

RTU studiju kurss "Valodas sistēmas un dati"**33000 Datorzinātnes, informācijas tehnoloģijas un enerģētikas fakultāte****Vispārējā informācija**

Kods	DE1091
Nosaukums	Valodas sistēmas un dati
Studiju kursa statuss programmā	Obligātais/Ierobežotās izvēles; Brīvās izvēles
Atbildīgais mācītbspēks	Oksana Ivanova - Doktors, Docents (praktiskais)
Mācītbspēks	Tatjana Kelebeka - Doktors, Docents Sintija Petroviča-Kļaviņa - Doktors, Docents
Apjoms daļās un kredītpunktos	1 daļa, 6.0 kredītpunkti
Studiju kursa īstenošanas valodas	LV, EN
Anotācija	Studiju kursā aplūktas galvenās valodas strukturālās apakšsistēmas (fonoloģija, morfoloģija, sintakse, semantika un pragmatika). Studējošie apgūst prasmi reprezentēt, iegūt un analizēt empīriskos valodas datus pētniecībai un tehnoloģiju attīstībai. Apvienojot teoriju ar praktisku darbu, studējošie apgūst korpusu izveides un vaicāšanas metodes, lingvistisko struktūru anotēšanas principus, kā arī dažādu datorizētu metožu (korpusu rīku, vārdšķiru un morfoloģiskās marķēšanas rīku u. c.) izmantošanu sistēmas līmeņa paraugu un valodas variācijas izpētei. Īpaša uzmanība tiek pievērsta formālu valodas sistēmu aprakstu sasaistīšanai ar uz datiem balstītām darba plūsmām, kas tiek izmantotas dabīgās valodas apstrādē (DVA) un valodas inženierijā, tostarp datu priekšapstrādē, anotēšanas vadlīniju izstrādē un reproducējamās analīzes veikšanā.
Mērķis un uzdevumi, izteikti kompetencēs un prasmēs	Studiju kursa mērķis ir attīstīt studējošo pamatkompetences lingvistisko sistēmu analizē, reprezentācijā un apstrādē, izmantojot uz datiem balstītas un datorizētas pieejas. Studiju kursa uzdevumi: - stiprināt studējošo izpratni par valodas strukturālajiem komponentiem (fonoloģiju, morfoloģiju, sintaksi, semantiku, pragmatiku); - nodrošināt studējošajiem praktiskās un pārnēsāmās prasmes lingvistisko datu vākšanā, anotēšanā un pārvaldībā digitālajā vidē; - attīstīt studējošo spēju izmantot datorizētus un statistiskus rīkus lingvistisko parādību analīzei un interpretācijai dažādās valodās; - veicināt kritisku izpratni par datu kvalitāti, anotēšanas standartiem un ētiskajiem principiem valodas datu tehnoloģijās; - nodrošināt studējošajiem iespēju integrēt teorētiskās valodniecības zināšanas ar empīrisku datu analīzi pētniecības, valodas modelēšanas un tehnoloģiju izstrādes vajadzībām.
Patstāvīgais darbs, tā organizācija un uzdevumi	Aktīva studējošo iesaiste patstāvīgā valodas sistēmu izpētē un datu analīzes uzdevumu veikšanā. Patstāvīgais darbs ietver nelielu lingvistisko datu kopu izstrādi un anotēšanu, korpusos balstītu vai datorizētu analīžu veikšanu, kā arī īsa raksta sagatavošanu par metodoloģiskiem jautājumiem. Studējošie izpilda praktiskos uzdevumus, izmantojot lingvistiskās programmatūras rīkus, un pielieto šos rīkus fonoloģisko, morfoloģisko, sintaktisko vai semantisko paraugu izpētei. Studējošajiem ir arī jāizstrādā gadījuma izpēte par izvēlēto valodas parādību, jāsniedz viena semināra prezentācija par sākotnējiem rezultātiem, jāiesniedz divi analītiski ziņojumi un noslēguma pētniecības projekts, kas apliecina spēju integrēt teorētiskās zināšanas ar uz datiem balstītām pieejām.
Literatūra	Obligātā. / Obligatory: Jurafsky, Daniel & Martin, James H. Speech and Language Processing. Stanford University, 2025. Bird, Steven, Klein, Ewan, & Loper, Edward. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, 2009. https://www.nltk.org/book/ Bender, Emily M., & Lascarides, Alex. Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics. Morgan & Claypool, 2019. Papildu. / Additional: Mitkov, Ruslan. The Oxford Handbook of Computational Linguistics (2nd ed.). Oxford University Press, 2022. Podesva, Robert J. and Sharma, Devyani. Research Methods in Linguistics. Cambridge University Press, 2013. Citi avoti. / Other sources: CLARIN Language Resources . Available at: https://www.clarin.eu/
Nepieciešamās priekšzināšanas	Angļu valoda B2 līmenī saskaņā ar Eiropas kopējām vadlīnijām, ja studiju kurss tiek apgūts angļu valodā. Angļu valodas prasmes, lai lasītu obligāto literatūru, ja studiju kurss tiek apgūts latviešu valodā.

Studiju kursa saturs

Saturs	Pilna un nepilna laika klātienē studijas		Nepilna laika neklātienē studijas	
	Kontakt stundas	Patstāv. darbs	Kontakt stundas	Patstāv. darbs
Ievads valodas sistēmās un lingvistiskajos datos. Lingvistisko datu loma valodas tehnoloģijās.	4	6	0	0
Fonoloģiskās sistēmas un datu reprezentācija. Fonoloģisko datu digitalizēšana un anotēšana.	4	6	0	0
Morfoloģija un morfoloģiskie dati. Morfoloģiskā marķēšana un anotācijas shēmas.	6	8	0	0

Sintakse: struktūra un anotācija. Koku bankas (tree banks), sintaktiskā analīze (parsing) un sintaktisko datu formāti.	6	8	0	0
Semantika un nozīmes reprezentācija. Semantiskā anotēšana un vārda nozīmes nošķiršanas datu kopas.	6	8	0	0
Datu avoti un korpusu izstrāde. Korpusu veidi. Izlases veidošana, metadati un reprezentativitāte.	4	6	0	0
Anotācijas principi un datu kvalitāte. Vadlīnijas, marķētāju vienprātība un ētiskie jautājumi. Rīki manuālai un automātiskai anotēšanai.	6	10	0	0
Datu formāti, standarti un savietojamība. Lingvistisko datubāzu un ontoloģiju sasaistīšana.	4	6	0	0
Kvantitatīvās un statistiskās metodes lingvistiskajos datos. Pamata statistika lingvistiskajiem pētījumiem. Biežuma analīze, kolokācijas un asociācijas rādītāji.	4	6	0	0
No lingvistiskajām sistēmām līdz skaitļošanas modeļiem. Valodas struktūras kartēšana dabisko valodu apstrādes (DVA) uzdevumu kontekstā. Noteikumos balstītas pieejas un mašīnmācīšanās metodes.	6	10	0	0
Daudzvalodu un starpvalodu dati. Daudzvalodu korpusu līdzināšana.	6	10	0	0
Gadījumu izpēti uz datiem balstītā valodniecībā. Korpusu pētnieciskie projekti, kas sasaista struktūru un datus. Reproducējamās darba plūsmas un rezultātu vizualizācija.	8	14	0	0
Kopā:	64	98	0	0

Sasniedzamie studiju rezultāti un to vērtēšana

Sasniedzamie studiju rezultāti	Rezultātu vērtēšanas metodes
Spēj demonstrēt padziļinātu izpratni par galvenajām valodas apakšsistēmām (fonoloģiju, morfoloģiju, sintaksi, semantiku un pragmatiku) un to savstarpējām saiknēm.	Eksāmens. Patstāvīgie analītiskie uzdevumi. Kritēriji: studenti demonstrē izpratni par galvenajām valodas apakšsistēmām un spēj analizēt tos, izmantojot atbilstošus analītiskus rīkus.
Spēj vākt, anotēt un pārvaldīt lingvistiskos datus atbilstoši noteiktajiem standartiem un anotācijas ietvariem.	Praktiskie datu anotēšanas uzdevumi. Kritēriji: tiek vērtēta studentu spēja izvērtēt lingvistisko datu kopu un anotēšanas praksi uzticamību, validitāti un atbilstību ētikas principiem.
Spēj integrēt valodniecības teoriju ar empirisku datu analīzi patstāvīgajā pētniecības projektā.	Analītiska esija par datu kvalitāti un pētniecības ētiku. Kritēriji: tiek vērtēta studentu spēja izmantot korpusos balstītus un datorizētus rīkus lingvistisko struktūru un paraugu analīzei un interpretācijai.
Spēj efektīvi komunicēt pētījuma rezultātus rakstiskā un mutiskā formā, kā arī piedalīties akadēmiskās diskusijās.	Dalība semināros un prezentācijā. Kritēriji: tiek vērtēta studentu prasme pamatot savu viedokli un prezentēt darba rezultātus, izmantojot jomai specifiskus jēdzienus un terminoloģiju.

Studiju rezultātu vērtēšanas kritēriji

Kritērijs	% no kopējā vērtējuma
Dalība semināros un prezentācija	10
Patstāvīgie analītiskie uzdevumi	20
Analītiska esija par datu kvalitāti un pētniecības ētiku	15
Praktiskie datu anotēšanas uzdevumi	25
Eksāmens	30
Kopā:	100

Studiju kursa plānojums

Daļa	KP	Stundas			Pārbaudījumi			Brīvās izvēles pārbaudījumi		
		Lekcijas	Prakt d.	Laborat	Ieskaite	Eksām.	Darbs	Ieskaite	Eksām.	Darbs
1.	6.0	32.0	32.0	0.0		*			*	