UČNI NAČRT PREDMETA / COURSE SYLLABUS	
Predmet:	Jezikovne tehnologije
Course title:	Language Technologies

Študijski program in stopnja Study programme and level	Modul Module	Letnik Academic year	Semester Semester
Informacijske in komunikacijske tehnologije, 2. stopnja	Tehnologije znanja	1	2
Information and Communication Technologies, 2nd cycle	Knowledge Technologies	1	2

Vrsta predmeta / Course type	Izbirni / Elective
Univerzitetna koda predmeta / University course code:	IKT2-714

Predavanja Lectures	Seminar Seminar	Sem. vaje Tutorial	Lab. vaje Laboratory work	Druge oblike	Samost. delo Individ. work	ECTS
30	**30**			**30**	**210**	**10**

**Navedena porazdelitev ur velja, če je vpisanih vsaj 15 študentov. Drugače se obseg izvedbe kontaktnih ur sorazmerno zmanjša in prenese v samostojno delo. / This distribution of hours is valid if at least 15 students are enrolled. Otherwise the contact hours are linearly reduced and transfered to individual work.*

Nosilec predmeta / Lecturer:	Prof. dr. Tomaž Erjavec

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

Pogoji za vključitev v delo oz. za opravljanje študijskih obveznosti:	**Prerequisites:**
Zaključen študijski program prve stopnje s področja naravoslovja, tehnike ali računalništva.	Student must complete first-cycle study programmes in natural sciences, technical disciplines or computer science.

Vsebina:	**Content (Syllabus outline):**
Uvod: Razvoj jezikoslovja in računalniškega jezikoslovja, kompleksnost jezika, ravni analize jezika, pregled aplikacij in metod. Jezikovni korpusi: Namen, zgodovina in tipologija, označevanje, uporaba, računalniški zapis, konkretni primeri. Metode računalniške obravnave: Regularni izrazi in končni avtomati, frazne gramatike, statistične metode, strojno učenje. Področja uporabe:	Introduction: Development of linguistics and computational linguistics, complexity of language, levels of linguistic analysis, overview of applications and methods. Language corpora: Purpose, history and typology, annotation, use cases, computer coding, specific examples. Methods of computer processing: Regular expressions and finite state automata, phrase-structure grammars, statistical methods, machine learning.

Iskanje in zajemanje informacij, strojno prevajanje, govorne tehnologije, digitalne knjižnice, itd.	Applications: Information retrieval and extraction, machine translation, speech technologies, digital libraries, etc.

Temeljna literatura in viri / Readings:

Izbrana poglavja iz naslednjih knjig: / Selected chapters from the following books:

- D. Jurafsky, and J.H. Martin. *Speech and Language Processing*, 2nd Edition. Prentice-Hall, 2008. ISBN 978-0131873216
- R. Mitkov (ed.). *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2003. ISBN 978-0-19-823882-9
- C. Manning, and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press. 1999. ISBN 0-262-13360-1
- N. Ide and J. Pustejovsky (eds.). *Handbook of Linguistic Annotation*. Springer. 2017. I SBN 978-94-024-0881-2

Cilji in kompetence:	**Objectives and competences:**
Cilj predmeta je posredovati splošno znanje o jezikovnih tehnologijah, to je metodah in aplikacijah obdelave naravnega jezika na računalniku. Predstavljena je zgodovina in osnovni pojmi jezikoslovja, raznovrstne aplikacije jezikoslovnih tehnologij in računalniške metode, ki se pri njih uporabljajo. Podrobno so obdelani jezikovni korpusi, velike zbirke označenih besedil, ki so osnovna infrastruktura, potrebna za raziskave in obdelavo posameznih jezikov. Obravnavana je tudi analiza jezikovnih korpusov z metodami strojnega učenja. Poudarek predmeta je na obravnavi slovenskega jezika. Slušatelji pridobijo osnovno teoretično razumevanje in nekaj praktičnih izkušenj s področij jezikovnih tehnologij in korpusnega jezikoslovja, kar je predpogoj za učinkovito delo na računalniški obdelavi jezikovnih podatkov.	The goal of this course is to introduce language technologies, i.e. methods and applications of computer processing of natural language. The course gives the history and basic concepts of linguistics, various applications of language technologies and the computational methods which they use. Particular attention is given to language corpora, large datasets of annotated texts, which serve as the basic infrastructure necessary for research and processing of individual languages. Also discussed is the analysis of language corpora with machine learning methods. The focus of the course is on the processing of Slovene language. Students will gain basic theoretical understanding and some practical knowledge of language technologies and corpus linguistics, which is a prerequisite for effective work on computer processing of language data.

Predvideni študijski rezultati:	**Intended learning outcomes:**
Študenti bodo z uspešno opravljenimi obveznostmi tega predmeta pridobili: • sposobnost analize, sinteze in predvidevanja rešitev ter posledic • obvladanje raziskovalnih metod, postopkov in procesov, razvoj kritične in samokritične presoje • zavezanost profesionalni etiki in regulativi • poznavanje zgodovine razvoja in razumevanje konceptov računalniškega jezikoslovja • osnovno poznavanje tradicionalnih in	Students successfully completing this course will acquire: • an ability to analyse, synthesise and anticipate solutions and consequences • to gain the mastery over research methods, procedures and processes, a development of the critical judgement • complying with professional ethics and regulatory body policies • knowledge of history and concept of computational linguistics

naprednih metod za obdelavo naravnih jezikov • pregledno znanje aplikacij jezikovnih tehnologij, njihovih lastnosti in omejitev z vidika možne uporabe v praksi, posebej za slovenski jezik • sposobnost integriranja znanja in obvladovanja kompleksnosti pri reševanju specifičnih problemov v računalniških aplikacijah	• basic understanding of traditional and advanced methods for natural language processing • overview knowledge of language technology applications, their features and limitations for possible applications in practice • ability to integrate knowledge and handle complexity when solving specific problems in computer applications

Metode poučevanja in učenja:	**Learning and teaching methods:**
Predavanja, seminar, konzultacije, samostojno delo	Lectures, seminar, consultations, individual work

Načini ocenjevanja:	Delež (v %) / Weight (in %)	**Assessment:**
Seminar	50 %	Seminar
Ustni izpit	50 %	Oral exam

Reference nosilca / Lecturer's references:

- Y. Scherrer, **T. Erjavec**. Modernising historical Slovene words. Natural language engineering, ISSN 1351-3249, 2016, vol. 22, no. 6, str. 881-905.
- **T. Erjavec**, The IMP historical Slovene language resources. *Language resources and evaluation*, 23 str., doi: 10.1007/s10579-015-9294-7, 2015.
- **T. Erjavec**, MULTEXT-East. V: Ide, N. (ur.), Pustejovsky , J. (ur.), Handbook of linguistic annotation. Amsterdam: Springer. 2017, str. 441-462.
- D. Divjak, S. Sharoff, **T. Erjavec**. 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
- **T. Erjavec**, N. Ljubešić, N. Logar. The slWaC corpus of the Slovene Web. *Informatica : an international journal of computing and informatics*, ISSN 0350-5596, Mar. 2015, vol. 39, no. 1, str. 35-42