UČNI NAČRT PREDMETA / COURSE SYLLABUS	
Predmet:	Korpusni pristop in jezikovne tehnologije v leksikografiji
Course title:	Corpus linguistics and the use of language technologies in lexicography

Študijski program in stopnja Study programme and level	Študijska smer Study field	Letnik Academic year	Semester Semester
Primerjalni študij idej in kultur, doktorski študij 3. stopnje	Leksikologija, leksikografija, slovničarstvo	Brez letnika	/
Comparative studies of ideas and cultures, doctoral study 3rd level	Lexicology, lexicography, gramaticography	Not specified	/

Vrsta predmeta / Course type	splošno izbirni / general elective
Univerzitetna koda predmeta / University course code:	57

Predavanja Lectures	Seminar Seminar	Vaje Tutorial	Klinične vaje work	Druge oblike študija	Samost. delo Individ. work	ECTS
60	30				90	6

Nosilec predmeta / Lecturer:	Izr. prof. dr. Tomaž Erjavec, izr. prof. dr. Darja Fišer

Jeziki / Languages:	**Predavanja / Lectures:**	slovenščina, angleščina / Slovenian, English
	Vaje / Tutorial:	/

Pogoji za vključitev v delo oz. za opravljanje študijskih obveznosti:	**Prerequisits:**
Ni posebnih pogojev.	None required.

Vsebina:	**Content (Syllabus outline):**
1. Uvod a. Humanistika in računalništvo b. Formalno in korpusno jezikoslovje c. Slovarski in drugi digitalni priročniki d. Jezikovni viri in jezikovne tehnologije 2. Korpusno jezikoslovje a. Namembnost, definicija, zgodovina b. Zvrsti korpusov s primeri c. Korpusne oznake d. Uporaba konkordančnikov e. Regularni izrazi 3. Strukturiranje jezikovnih virov a. Standardi in odprta koda b. Nabori znakov c. Standard XML d. č. Priporočila TEI e. Jezikoslovne oznake 4. Izdelava jezikovnih virov a. Proces izdelave korpusa b. Licence, avtorske pravice in varovanje zasebnosti	1. Introduction a. Humanities and computer science; b. Formal and corpus linguistics; c. Dictionaries and other digital manuals; d. Language resources and technologies. 2. Corpus linguistics a. Purpose, definition, historical development; b. Types of corpora with representative examples; c. Corpus labels; d. The use of concordance software; e. Regular expressions. 3. Language resources management a. Standards and Open source; b. Character encoding sets; c. XML standard; d. TEI guidelines; e. Linguistic annotations. 4. Creation and development of language resources

c. Zasnova označevalskega projekta
d. Okolja za označevanje
e. Množičenje
f. Programi osnovani na pravilih
g. Strojno učenje

Seminarji

- Seminarji potekajo vzporedno s predavanji in se navezujejo na posamezne tematske sklope predavanj. Poudarek je na samostojnem raziskovanju in predstavitvi izbranih problemov, kar vključuje seznanjanje z izbrano literaturo, uporabo jezikovnih virov (predvsem korpusov) pri raziskovanju problema in predstavitev rezultatov, ki jih obiskovalci seminarja skupno analiziramo.

Povezava z drugimi predmeti

- Za uspešno razumevanje snovi je potrebno osnovno poznavanje dela z računalniki in sposobnost logičnega mišljenja. Predmet se smiselno povezuje s predmetoma Leksikologija, leksikografija in slovničarstvo sodobnega jezika in Zgodovinska leksikologija in leksikografija ter zgodovinska slovnica.

a. The setting up of a language corpus;
b. Licenses, copyright and privacy laws;
c. Corpus annotation;
d. Annotation environment;
e. Crowdsourcing;
f. Rule-based programmes;
g. Machine learning.

Seminar classes:

- Seminar classes accompany the lectures and complement them by exploring individual chapters covered by the course syllabus. The main objective is to provide the student with the opportunity to conduct individual research on the selected topic. This also involves individual study of relevant reference works and independent use of language resources (especially language corpora). The results of the student's own research are analysed as part of group work in seminars.

Cross-curricular integration:

- Basic computer knowledge and logical thinking ability are necessary for a successful participation in the course. Within the module the course is cross-referenced by »Lexicology, lexicography, contemporary grammar« and »Historical lexicology, historical lexicology and historical grammar«.

Temeljni literatura in viri / Readings:

Na seznamu je osnovna literatura, poleg katere bodo študentke in študenti prejeli še dodaten seznam besedil, ki bodo prišla v poštev za posamezna predavanja in seminarsko delo.

The following list contains basic reference works. A series of additional readings for individual lectures and/or seminars will be supplied subsequently.

- Erjavec, Tomaž. 2013: Korpusi in konkordančniki na strežniku nl.ijs.si. Slovenščina 2.0, 1/1, str. 24-49. http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf.
- Finlayson, Mark A., Erjavec, Tomaž. Overview of annotation creation: processes and tools. V: IDE, Nancy M. (ur.), PUSTEJOVSKY, James (ur.). Handbook of linguistic annotation. Amsterdam: Springer. 2017, str. 167-192. https://arxiv.org/abs/1602.05753
- Fišer, Darja, Ljubešić, Nikola, Erjavec, Tomaž. The Janes project: language resources and tools for Slovene user generated content. Language resources and evaluation. 2020, vol. 54, str. 223–246. https://rdcu.be/7RX4
- Fišer, Darja, Ljubešić, Nikola. Distributional modelling for semantic shift detection. International journal of lexicography, ISSN 0950-3846, June 2019, vol. 32, no. 2, str. 163-183
- Gorjanc, Vojko, Fišer, Darja 2013: Korpusna analiza. 2., predelana in razširjena izd. Ljubljana: Znanstvena založba Filozofske fakultete.
- Logar, Nataša in dr. 2012: Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES:

gradnja, vsebina, uporaba. Zbirka Sporazumevanje. Ljubljana: Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede. https://knjigarna.fdv.si/i_578_korpusi-slovenskega-jezika-gigafida-kres-ccgigafida-in-cckres-gradnja-vsebina-uporaba

- Raziskovalna infrastruktura CLARIN.SI: http://www.clarin.si/
- Standard XML: http://en.wikipedia.org/wiki/XML
- Priporočila TEI: https://tei-c.org/

Cilji in kompetence:

Sodobna leksikografija in slovničarstvo sta nepredstavljiva brez računalniških orodij, tako pri raziskovanju jezikovnega materiala kot za strukturiranje in računalniški prikaz rezultatov dela. Predmet zato obravnava hitro razvijajoče se področje digitalne humanistike, osredotočeno na korpusno jezikoslovje in jezikovne tehnologije slovenskega jezika. Na teh področjih se je v zadnjih letih zgodil velik premik, tako da je sedaj dostopno večje število korpusov, od referenčnega korpusa sodobnega jezika Gigafida, govornega korpusa GOS, korpusa starejših besedil IMP itd., obstaja pa tudi že večje število (spletnih) orodij za jezikoslovne označevanje, npr. lematizatorji ter oblikoskladenjski in skladenjski označevalniki. Cilj predmeta je študentom in študentkam podati znanja, da bodo obstoječe korpuse in orodja znali uporabljati ter individualno ali v okviru projektov izdelovati nove. Predmet bo obravnaval tri tematike: korpusno jezikoslovje, strukturiranje jezikovnih virov ter njihovo jezikoslovno označevanje. Pri korpusnem jezikoslovju bo poudarek na spoznavanju instrumentarija, ki ga ponujajo sodobni konkordančniki, od konkordanc in frekvenčnih leksikonov do ključnih besed in kolokacij. Tu bo ključno razumevanje regularnih izrazov, korpusnih oznak in specifik ter namembnosti dostopnih korpusov in konkordančnikov slovenskega jezika. Za bolj poglobljeno razumevanje korpusov in digitalnih slovarskih baz ter formalnih opisov jezikovnih modelov so potrebna osnovna računalniška znanja s področja zapisa znakov in strukturiranja besedilnih podatkov. Za prvo je glavni standard Unikod, ki omogoča kodiranje vseh svetovnih abeced, za drugo pa standard XML, ki je meta-jezik za označevanje polstrukturiranih podatkov. V XML je mogoče definirati sheme, ki

Objectives and competences:

Computer-aided research of language material and its presentation is an integral part of contemporary lexicography and grammaticography. The course is therefore devoted to the area of digital humanities, with special emphasis on corpus linguistics and Slovenian language technologies. Over the past few years both research areas have witnessed substantial progress, which has resulted in the availability of a wide range of language corpora (e.g. reference corpus Gigafida, spoken language corpus GOS, two corpora IMP for historical Slovene etc.), accompanied by a large number of (online) tools for linguistic annotation such as lemmatizers, morpho-syntactic and syntactic annotation tools etc. The main objective of the course is to equip students with sufficient knowledge to encourage independent use of the available corpora and other tools in contemporary linguistic research and development of new language technologies. The course will cover three thematic fields: corpus linguistics, language resource management and linguistic annotation. The first part will present the apparatus offered by contemporary concordance programs (concordance, frequency lexicon, key words, and collocations), which will require basic comprehension of regular expressions, corpus annotations and specifications, and the functionalities of the available corpora and concordancers for Slovene. For a more sophisticated insight about the corpora, digital dictionary databases and formal descriptions of language models basic computer knowledge is required, both for encoding and structuring of textual data. The most widely used standard for character encoding is Unicode, in which most of the world's writing systems can be represented, while XML, as a meta-language, serves the

določajo besedišče in medsebojne odvisnosti oznak za posamezne zvrsti dokumentov, pri čemer obstaja že večje število standardiziranih shem. Za strukturiranje in označevanje besedil v humanističnih študijah so se najbolj uveljavila Priporočila za kodiranje besedil TEI (Text Encoding Initiative Guidelines), s katerimi je mogoče zapisati zelo raznovrstna besedila, in to v poljubnem jeziku, uporabljajo pa se tudi v večini korpusov slovenskega jezika. V sklopu predavanj bodo obravnavane osnove Unikoda, XML, XML shem in TEI, s čimer bodo študenti in študentke dobili dobro osnovo za samostojno uporabo teh standardov in priporočil. V zadnjem sklopu predavanj se bomo podrobneje posvetili metodam za izdelavo jezikovnih virov, predvsem korpusov, od zbiranja besedil, njihove obdelave in ročnega označevanja. Podrobneje bodo obravnavane avtomatske metode označevanja s poudarkom na metodah, ki temeljijo na strojnem učenju, saj je to v zadnjih letih postala najuspešnejša metoda za jezikoslovno označevanje.	annotation of semistructured data. XML makes it possible to define schemas (of which several standardised models are in use) that specify the lexis and reciprocality of annotations for individual types of documents. Text encoding and linguistic annotation in humanities generally follow TEI Guidelines (Text Encoding Initiative Guidelines), which enable the generation of a highly diverse range of texts in any given language and are utilised by the majority of Slovene language corpora. The course will cover the basics of Unicode, XML, XML Schemas and TEI, which will provide students with a good foundation for future confident use of standards and guidelines. The last series of lectures will be devoted to various approaches and methods in the development of language resources, predominantly language corpora, involving text collecting, data processing and manual annotation. More detailed examination will be devoted to automatised annotation with particular emphasis on machine learning, which in the last few years has proved to be the most successful method in linguistic annotation.

Predvideni študijski rezultati:	Intended learning outcomes:
• poglobljena seznanjenost in praktične kompetence za uporabo instrumentarija korpusnega jezikoslovja • seznanjenost in praktične kompetence pri strukturiranju jezikovnih virov • poznavanje postopkov ročnega in strojnega jezikoslovnega označevanja • specializirana informacijskotehnološka znanja	• Detailed familiarity with the apparatus of corpus linguistics and the practical ability to use the resources; • Managing language resources; • Principles of manual and automated linguistic annotation; • Specialised knowledge in information technology.

Metode poučevanja in učenja:	Learning and teaching methods:
Oblike dela: ☒Frontalna oblika poučevanja ☐Delo v manjših skupinah oz. v dvojicah ☒Samostojno delo študentov ☒e-izobraževanje Metode (načini) dela: ☒Razlaga ☒Razgovor/ diskusija/debata ☐Delo z besedilom ☒Proučevanje primera ☐Igra vlog	Types of learning/teaching: ☒Frontal teaching ☐Work in smaller groups or pair work ☒Independent students work ☒e-learning Teaching methods: ☒Explanation ☒Conversation/discussion/debate ☐Work with texts ☒Case studies ☐Roleplay

☐Druge vrste nastopov študentov ☐Reševanje nalog ☐"Terenske vaje" (npr. obiski podjetij) ☐Vključevanje gostov iz prakse	☐Different presentation ☐Solving exercises ☐Field work (e.g. company visits) ☐Inviting guests from companies

Načini ocenjevanja:	Delež (v %) / Weight (in %)	Assessment:
Krajši pisni izdelki		Short written assignments
Daljši pisni izdelki	80	Long written assignments
Javni nastop ali predstavitev		Presentations
Končno ocenjevanje (pisni/ustni izpit)	20	Final examination (written/oral)
Drugo		Other

Reference nosilca / Lecturer's references:

- **ERJAVEC, Tomaž**. The IMP historical Slovene language resources. *Language resources and evaluation*, ISSN 1574-020X, 2015, vol. 49, no. 3, str. 753-775, doi: 10.1007/s10579-015-9294-7. [COBISS.SI-ID 28321575]
- **ERJAVEC, Tomaž**. MULTEXT-East. V: IDE, Nancy M. (ur.), PUSTEJOVSKY, James (ur.). *Handbook of linguistic annotation*. Amsterdam: Springer. 2017, str. 441-462. [COBISS.SI-ID 30614055]
- DIVJAK, Dagmar, SHAROFF, Serge, **ERJAVEC, Tomaž**. Slavic corpus and computational linguistics. *Journal of Slavic linguistics*, ISSN 1068-2090, 2017, vol. 25, no. 2, str. 171-198, doi: 10.1353/jsl.2017.0008. [COBISS.SI-ID 31281191]
- PRUNČ, Erich, OGRIN, Matija, **ERJAVEC, Tomaž**. Kapelski pasijon v elektronski znanstvenokritični izdaji. "Nov" rokopis, nove raziskovalne poti. V: KRŽIŠNIK, Erika (ur.), HLADNIK, Miran (ur.). *Toporišičeva obdobja*, (Obdobja, ISSN 1408-211X, Simpozij, = Symposium, 35). 1. natis. Ljubljana: Znanstvena založba Filozofske fakultete. 2016, str. 395-402, ilustr. http://centerslo.si/wp-content/uploads/2016/11/PruncOgrErj.pdf. [COBISS.SI-ID 62810978]
- **FIŠER, Darja**, LJUBEŠIĆ, Nikola, **ERJAVEC, Tomaž**. The Janes project: language resources and tools for Slovene user generated content. *Language resources and evaluation*, ISSN 1574-020X, 2020, vol. 54, no. 1, str. 223-246, ilustr., doi: 10.1007/s10579-018-9425-z. [COBISS.SI-ID 68029026]
- **FIŠER, Darja**, LJUBEŠIĆ, Nikola. Distributional modelling for semantic shift detection. *International journal of lexicography*. June 2019, vol. 32, no. 2, str. 163-183, ilustr., tabele. ISSN 0950-3846. https://academic.oup.com/ijl/advance-article/doi/10.1093/ijl/ecy011/5051703, DOI: 10.1093/ijl/ecy011. [COBISS.SI-ID 67380066]
- **FIŠER, Darja**, SAGOT, Benoît. Constructing a poor man's wordnet in a resource-rich world. *Language resources and evaluation*. 2015, vol. 49, str. 601-635, ilustr. ISSN 1574-020X. http://link.springer.com/article/10.1007/s10579-015-9295-6/fulltext.html, DOI: 10.1007/s10579-015-9295-6. [COBISS.SI-ID 56782434]
- GORJANC, Vojko, **FIŠER, Darja**. *Korpusna analiza*. 2., predelana in razširjena izd. Ljubljana: Znanstvena založba Filozofske fakultete, 2013. 88 str., ilustr. ISBN 978-961-237-610-9. [COBISS.SI-ID 269429504]
- GORJANC, Vojko, **FIŠER, Darja**. Twitter in razmerja moči: diskurzna analiza kampanj ob referendumu za izenačitev zakonskih zvez v Sloveniji. *Slavistična revija: časopis za jezikoslovje in literarne vede*. [Tiskana izd.]. okt.-dec. 2018, letn. 66, št. 4, str. 473-495, ilustr. ISSN 0350-6894. https://srl.si/ojs/srl/article/view/2018-4-1-5. [COBISS.SI-ID 68754274]
- MILIČEVIĆ, Maja, LJUBEŠIĆ, Nikola, **FIŠER, Darja**. Nestandardno zapisivanje srpskog jezika na Tviteru: mnogo buke oko malo odstupanja?. *Anali Filološkog fakulteta*. 2017, vol. 29, no. 2, str.

111-136, ilustr. ISSN 0522-8468. http://doi.fil.bg.ac.rs/pdf/journals/analiff/2017-2/analiff-2017-29-2-8.pdf, DOI: 10.18485/analiff.2017.29.2.8. [COBISS.SI-ID 66613858]