Утвърдил: ……..

Декан

Дата .....

СОФИЙСКИ УНИВЕРСИТЕТ “СВ. КЛИМЕНТ ОХРИДСКИ”

Факултет: Славянски филологии

Специалност: *(код и наименование)*

.....

Магистърска програма: *(код и наименование)*

.....

УЧЕБНА ПРОГРАМА

Дисциплина:

(код и наименование) Лингвистична анотация

Преподавател: проф. д-р Андрей Бояджиев

Асистент:

Учебна заетост	Форма	Хорариум
Аудиторна заетост	Лекции	30
	Семинарни упражнения	
	Практически упражнения (хоспетиране)	
Обща аудиторна заетост		**30**
Извънаудиторна заетост	Реферат	
	Доклад/Презентация	
	Научно есе	
	Курсов учебен проект	
	Учебна екскурзия	
	Самостоятелна работа в библиотека или с ресурси	120
Обща извънаудиторна заетост		**120**
ОБЩА ЗАЕТОСТ		**180**
Кредити аудиторна заетост		**2**
Кредити извънаудиторна заетост		**6**
ОБЩО ЕКСТ		8

№	Формиране на оценката по дисциплината[1]	% от оценката
1.	Workshops {информационно търсене и колективно обсъждане на доклади и реферати)	
2.	Участие в тематични дискусии в часовете	20
3.	Демонстрационни занятия	
4.	Посещения на обекти	
5.	Портфолио	
6.	Тестова проверка	
7.	Решаване на казуси	20
8.	Курсов проект	20
9.		
10.		
11.		
12.	Изпит	40

Анотация на учебната дисциплина:

Маркиращите езици са в основата на лингвистичната анотация. Независимо от техния формат, съвременните езикови корпуси разчитат на тези езици за назоваване, сегментация, търсене, извличане, форматиране и публикация на резултатите от данните. Този уводен курс предлага преглед на най-важните технологии, подходи и формати в съвременната компютърна лингвистика в това отношение. Разчита се на самостоятелната работа и активното участие на студентите. В зависимост от техните интереси в курса ще се акцентира върху определени технологии за маркиране на езикови данни.

Предварителни изисквания:

Няма

Очаквани резултати:

Студентите следва да могат да анотират самостоятелно лингвистични данни с цел подготовка за работа по различни проекти в бизнес среда или в академичната общност.

Учебно съдържание

№	Тема:	Хорариум
1.	Маркиращите езици. Понятие. История	2
2.	XML. Основни характеристики. Семейството на XML.	4

[1] В зависимост от спецификата на учебната дисциплина и изискванията на преподавателя е възможно да се добавят необходимите форми, или да се премахнат ненужните.

	Търсене, трансформация, публикуване	
3.	XML, електронни издания, корпуси и бази от данни	2
4.	XML и лингвистичната анотация	2
5.	Markdown. Понятие. Основни характеристики. Markdown в лингвистиката	1
6.	JSON (JavaScript Object Notation). Понятие. Основни характеристики. JSON в лингвистиката	2
7.	Форматът CoNLL-U за лингвистична анотация	4
8.	NER (Named-Entity Recognition). Понятие. Основни характеристики. NER в лингвистиката	2
9.	BIO/IOB формат. Понятие. Основни характеристики. BIO/IOB в лингвистиката	1
10.	TeX и LaTeX. Понятие. Основни характеристики. Дистрибуции и формати	2
11.	Технологията TeX и използването ѝ в лингвистиката	2
12.	Метаданни и онтологии. Контролирани речници. Понятия. Формати и препоръки за представяне на метаданни	4
13.	Указания за изготвяне на изпита / курсовата задача	2

Конспект за изпит

№	Въпрос
1	Изпитът представлява курсов учебен проект, който се състои от лингвистична анотация и документацията към нея и е изцяло с практическа насоченост. Студентите представят анотиран текст според някой от форматите, разгледани в курса, като анотацията следва да обхваща граматическа и/или семантична информация, снабдена с метаданни за изработения корпус.

Библиография

Основна:

About: Inside-Outside-beginning. In: DBPedia: <https://dbpedia.org/page/Inside%E2%80%93outside%E2%80%93beginning_(tagging)>

Adobe Inc. 2021. Syntax for OpenType features in CSS. <https://helpx.adobe.com/bg/fonts/using/open-type-syntax.html>

Banerjee, E. A. Cairncross, N. Haket, and S. Kidwai. A crash course in LATEX (for linguistics. University of Cambridge <https://www.mmll.cam.ac.uk/files/copil_presentation_4.pdf>

Carpintero, D. 2023. Named Entity Recognition to Enrich Text. In: OpenAI Cookbook <https://cookbook.openai.com/examples/named_entity_recognition_to_enrich_text>

CSS Tutorial. <https://www.w3schools.com/css/default.asp>

Comparison of document markup languages. <https://en.wikipedia.org/wiki/Comparison_of_document_markup_languages>
Cover Pages: Extensible Markup Language <http://xml.coverpages.org/xml.html>; General Introductions and Overviews <https://xml.coverpages.org/general.html>
di Buono, M., H. Gonçalo Oliveira, V. Barbu Mititelu, B. Spahiu and G. Nolano. 2022. Paving the way for enriched metadata of linguistic linked data. Semantric Web 13(6). 1133–1157. <https://content.iospress.com/articles/semantic-web/sw222994>
General-purpose markup languages: https://en.wikipedia.org/wiki/General-purpose_markup_language
Goldfarb, C. 1996. The Roots of SGML -- A Personal Recollection. <http://www.sgmlsource.com/history/roots.htm>
HTML Tutorial. <https://www.w3schools.com/html/default.asp>
Keith, Jeremy. 2010. A Brief History of Markup. <https://alistapart.com/article/a-brief-history-of-markup/>
Krishman, V. and V. Ganapathy. 2005. Named Entity Recognition. <https://cs229.stanford.edu/proj2005/KrishnanGanapathy-NamedEntityRecognition.pdf>
LaCara, N. 2018. LATEX for Linguistics <http://individual.utoronto.ca/nlacara/misc/lfl.pdf>
LaTeX/ Linguistics: https://en.wikibooks.org/wiki/LaTeX/Linguistics
Lightweight markup languages. <https://en.wikipedia.org/wiki/Lightweight_markup_language>
List of document markup languages. <https://en.wikipedia.org/wiki/List_of_document_markup_languages>
Lund, Gunnar. 2000. Markdown for linguists. <https://gunnarnl.github.io/posts/markdown-for-linguists.html>
Markdown help. Linguistics. <https://linguistics.stackexchange.com/editing-help>
Ramshaw, L. and M. Markus. 1995. Text Chunking Using Transformation-Based Learning. In: *Third Workshop on Very Large Corpora.* MIT <*https://aclanthology.org/W95-0107*>
Schalley, A. C. 2019. Ontologies and ontological methods in linguistics. *Language and Linguistics Compass* 13(11). <https://compass.onlinelibrary.wiley.com/doi/full/10.1111/lnc3.12356>
Schulze, Jessica. 2024. What Is Named Entity Recognition (NER) and How Does It Work? <https://www.coursera.org/articles/named-entity-recognition>
Stührenberg, M. 2012. The TEI and Current Standards for Structuring Linguistic Data. *Journal of Text Encoding Initiative* 3. <https://journals.openedition.org/jtei/523**>**
TEI Consortium, eds. 2023a. "A Gentle Introduction to XML." *Guidelines for Electronic Text Encoding and Interchange*. <https://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html>
TEI Consortium, eds. 2023b. "Language Corpora". *Guidelines for Electronic Text Encoding and Interchange*. <https://tei-c.org/release/doc/tei-p5-doc/en/html/CC.html>
TEI Consortium, eds. 2023c. "Simple Analytic Mechanisms". *Guidelines for Electronic Text Encoding and Interchange*. <https://tei-c.org/release/doc/tei-p5-doc/en/html/CC.html>
Universität des Saarlandes. Corpus building with XML and TEI: Introduction to XML. <http://fedora.clarin-d.uni-saarland.de/teaching/Corpus_Linguistics/Tutorial_XML.html>
Wiki Syntax help. MediaWiki, Help: Formating. <https://www.mediawiki.org/wiki/Help:Formatting>
Wishart, Ryder A. 2020. Using XML for Linguist Research. <https://ryder.dev/xml-for-linguistics/>
XML Tutorial. <https://www.w3schools.com/xml/default.asp>
XPath Tutorial. <https://www.w3schools.com/xml/xpath_intro.asp>
XSLT Tutorial. <https://www.w3schools.com/xml/xsl_intro.asp>
Бончев, Б. 2015. *XML технологии*. София: УИ „Св. Климент Охридски"

Софтуерни стандарти и препоръки:

Cone, Matt. 2024. Markdown Guide. <https://www.markdownguide.org/>.
CSS. Casscading Style Sheets. <https://www.w3.org/Style/CSS/Overview.en.html>
CSS Writing Modes Level 3. W3C Recommendation, 10 December 2019. <https://www.w3.org/TR/css-writing-modes-3/>
FoLiA. Format for Linguistic Annotation. <https://proycon.github.io/folia/>.
HTML. HyperText Markup Language. <https://html.spec.whatwg.org/>
JSON. JavaScript Object Notation. <https://www.json.org/json-en.html>.
LaTeX Core Documentation. <https://www.latex-project.org/help/documentation/>
Ontologies of Linguistic Annotation <https://acoli-repo.github.io/olia/>
OWL. Web Ontology Language. <https://www.w3.org/2001/sw/wiki/OWL>
TEI Consortium, eds. 2023. *Guidelines for Electronic Text Encoding and Interchange.* <http://www.tei-c.org/P5/>.
The Unicode Consortium. *The Unicode Standard*. <https://www.unicode.org/versions/latest/>
Universal Dependencies. <https://universaldependencies.org/>
XML. *Extensible markup Language*. <https://www.w3.org/XML/>
XSL. *The Extensible Stylesheet Language Family*: XSL Transformations (XSLT), The XML Path Language (XPath), XSL Formating Objects (XSL-FO) <https://www.w3.org/Style/XSL/>

Софтуерни приложения:

CLarK. An XML Based System For Corpora Development. <http://bultreebank.org/clark/> .
ConTeXT. The TeX macro language and interpreter. <https://wiki.contextgarden.net/Main_Page>.
CTAN. Comprehensive TeX Archive Network <https://ctan.org/>.
eXist-db. The Open Source Native XML Database. <http://exist-db.org/>.
LuaTeX. The extended version of pdfTeX using Lua as an embedded scripting language. <https://www.luatex.org/>.
MindForger. Thinking Notebook. Large Language Models from OpenAI integration. <https://www.mindforger.com/>.
NotePad++. Source code editor (Windows). <https://notepad-plus-plus.org/>.
Overleaf. Collaborative online LaTeX editor. <https://www.overleaf.com/>.
OWLGrEd. Ontology editor. <http://owlgred.lumii.lv/>.
Oxygen. XML Editor. <https://www.oxygenxml.com/>.
Protégé. Ontology editor. <https://protege.stanford.edu/>.
Pulsar. A Community-led Hyper-Hackable Text Editor. <https://pulsar-edit.dev/>.
Scite. Code editor. <https://www.scintilla.org/SciTE.html>.
StackEdit. In-browser Markdown editor <https://stackedit.io/>.
TEI Publisher. The Instant Publishing Toolbox. <https://teipublisher.com/>.
TeX / LaTeX дистрибуции: TeX Live (Linux, MacOS, Unix). <https://ctan.org/pkg/texlive-repo> ; MikTeX (Windows, Linux, MacOs). <https://miktex.org/>.
Visual Studio Code. Code editing. <https://code.visualstudio.com/>.

Дата: 29 март 2024 г.

Съставил:
проф. д-р Андрей Бояджиев