Diskurso ryšiai ir tekstų anotavimas

Doc. dr. Giedrė Valūnaitė Oleškevičienė

Mokslinis tyrimas vykdomas pagal podoktorantūros projektą Nr. 09.3.3-LMT-K-712-02-0144 „Diskurso ryšiais anotuoto tekstyno, grįsto socialinių medijų tekstais, naudojimas aukštesnio lygio užsienio kalbos mokymui(si)“, finansuojamas Europos socialinio fondo lėšomis pagal priemonę Nr. 09.3.3-LMT-K-712 „Mokslininkų, kitų tyrėjų, studentų mokslinės kompetencijos ugdymas per praktinę mokslinę veiklą“.

MYKOLO ROMERIO UNIVERSITETAS

Kuriame Lietuvos ateitį

2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Kas tai yra tekstynas

- Bet koks, net ir pats mažiausias, elektroninę formą turinčių tekstų rinkinys yra tekstynas.
- Mokslo kalba: Tai - toks tekstų rinkinys, kuris yra pakankamai didelis, matuojant pagal šių dienų kompiuterinių technologijų galimybes, ir sudarytas ne dėl kokio specialaus tyrimo, bet nepriklausomai nuo jo panaudojimo tikslų (Sinclair 1999).

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Tekstynas

- Tekstynas sudarytas iš tekstų
- Tekstynų analizės priemonės, įvairios programos ir įrankiai yra neatskiriama tekstynų dalis

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Tekstynų tipai

- Bendrojo pobūdžio ir specialieji
- Sakytinės ir rašytinės kalbos
- Vienakalbiai ir paraleliniai (daug kalbų)
- Senųjų raštų ir dabartinės kalbos tekstynai

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Tekstynų dydis

- Pirmieji tekstynai siekė 1 milijoną žodžių apimtį
- Dabartiniai siekia pusę bilijono
- Svarbus reprezentatyvumas
- Priklauso nuo tyrimų tikslų

Lietuviški tekstynai

- Du pagrindiniai tekstynai:
- Akademinės kalbos tekstynas (Corpus Academicum Lithuanicum – CorALit http://coralit.lt, talpina apie 9 milijonus žodžių) sudarytas VU. Apima akademinius tekstus biomedicinos srityje, humanitarinių mokslų, fizinių mokslų, socialinių mokslų ir technologijų srityse.
- Šiuolaikinės lietuvių kalbos tekstynas (http://tekstynas.vdu.lt), talpina 102 milijonų žodžių ir apima publicistinius tekstus, grožinės ir ne grožinės literatūros tekstus, sakytinės kalbos tekstus.
- Tačiau lygiagretieji tekstynai apima tik anglų – lietuvių(70,813 lygiagrečių sakinių) ir lietuvių –anglų (1,614 lygiagrečių sakinių)

Tekstynų analizės priemonės

- Tekstų lemavimo, anotavimo ir sintaksinės analizės programos
- Įvairios statistinės priemonės - sąrašų bei konkordansų generavimo, tekstų paralelinimo, kolokacijų paieškos ir kt. (Utka, 2000)

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Tekstynų anotavimas

- Tai - teksto kodavimas susijęs su struktūriniais jo elementais
- Kontraversiškas mokslininkų požiūris
- Kritika:

➢Apima ir interpretacijos atvejus (pažymos vietą ir pobūdį lemia subjektyvi tyrėjo nuomonė).

➢Interpretacijos esama ir gramatinėse anotacijose, nes jos remiasi tradicine, daugeliu atveju subjektyvia nuomone ir susitarimu paremta gramatika (Sinclair, 2000).

➢Kompiuteris dirba su žymomis ir ignoruoja pačią kalbą.

Tekstynų anotavimas

Palaikantis požiūris:

- Tinka kurti naujoms hipotezėms, išryškėjančioms iš darbo su dideliais tekstų masyvais.
- Anotacijos labai palengvina paiešką ir bet kokią lingvistinę tekstyno analizę.

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Tekstynų lingvistikos ryšys su kitais mokslais

- Leksikografija (žodynų rengimas nuo žodžių atrankos iki jų aprašo)
- Leksikologija (leksinių vienetų analizė)
- Kontrastyvinė lingvistika
- Vertimo teorija
- Sociolingvistika (lyčių skirtumai vartojant kalbą)
- Stilistika ir literatūrinė analizė
- Teksto ir diskurso analizė
- Užsienio kalbos mokymo teorija

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Tekstynų lingvistikos privalumai mokant užsienio kalbos

- Naudodami tekstynus užsienio kalbų mokyme(si) tiek besimokantieji tiek mokytojai gali identifikuoti tam tikras taisykles ir išimtis remdamiesi tekstynų duomenimis.
- Kitas privalumas, kad tekstynai suteikia realią informaciją apie realų kalbos vartojimą tam tikruose kontekstuose (Aston, 2001).

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Konkordansas

- Konkordansas – tai sąrašas eilučių, kuriose rastas tiriamas žodis ar žodžių junginys (Utka, 2000).
- Žodžio konkordansai – svarbiausia iš tekstyno gauta informacija.
- Pavartojimo kontekstas, kurį rodo konkordansas, padeda nustatyti tikslią žodžio reikšmę.
- Atskiri turi fiksuotą reikšmę, tačiau ją pakeičia tam tikromis vartojimo aplinkybėmis.

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Konkordanso vaizdas

dictionary like we all do and the first	meaning	in the dictionary is all to do with #
a proper Italian pronunciation for that	meaning	leaf or sheet and so really in its basi
heet and so really in its basic #	meaning	portfolio means a case for carrying loo
ussed before she'd also assessed orally	meaning	they want to check if her teeth are oka
ather different it has a very different	meaning	and that is that you can't override
tes following a step input oscillates	meaning	it it goes backwards and forwards and that
o push out the residual demand curve	meaning	it would comm not only command a
e static properties of the equilibrium	meaning	we shifted one of the curves and look
kets i'm going to use the word bundles	meaning	a collection of or a combination of tw
y're falling apart i think what you're	meaning	is that you've got an i # an interfacia
and it fits the syntax and it fits the	meaning	that's it i'll # i'll go for that hy
defined concepts which we all know the	meaning	of but don't need or not even able to d
pecies and i take you understand the	meaning	of those numbers here minus-two m
index and a little label at the top N	meaning	the time index so that's the times th
en out # and that's the same same	meaning	the reason the word's choosed in this
holism bec # for our intentionality	meaning	our human concepts our human understar
e that it has managed to bleed all the	meaning	out of maternity we don't have matria
but you under what you understand the	meaning	of what i'm saying here no i don't
but in economics it has a very specific	meaning	input-output models were invented by

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Žodžių vartojimo dažnumas

- Juo remiantis galima pasirinkti ko mokyti(s) (Aston, 2001).
- Pavyzdžiui, kokius esminius žodžius įtraukti į mokymo(si) programą.

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Kolokacija

- Žodis su reikšminiais junglumo partneriais (žodžiai vartojami kartu tam tikrose frazėse).
- Pavyzdžiui, *shed* kolokacijos su žodžiais *light, tears, jobs, blood, image, pounds, staff, skin, and clothes.*
- Dažniausiai skirtingos kolokacijos turi skirtingas reikšmes: *shed blood* reiškia kentėti*, shed pounds* reiškia numesti svorio*, shed image* reiškia pakeisti žmonių nuomonę apie save.

Kolokacijos taikymo pavyzdys

- “I will open the air-conditioner,” – neteisinga
- Kolokacija – “switch on air-conditioner”, bet “open the gate, the window, etc.”

Koligacija

- Žodis su nereikšminiais junglumo partneriais (jungtukais, dalelytė-mis)
- *head* turi koligacijas su*: of, over, on, back, and off.* Koligacijos taip pat keičia žodžio reikšmę.
- Pavyzdžiai: *head of department, hit someone over the head, throw one's head back.*

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Koligacijos taikymo pavyzdys

- The building is adjacent . . . the train station. (to) prijungtas
- It is usually a good idea to abide . . . the law. (by) laikytis įstatymo
- You should give clear indication . . . your intentions. (of) savo ketinimų

Semantinis pirmumas

- Žodžio galėjimas jungtis su tam tikros leksinės aplinkos žodžiais.
- Tai susiję su semantinio lauko teorija.
- Žodžio vartojimas gali būti susijęs su tam tikromis aplinkomis ar situacijomis, kaip pavyzdžiui, “at the post office,” “at the airport,” “in the supermarket,” “in the office”.
- Saudo Arabija (forma Saudo Arabijoje) linkęs jungtis su būsenos veiksmažodžiais, reiškiančiais mirtį (žūti, mirti). Tai susiję su dažniausiu karo, teroro kontekstu (Kamandulytė, 2006).

Semantinė prozodija

- numanomas žodžio kontekstas, konotacija (Sinclair, 2000)
- *cause* dažniausia jungiasi (kolokuoja) su negatyvią reikšmę nešančiais žodžiais *accident, concern, damage, death, trouble* – tokiu būdu jis turi negatyvią semantinę prozodiją.
- p*rovide,* dažniausia jungiasi (kolokuoja) su pozityvią reikšmę nešančiais žodžiais *aid, care, food, opportunities, relief, support* – tokiu būdu jis turi pozityvią semantinę prozodiją.

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Registras ir žanras

- Pavyzdžiui, 12 dažniausiai vartojamų veiksmažodžių (*say, get, go, know, think, see, make, come, take, want, give, mean)* tekstyne, apimančiame pokalbio, grožinės literatūros, laikraščių ir akademinį žanrus, pasiskirsto labai skirtingai šiuose žanruose.
- 45 procentai visų veiksmažodžių vartojami pokalbių žanre ir tik 11 procentų akademiniame žanre.

Registro ir žanro taikymas

- Biber ir Conrad (2001) teigia kad šiems veiksmažodžiams turėtų būti teikiama pirmenybė mokant anglų kalbos kaip užsienio kalbos pradiniuose lygiuose.

Diskursas

- Tai išplėstas minties apie kurį nors dalyką išreiškimas; kalbinis arba nekalbinis procesas, turintis vienokią ar kitokią prasmę.

Diskurso žymikliai

- Diskurso žymikliai patraukė tyrėjų dėmesį kaip priemonės naudojamos diskurso jungimo ir valdymo užtikrinimui.

Diskurso žymikliai - apibrėžimas

- Tai leksiniai vienetai arba gramatinės formos naudojamos susieti atskiras diskurso dalis (Halliday, 1992).
- Teksto siejimo žodžiai ir požiūrio raiškos žodžiai.

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Diskurso žymikliai

Apima:

- paprastus jungtukus (bet, ir, ar);
- sudėtinius jungtukus (kita vertus);
- kitas jungiamąsias frazes (visų pirma)
- ir požiūrio raiškos žodžius (žinoma, savaime suprantama, tiesą sakant).

Terminas

- Mokslininkų siūlomi terminai
- 'Particles‘ (dalelytės) Svartvik (1980);
- 'connectives' (jungtukai) Chalker (1984);
- 'discourse particles' (diskurso dalelytės) Schourup(1985),
- ‘discourse connectives‘ (diskurso jungtukai) Prasad, et al (2007) pagal **PDTB anotavimo sistemą**.

Diskurso struktūra

- Pagal Webber and Joshi (1998) diskurso jungtukai jungia du abstrakčius objektus tokius kaip įvykiai, būsenos ar teiginiai (Asher, 1993) kaip jų jungiamus argumentus.
- Pavyzdys: The federal government suspended sales of U.S. savings bonds because **Congress hasn't lifted the ceiling on government debt**.

Diskurso jungtukai

- Išreikšti – prijungiamieji (nes), sujungiamieji (ir) ir prieveiksmiai (pavyzdžiui).
- Numanomi - Projects already under construction will increase Las Vegas’s supply of hotel rooms by 11,795, or nearly 20%, to 75,500. Implicit = so (taigi) By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs.

Diskurso jungtukų prasmių hierarchija

Diskurso ryšiai									
Laikas		Kontingencija		Palyginimas		Plėtinys			
Sinchroniškas	Asinchroniškas	Priežastingumas	Sąlyga	Kontrastas	Nuolaida	Sujungimas	Paaiškinimas	Pavyzdys	Kita

Keturios pagrindinės prasmių klasės

- Laikas
- Kontingencija
- Palyginimas
- Plėtinys

Laikas

- But a Soviet bank here would be crippled unless Moscow found a way to settle the $188 million debt, which was lent to the country's short-lived democratic Kerensky government before **the Communists seized power in 1917**. (asinchroniškas)

Kontingencija

- In addition, its machines are typically easier to operate, so **customers require less assistance from software**. (pasekmė)

Palyginimas

- Operating revenue rose 69% to A$8.48 billion from A$5.01 billion. <u>But</u> **the net interest bill jumped 85% to A$686.7 million from A$371.1 million**. (kontrastas)

Plėtinys

- A Lorillard spokewoman said, “This is an old story. In fact, **we’re talking about years ago before anyone heard of asbestos having any questionable properties**.” (paaiškinimas)

PDTB anotavimo schema

Ar yra diskurso ryšys tarp dviejų diskurso segmentų?				
TAIP			NE	
Ar yra leksiškai išreikštas diskurso ryšys tekste?			Ar yra sąsajos tarp diskurso segmentų?	
TAIP		NE	TAIP	NE
Ar tai yra jungtukas?	Ar tai alternatyvi leksikalizacija?			
⇩	⇩	⇩	⇩	⇩
Explicit	**Altlex**	**Implicit**	**EntRel**	**NoRel**

Alternatyvi leksikalizacija

- After trading at an average discount of more than 20% in late 1987 and part of last year, country funds currently trade at an average premium of 6%. **AltLex** [**The reason**:] **Share prices of many of these funds this year have climbed much more sharply than the foreign stocks they hold**.

Vientisumo ryšys

- Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29. **EntRel Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group**.

Nėra ryšio

- Mr. Rapanelli met in August with U.S. Assistant Treasury Secretary David Mulford. **NoRel Argentine negotiator Carlos Carballo was in Washington and New York this week to meet with banks**.

Vertimo analizė (I)

- Kai kurie numanomi originalaus teksto jungtukai yra verčiami išreikštais jungtukais.
- *that's okay, right <u>but</u>* ***We want more*** (annotated **implicit** Comparison: Concession: Arg2_as_denier)
- *Nebogai, tiesa [Bet]* ***mes norim daugiau*****.** (annotated **explicit** Comparison: Concession: Arg2_as_denier)
-

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Vertimo analizė (II)

- Kai išreikšti jungtukai originaliame tekste neišverčiami, galimas prasmės praradimas
- *only looking at race doesn't really contribute to our development of diversity. [So]* ***if we're trying to use diversity as a way to tackle some of our more intractable problems, we need to start to think about diversity in a new way****.* (annotated **explicit** Contingency: Cause: Result)
- *žiūrėti tik į rasę nepadeda bandant prisidėti prie įvairumo vystymo. [taigi]* ***Bandome įvairumą naudoti sprendžiant kai kurias sudėtingesnes problemas, turime pradėti kitaip galvoti apie įvairumą****.* (annotated **implicit** Contingency: Cause: Result)

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

Literatūra (I)

- Asher, N. (1993). Reference to Abstract Objects. Kluwer, Dordrecht.
- Aston, G. (2001). Learning with corpora: An overview. In G. Aston (ed.), *Learning with corpora (pp. 7–45). Houston: Athelstan.*
- Biber, D. & Conrad, S. (2001). Quantitative corpus-based research in TESOL: Much more than bean counting. *TESOL Quarterly 35, 2, 331–5.*
- Chalker, S. (1984). Current English Grammar. London: Mc Millan publishers.
- Halliday, M.A.K. and Hasan, R. 1992. *Cohesion in English.* Longman: Longman group Limited.
- Kamandulytė, L. (2006). Onimų reikšmės tyrimai tekstynų lingvistikos metodu, Lituanistica, t. 65, 1: 38-47.
- Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., & Webber, B. L. (2007). The Penn Discourse Treebank 2.0 Annotation Manual. https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf

Literatūra (II)

- Quirk, R.; Green Baum, S.; Leech, G and Svartvik, J. (1985). Comprehensive Grammar of the English Language. London: Longman.
- Schourup, L. (1982). Common Discourse particles in English. Published dissertation (1985). Ohio State. University. New York: Garland.
- Sinclair J. McH. (1999). The Lexical Item. Contrastive Lexical Semantics, ed. by E.Weigand, vol. 17, Current Issues in Linguistic Theory. Amsterdam/Philadelphia: John Benjamins.
- Sinclair J. McH. (2000). Current Issues in Corpus Linguistics. Ms.
- Utka A. (2000). Kalbinė programinė įranga ir jos galimybės. Darbai ir Dienos 24: 275-285.
- Webber, B. and Joshi, A. (1998). Anchoring a lexicalized tree-adjoining grammar for discourse. In M. Stede, L. Wanner, and E. Hovy, editors, Discourse Relations and Discourse Markers: Proceedings of the Conference, pages 86–9 for Computational Linguistics, Somerset, New Jersey.

Kuriame Lietuvos ateitį

2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa