tekstynai). Balansas pasiekiamas nustatant įvairių šaltinių proporcijas pagal tam tikrus kriterijus. Galimi kriterijai: elitiškumas (vertingiausi kūriniai), skaitomumas (bestseleriai, publicistika), demografiniai rodikliai (kuo įvairesni autoriai), prieinamumas ir t. t. Kalbant apie reprezentatyvumą kyla klausimas, ką turi atspindėti tekstynas. Pasakymas, kad jis turi atspindėti kalbą ar jos įvairovę, yra neinformatyvus, todėl patogu išskaidyti į keturias vartojimo sritis: kalbėjimą, rašymą, klausymą ir skaitymą ir atsižvelgti į kiekvienos srities vartotojų skaičių. Oportunistiniai ir reprezentatyvūs tekstynai dažnai susiję kaip du skirtingi tekstyno sudarymo etapai: iš pradžių sudaromas oportunistinis tekstynas, o po to pagal tam tikras proporcijas atrenkami tekstai balansuotam tekstynui.

Anotuoti bei koduoti ir **neanotuoti**. Teksto žymėjimas atsirado tada, kai senieji kompiuteriai nesugebėjo apdoroti teksto kaip teksto, o tik jame esančias žymes. Dabar

žymės dažnai naudojamos iš įpročio. Pagrindinis žymėto teksto trūkumas, kad kompiuteris dirba tik su žymėmis ir ignoruoja pačią kalbą, tai trukdo atsirasti naujoms hipotezėms, į tekstą žiūrima per tam tikros, dažnai iki tekstynų atsiradimo sukurtos teorijos rėmus. Žinoma, pažymos gali ir palengvinti paiešką ir bet kokią lingvistinę tekstyno analizę.

Dėl visų anksčiau paminėtų tekstynų tipų iki šiol vyksta ginčai, kurie iš jų geresni ir tinkamesni. Tekstynų tipai, dėl kurių tinkamumo nediskutuojama yra tokie:

Bendrieji ir **specialieji**. Bendrieji tekstynai skirti daugialypei analizei, juos sunku parengti, nes neaiškus adresatas. Specialieji tekstynai gali būti skirstomi pagal tikslus (pvz., kalbos mokymo tekstynai) arba pagal juose esančių tekstų pobūdį (pvz., vaikų kalbos tekstynai).

Sakytinės ir **rašytinės** kalbos tekstynai. Sakytinės kalbos tekstynai dažnai turi ir pirminį garsinį pavidalą. Dauguma tekstynų yra rašytinės kalbos.

Senųjų raštų ir **dabartinės kalbos** tekstynai.

Vienos ir **kelių kalbų** tekstynai. Paralelius tekstynus sudaro originalo ir vertimo tekstai, sulygiuoti papastraipiui ar pasakiniui. Palyginamieji tekstynai, sudaryti iš originalių dviejų ar daugiau kalbų tekstų, sudarytų pagal panašius kriterijus: pagal panašų turinį, panašias temas ir t. t.

Programinė įranga darbui su tekstynais

Pagal paskirtį programinę įrangą galima suskirstyti į tokias grupes:

Konverteriai – keičia raidžių kodavimus, pvz., iš KBL į Baltic Rim, keičia failų tipą iš HTML ar SGML į tekstinius ir atvirkščiai ir pan.

Dažninių sąrašų generatoriai.

Konkordavimo programos – apie šias dvi bus kalbama toliau.

Sintaksinės analizės programos (angl. parsers) – identifikuoja žodžius sakinyje, nustato jų sintaksinį priklausymą, sugrupuoja į aukštesnio lygio vienetus (žodžių junginius, prijungiamuosius sakinius) ir juos atitinkamai pavadina. Sintaksinė analizė gali remtis tikimybiniais modeliais arba taisyklėmis.

Klaidų tikrintuvai.

Skiemenuokliai – apie šiuos du bus kalbama kitose paskaitose.

Lemuokliai – įvairias vieno žodžio formas sujungia į vieną antraštinę – lemą. Sulemuotas žodžių sąrašas leidžia lingvistui skirtingas žodžių formas traktuoti kaip vieną. Lietuvių kalbos lemuoklį yra sukūręs V. Zinkevičius.

Morfologiniai analizatoriai – nustato žodžio gramatines charakteristikas. Paprastai nagrinėja atskirus žodžius, tačiau, jei yra daugiareikšmiškumas, gali nagrinėti ir kontekstą. Pvz., žodis “galvos” gali turėti tris gramatines interpretacijas: daiktavardžio vienaskaitos kilmininkas, daiktavardžio daugiskaitos vardininkas ir veiksmažodžio būsimasis laikas.

Anotatoriai – prie žodžių ar kitų tekstinių vienetų prirašo žymes, aiškinančias jų bruožus. Paprastai jie panašūs į morfologinius analizatorius.

Paralelinimo programos – skirtos vertimo ir originalo kalbų tekstams lygiagretinti. Paprastai paralelinami sakiniai, bet galima ir žodžių lygmenyje.

Mašininio vertimo sistemos – apie jas bus kalbama kitoje paskaitoje.

Dažniniai žodžių sąrašai

Tai paprasčiausias ir istoriškai pirmasis statistinio tekstų apdorojimo rezultatas. Šie sąrašai gali būti dviejų tipų: žodžių (nekaitomų žodžių ir kaitomų žodžių formų) ir lemų (antraštinių žodžių) sąrašai. Žodžių sąrašai gaunami sujungus ir tam tikra tvarka išrikiavus nekaitomus žodžius ir identiškas kaitomų žodžių formas. Lemų sąrašai generuojami iš morfologiškai anotuoto tekstyno arba lemuojant žodžių sąrašą. Morfologiškai anotuoti tekstyną užima daug laiko, o lemuojant žodžių sąrašą susiduriama su homonimais (pvz., žodis *laužo* kilęs iš *laužas* ar *laužyti*). Žodžių sąraše išsaugoma informacija apie atskirų to paties žodžio morfologinių formų dažnį, o lemų sąrašuose ši informacija dingsta, tačiau lemų sąrašai yra trumpesni.

Sąrašo elementai gali būti surūšiuojami dažnio mažėjimo ar didėjimo, abėcėlės arba pirmojo pasirodymo tekste tvarka. Pagal dažnį surūšiuotame sąraše išryškėja dažniausi žodžiai ir jų formos. Lentelėje pateikta 50 dažniausių lietuvių kalbos žodžių sutiktų 60 mln. žodžių tekstyne:

Eil. Nr.	Žodis	Dažnis (tūkstančiais žodžių)	Eil. Nr.	Žodis	Dažnis (tūkstančiais žodžių)
1	ir	1909	26	dar	138
2	kad	485	27	po	130
3	į	455	28	už	125
4	iš	384	29	per	124
5	su	334	30	dėl	122
6	o	307	31	bei	120
7	buvo	307	32	tačiau	117
8	tai	296	33	kas	113
9	kaip	290	34	jos	112
10	yra	272	35	a	108
11	tik	233	36	to	102
12	ar	225	37	metų	99
13	ne	216	38	labai	97
14	Lietuvos	216	39	gali	94
13	savo	211	40	mūsų	94
16	bet	207	41	būti	93
17	jis	178	42	turi	92
18	apie	178	43	d	91
19	m	172	44	arba	90
20	nuo	163	45	jie	90
21	taip	162	46	prie	89
22	jo	151	47	iki	89
23	kai	149	48	ji	89
24	jų	147	49	pat	88
25	jau	142	50	nors	83

Pagal abėcėlę surūšiuotame sąraše greta atsiduria giminiški žodžiai. Be to, lengviau rasti konkretų žodį, jei sąrašas spausdintas. Žodis *filologija* pagal abėcėlę surūšiuotame sąraše atrodo taip:

Žodis	Dažnis
filologija	26
filologiją	43
filologijai	8
filologijas	1
filologijoje	10
filologijomis	1
filologijos	367

Iš 60 mln. žodžių tekstyno buvo gautas 1,2 mln. žodžių ilgio dažninis žodžių sąrašas. Santykis tarp viso tekstyno žodžių kiekio ir skirtingų žodžių kiekio rodo tekstyno žodingumą. Paprastai šis rodiklis taikomas atskiriems autoriams. Dažniniame sąraše žodžiai ir žodžių formos pasiskirsto labai netolygiai. Žr. lentelę:

Žodžių dažnumas	Sąrašo dalis	Tekstyno dalis
>10000	0,04%	37,36%
1000-9999	0,47%	27,33%
100-999	3,25%	21,24%
10-99	9,45%	9,94%
2-9	36,08%	3,09%
1	45,54%	1,04%

Prasminių žodžių sąrašas, tai tam tikram tekstui sudarytas žodžių dažnių sąrašas, iš kurio atrinkti tie žodžiai, kurie sutinkami dažniau, nei visame tekstyne.

Konkordansas

Konkordansas yra sąrašas eilučių, kuriose buvo rastas tiriamasis žodis ar žodžių junginys, paimtas iš teksto ar tekstyno. Konkordansas gali būti naudojamas žodžio junginiams ar kontekstui tirti. Pvz.:

```
tafizika turi virsti grynąja matematika, o metafizinis - individuali
s gamtotyros metodą. Grynoji matematika irgi yra grynojo santykio pa
ai kokie paradoksai! Grynoji matematika suartėja su poezija ir misti
odis yra "santykis". Grynoji matematika išreiškia santykius, tačiau
, kad mokslas kaip taikomoji matematika išreiškia kiekybinį esinijos
žodžio prasme, yra taikomoji matematika. Todėl ir sakoma, kad moksli
e visi kultūros komponentai. Matematika, fizika, sakysime, yra speci
mą. Dabartinis mokslas, ypač matematika ir fizika, ėmė tyrinėti prob
sitete. Kad domėjosi fizika, matematika ir gamtos mokslais, parašė p
inieriniai dalykai (aukštoji matematika, inžinerinė grafika, fizika,
lavinant protą (specialybė - matematika, fizika, chemija ir pan.), j
disciplinų (fizika, chemija, matematika, kalbos, informatika ir pan.
```

Lietuvių kalbos tekstynai

Lietuvių kalbos tekstynų radimosi pradžia galima laikyti L. Grumadienės ir V. Žilinskienės parengtą 1,2 mln. žodžių tekstyną, kuris buvo lemuotas ir jo pagrindu išleisti keli žodynai ([3], [4]).

Viešu ir visiems prieinamu bendrojo pobūdžio lietuvių kalbos tekstynu laikytinas VDU dabartinės lietuvių kalbos tekstynas. Jis sumanytas kaip neanotuotas,

nekoduotas, daugiausiai ištisų periodikos ir knygų tekstų, didelės temų ir kitokios įvairovės autentiškos lietuvių kalbos tekstynas. Apimtis 2002 m. viršijo 100 mln. žodžių, bet ir toliau didinamas. Juo galima naudotis visu arba dalimis (periodika, ne periodika). Jo pagrindu numatoma sukurti kelių dešimčių ir kelių milijonų žodžių tekstynus. Šimtinis ir toliau liks nekoduotas, dešimtinis bus koduotas pagal SGML standartą vartotojui pateikiant teksto bibliografiją ir struktūrinius vienetus (pavadinimus, pastraipų ir sakinių ribas), mažasis tekstynas bus morfologiškai anotuotas. Minėtą tekstyną galima rasti internete [5]. Jame galima atlikti žodžio ar jo dalies paiešką, sužinoti žodžio dažnį kiekvienoje tekstyno dalyje ir dažniausius žodžio junginius, gauti pasirinkto pločio (30, 70, 150 arba 300 simbolių) konkordansą.

Teorinės tekstynų lingvistikos nuostatos

Iki šiol tebesiginčijama, ar tekstynų lingvistika yra atskira lingvistikos šaka, ar tik metodologija. Tekstynų lingvistika neturi tokio aiškiai apibrėžto išskirtinio analizės objekto, kokį turi kitos lingvistikos šakos. Tačiau tekstynų lingvistika toli peržengia metodologijos ribas. Toliau pateiktos teorinės tekstynų lingvistikos nuostatos leidžia visai naujai pažvelgti į kelis tūkstantmečius tyrinėtą kalbą.

Kalbos fraziškumas. Egzistuoja aibė daugiau ar mažiau sustabarėjusių frazių (pvz., *gultis į ligoninę*, *sėsti į kalėjimą*), kurios skiriasi apimtimi, frazių ribos neryškios, įvairovė didesnė frazių kraštuose, gali kisti žodžių tvarka, tačiau dažniausiai vyrauja viena kuri elementų forma, gana fiksuotas kontekstas. Apie 80% teksto kuriama parenkant žodžių junginius.

Išplėstinis reikšmės vienetas. Frazės prasmė nėra žodžių reikšmių suma.

Desemantizacija – kai žodžiai netenka savo pirminės prasmės, o naudojami tik junginiuose. Pvz., žodis *prasmė* dažniausiai naudojamas junginiuose *ta prasme*, *šia prasme* ir pan., kur jis neturi jokios prasmės, o yra tik delšžodis. Iš čia seka dar vienas įdomus rezultatas, kad vadovaujantis kalbos jausmu mes dažnai manom, kad žodis turi vienokią reikšmę, o remiantis tekstynų tyrinėjimais – kitokią. Pvz., *aistra* asocijuojasi ne su meile, o su politika, *klausimas* ne su paklausimu, o su reikalu.

Leksikos ir semantikos vienovė. Anksčiau buvo manoma, kad iš pradžių sukuriama sintaksinė sakinio struktūra, kuri užpildoma leksiniais vienetais. Leksikos ir gramatikos atskyrimas jas nagrinėti kaip visumą. Kad reikia jungti, rodo kalbos dalių neapibrėžtumas (ar *daugiabutis* daiktavardis ar būdvardis), koks pasiskirstymas tarp lemos reikšmės ir jos formų reikšmių, nes lema būna ir visai nevartojamas žodis, pvz., *omenyje*.

Literatūra

1. Marcinkevičienė, R. (2000). *Tekstynų lingvistika: Teorija ir praktika.* VDU leidykla. Kaunas. *arba* Marcinkevičienė, R. (2000). Tekstynų lingvistika: Teorija ir praktika. *Darbai ir dienos*, 2000.24, VDU leidykla, Kaunas, p. p. 7-64.

2. Utka, A. (2000). Kalbinė įranga ir jos galimybės. *Darbai ir dienos*, 2000.24, VDU leidykla. Kaunas, p. p. 275-285.

3. Žilinskienė, V. (1995). *Atgalinis dabartinės lietuvių kalbos žodynas*. Vilnius, Matematikos ir informatikos institutas.

4. Grumadienė, L., V. Žilinskienė (1997), *Dažninis dabartinės rašomosios lietuvių kalbos žodynas (mažėjančio dažnio tvarka)*. Vilnius.

5. http://donelaitis.vdu.lt/.