ВСТУП ДО КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ

(INTRODUCTION TO COMPUTATIONAL LINGUISTICS)

Силабус курсу, осінній семестр 2019-2020 н. р.

Загальна інформація про курс

ВИКЛАДАЧ

Старко
Василь

v.starko@ucu.edu.ua

1. Загальна інформація про курс:
   a. Вступ до комп'ютерної лінгвістики (Introduction to Computational Linguistics)
   b. Старко Василь
   c. Кандидат філологічних наук, доцент кафедри філології Гуманітарного факультету Українського католицького університету. Дисертацію захистив за спеціальністю "Загальне мовознавство" в Інститут мовознавства ім. О.О. Потебні НАНУ. стипендіат програми Фулбрайта для молодих викладачів та дослідників (Державний університет Нью-Йорку в Олбані, США) та програми академічних обмінів Erasmus Mundus (дослідник на посаді post-doc, Гумбольдтський університет Берліна, ФРН). Керівник корпусної групи БрУК (https://r2u.org.ua/corpus), співавтор Великого електронного словника української мови (https://github.com/brown-uk/dict_uk), засобу перевіряння орфографії, граматики й стилю «Правописник LanguageTool» (https://languagetool.org/uk/), словникових вебсайтів http://e2u.org.ua/ та http://r2u.org.ua/. Володіє кількома іноземними мовами. Викладає університетські дисципліни англійською мовою. Має досвід викладання понад 6 років у вищих навчальних закладах. Автор понад 70 наукових праць в Україні та за кордоном.
   d. v.starko@ucu.edu.ua , роб. тел. (032)240-36-42
   e. https://cms.ucu.edu.ua/course/view.php?id=2100
   f. Онлайн-консультації: через форум курсу в CMS УКУ та електронну пошту.

2. Коротка анотація. Курс має практичне й сучасне спрямування та охоплює український і закордонний доробок у галузі комп'ютерної лінгвістики. Курс викладається англійською мовою за винятком двох тем, пов'язаних із українськими комп'ютернолінгвістичними ресурсами.

3. Мета та цілі курсу – ознайомити студентів з основними поняттями, напрямами й підходами в комп'ютерній лінгвістиці, а також практичними засобами автоматичного опрацювання мовних даних. Завдання курсу:

   - дати студентам необхідні теоретичні знання
   - навчити базових інструментів, методів й алгоритмів комп'ютерної лінгвістики
   - навчити грамотно користуватися практичними системами й засобами автоматичного опрацювання мовних даних
   - навчити аналізувати й порівнювати отримані дані
   - вказати шляхи для самостійного поглиблення знань

4. Формат курсу – змішаний, включає лекції, практичні заняття та консультації, має супровід у систему CMS УКУ.

5. Результати навчання

За результатами курсу студенти будуть:

знати:

- основні поняття комп'ютерної лінгвістики
- теоретичні основи й підходи до автоматичного опрацювання природної мови на різних рівнях
- основні напрями та сучасний стан комп'ютерної лінгвістики в Україні й світі
- основні комп'ютерні системи для роботи з мовою

вміти:

- пояснити застосування й роботу базових алгоритмів комп'ютерної лінгвістики
- застосовувати комп'ютерні засоби в лінгвістичних дослідженнях
- оцінювати й порівнювати результати роботи цих засобів

6. Обсяг курсу – 48 год. аудиторних (14 год. лекційних і 34 год. практичних), 60 год. самостійної роботи.

7. Ознаки курсу:

Рік викладання	**семестр**	**спеціальність**	**Курс (рік навчання)**	**Нормативний\ вибірковий**
2019	1	філологія	3	нормативний (Н)

8. Пререквізити: Базові знання з лінгвістики, володіння англійською на рівні B2.

9. Технічне й програмне забезпечення /обладнання. Мультимедійний проєктор та інтернет-зв'язок для проведення занять. Кожен студент повинен мати змогу користуватися персональним комп'ютером — ноутбуком або стаціонарним комп'ютером.

10. Політики курсу — суворе дотримання принципів академічної доброчесності, а у випадку їх порушення — реагування відповідно до Положення.

11. Схема курсу

Матеріали розміщено на сторінці курси в CMS UCU

Тиж./акад. год.	**Тема, план, тези**	**Форма заняття/ Формат**	**Література/ Ресурси в інтернеті**	**Завдання, год**	**Вага оцін-ки**	**Термін виконання**
Тиждень 1 2 год	Introduction. Projects of the r2u group for Ukrainian. Dictionary websites. Spellchecker. VESUM. Ukrainian Brown Corpus.	лекція F2F	4, с. 54-74; 5, с. 4-22, 192-205; 6; https://github.com/brown-uk	Переглянути презентацію, 2 год.		
Тиждень 2 4 год	Exploring web resources for Ukrainian	практична F2F	r2u.org.ua, r2u.org.ua/vesum, e2u.org.ua, languagetool.org/uk	Навчитися користуватися онлайн-ресурсами, 4 год.	8	1 тиждень
	Homework: dictionary search in r2u.org.ua; checking texts using languagetool.org/uk	домашнє завдання		Опрацювання літератури, виконання практичних завдань 4 год.		

Тиждень 3 2 год	Morphological analysis. Morphological word. POS tags for English and Ukrainian. Universal POS tags. Tokenization, stemming, lemmatization. Stemmers for English and Ukrainian.	лекція F2F	1, с. 124-134; 4, с. 127-137; 7, Ch. 8 (8.1-8.3); http://www.senyk.poltava.ua/projects/ukr_stemming/stemming_about.html, http://snowballstem.org/demo.html, https://github.com/brown-uk/dict_uk	Переглянути презентацію, 2 год.		
Тиждень 4 4 год	Using VESUM's web interface. Using LanguageTool for morphological analysis.	практична F2F	6; r2u.org.ua/vesum, languagetool.org	Робота з онлайн-інструментами з супроводом викладача		
	Finding lemmas for Ukrainian words. Identification of grammatical tags for Ukrainian and English wordforms.	домашнє завдання	r2u.org.ua/vesum, languagetool.org	4 год.	8	1 тиждень
Тиждень 5 2 год	GRAC: a corpus of Ukrainian	лекція F2F	uacorpus.org	Переглянути презентацію, 2 год.		
Тиждень 6 4 год	Learning different types of searches and other functionality of the GRAC corpus	практична F2F	http://www.parasolcorpus.org/bonito/run.cgi/first_form	Робота з онлайн-інструментами з супроводом викладача		
	Письмове контрольне опитування	опитування			10	
	Form queries to GRAC using the Corpus Query Language (CQL)	домашнє завдання	http://www.parasolcorpus.org/bonito/run.cgi/first_form	4 год.	8	1 тиждень
Тиждень 7 2 год	Corpora and corpus linguistics	лекція F2F	3, с. 8-27	Переглянути презентацію, 2 год.		
Тиждень 8 4 год	Survey of key corpora for Ukrainian and English (Brown, BNC, COCA, and others) and exploring their functionality	практична F2F	http://www.mova.info/corpus.aspx http://www.mova.info/pcorpus_UA.aspx https://mova.institute/ https://www.english-corpora.org/	Робота з онлайн-інструментами з супроводом викладача		
Тиждень 9 2 год	Syntactic parsing. Types of grammars, constituency, CFGs and their components, constructing a CFG.	лекція F2F	2, с. 95-144; 4, с. 139-143; 7, Ch. 12	Переглянути презентацію, 2 год.		
Тиждень 10 4 год	Parsing using CFGs. CFG rules for different sentence structures. Modifying and creating CFGs.	практична F2F	7, Ch. 12	Робота в аудиторії з супроводом викладача		
	Письмове контрольне опитування	опитування			10	
	Building parse trees using a given CFG for a set of sentences.	домашнє завдання		2 год.	8	1 тиждень
Тиждень 11 2 год	Dependency parsing. Shallow parsing. Types of notation. Dependency parsers.	лекція F2F	7, Ch. 13 2, с.117-123	Переглянути презентацію, 2 год.		
Тиждень 12 4 год	Using online tools for dependency parsing for Ukrainian and English	практична F2F	http://nlp.stanford.edu:8080/parser, https://mova.institute/%D0%B0%D0%BD%D0%B0%D0%BB%D1%96%D0%B7%D0%B0%D1%82%D0%BE%D1%80	Робота з онлайн-інструментами з супроводом викладача		
	Building dependency trees for a set of sentences.	домашнє завдання		2 год.	8	1 тиждень
Тиждень 13 2 год	Machine translation. Applications and approaches: rule-based, statistical, and neural MT. Advantages and disadvantages of NMT.	лекція F2F		Переглянути презентацію, 2 год.		
Тиждень 14 4 год	Overview of Neural Machine Translation (NMT). Evaluating MT.	практична F2F	Відео лекцій проф. К. Меннінга про НМТ https://youtu.be/IxQtK2	Перегляд і обговорення відео		

			SjWWM?list=PL3FW7Lu3i5Jsnh1rnUwq_TcylNr7EkRe6&t=893 https://youtu.be/6_MO12fPC-0?list=PL3FW7Lu3i5Jsnh1rnUwq_TcylNr7EkRe6&t=3654			
	Письмове контрольне опитування	опитування			10	
Тиждень 15 4 год	Semantic relations and WordNet. WordNet as a lexical database and its structure. Word relations, senses, and disambiguation. Word similarity.	практична F2F	2, с. 189-202; 7, Ch. 19, 7, Appendix C., pp. 1-6 https://wordnet.princeton.edu/, http://wordnetweb.princeton.edu/perl/webwn	Робота в аудиторії з супроводом викладача		
	Іспит	Іспит F2F			30	

12. Система оцінювання та вимоги

Тип завдання	**Кількість**	**Бали**
Контрольні опитування	3	**30**
Домашні завдання	5	**40**
Іспит	1	**30**

Література:

1. Волошин В.Г. Комп’ютерна лінгвістика: Навч. посібник, 2004.
2. Дарчук Н.П. Комп’ютерна лінгвістика (автоматичне опрацювання тексту), 2008
3. Жуковська В.В. Вступ до корпусної лінгвістики, 2013.
4. Карпіловська Є.А. Вступ до прикладної лінгвістики: комп’ютерна лінгвістика, 2006.
5. Перебийніс В.І., Сорокін В.М. Традиційна та комп'ютерна лексикографія, 2009.
6. Старко В. Комп’ютерні лінгвістичні проекти гурту r2u: стан та застосування // Українська мова. — 2017. — № 3. — С. 86–100.
7. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd ed., 2008; 3rd ed., draft (https://web.stanford.edu/~jurafsky/slp3/)