Утвърдил: ……..

Декан

Дата .....

СОФИЙСКИ УНИВЕРСИТЕТ “СВ. КЛИМЕНТ ОХРИДСКИ”

Факултет: .....

Специалност: *(код и наименование)*

.....

Магистърска програма: *(код и наименование)*

С	Л	Б	З	0				

Компютърна лингвистика. Интернет технологии в хуманитаристиката
(задочна форма на обучение)

УЧЕБНА ПРОГРАМА

Дисциплина:

(код и наименование) **Основи на трибанкинга**

Преподавател: проф. д-р Петя Осенова

Асистент:

Учебна заетост	Форма	Хорариум
Аудиторна заетост	Лекции	20 ч.
	Семинарни упражнения	
	Практически упражнения (хоспетиране)	
Обща аудиторна заетост		**20**
Извънаудиторна заетост	Реферат	
	Доклад/Презентация	30 ч.
	Научно есе	
	Курсов учебен проект	
	Учебна екскурзия	
	Самостоятелна работа в библиотека или с ресурси	40 ч.
Обща извънаудиторна заетост		**70**
ОБЩА ЗАЕТОСТ		**90**
Кредити аудиторна заетост		**1**
Кредити извънаудиторна заетост		**2**
ОБЩО ЕКСТ		3

№	Формиране на оценката по дисциплината[1]	% от оценката
1.	Участие в тематични дискусии в часовете	50 %
2.	Изпит	50 %

Анотация на учебната дисциплина:

Курсът запознава студентите с основите на трибанкинга. В този процес се включват синтактично анотирани ресурси (трибанки) и автоматични анализатори (парсери), обучени върху тях, за правене на автоматичен синтактичен анализ (парсинг).

Проследява се процесът от създаването на трибанки (история, видове представяния и начини на създаване) към използването им (търсене в тях) и приложенията им (за обучение на парсери, в машинния превод, за извличане на знание).

Курсът също така обръща внимание на добрите практики в момента както в световен мащаб, така и за българския език.

Предварителни изисквания:

Необходими са основни знания по традиционна българска граматика, и по-специално синтаксис; знания по XML.

Предимство са знанията в следните области: българска формална граматика, и по-специално формален синтаксис; лингвистични подходи към моделирането на езика; автоматичен анализ на данни с използване на лингвистично знание.

Очаквани резултати:

Студентите ще имат знание за това какво представляват важни ресурси като трибанките и парсерите (видове, кодиран езиков модел и др.), кои са трудностите при създаването на трибанка и обучението на парсер; как могат да се използват трибанките и парсерите за определени задачи при обработката на естествен език.

Студентите ще имат уменията: да създадат ресурс от типа „трибанка“, да търсят в подобен ресурс и да преценяват кой езиков модел е най-подходящ за конкретна

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

задача. Също така те ще могат да използват някои варианти на вече имплементирани парсери и да се запознаят с възможностите за ползване и на други имплементации в бъдеще в зависимост от знанията си по програмиране.

Учебно съдържание

№	Тема:	Хорариум
1.	Същност и история на т.нар. *трибанки* (treebanks)	1 ч.
2.	Конституентни и депендентни модели на трибанките: предимства и недостатъци. Подходи с оглед на задачите на компютърните приложения (машинен превод, парсинг и др.)	1 ч.
3.	Трансформации между двата модела: добрите практики и проблемът със загуба на знание	1 ч.
4.	Видове трибанки (според лингвистичната теория, начина на създаване, целта за създаване и др.)	2 ч.
5.	Класическата трибанка за английския език: Penn Treebank – история, схема на кодиране, важност, използване, особености	1 ч.
6.	Трибанката за българския език BulTreeBank: – история, схема на кодиране, важност, особености	2 ч.
7.	Трибанката за българския език BulTreeBank: трансформации и използване	2 ч.
8.	Проектът за Универсални зависимости: същност и значимост.	1 ч.
9.	Проектът за Универсални зависимости и кодиране на морфологията за много езици	1 ч.
10.	Проектът за Универсални зависимости и кодиране на синтаксиса за много езици	2 ч.
11.	Проектът за Универсални зависимости и кодиране на явления между морфологията, синтаксиса и семантиката за много езици	1 ч.
12.	Видове търсене в трибанки. Програми за търсене.	2 ч.
13.	Компютърни приложения с използване на трибанки: парсинг, извличане на знание	1 ч.
14.	Добрите практики в парсинга (с оглед на т.нар. parsebanks, cashbanks)	1 ч.
15.	Добрите практики за парсинга в българския език.	1 ч.

Конспект за изпит

№	Въпрос

1.	Формално определение на понятието „трибанка“. Прилики и разлики с понятието граф
2.	Особености на конституентните трибанки
3.	Особености на дependентните трибанки
4.	Видове трибанки според различни критерии
5.	Трансформация между конституентна и дependентна трибанка в двете посоки: предимства и слабости
6.	Penn Treebank – същност, значение за трибанкинга, използване, особености
7.	Трибанката за българския език BulTreeBank – същност, особености, значение, разновидности, приложение
8.	Проектът за Универсални зависимости: същност, значение, кодиране на езиковата информация, използване
9.	Видове търсене в трибанки. Сложност на търсенето.
10.	Компютърни приложения с използване на трибанки (вкл. parsebanks, cashbanks)
11.	Добри практики в парсинга – за българския език и за други езици

Библиография

Основна:

1. Осенова и Симов 2007: П. Осенова и Кирил Симов. *Формална граматика на българския език.* ИПОИ, София. Свободно достъпна на следния линк: http://bultreebank.org/wp-content/uploads/2017/04/FormalGrammarBG.pdf
2. Осенова 2016: П. Осенова. *Граматическо моделиране на българския език (с оглед на автоматичната обработка на естествен език)*. Парадигма, София. Свободно достъпна на следния линк: https://www.academia.edu/30647536/Граматическо_моделиране_на_българския_език
3. Abeille (ed.) 2003: Ann Abeille. *Treebanks: Building and Using Parsed Corpora.* Text, speech and language technology, vol. 20. Dodrecht: Kluwer Academic Publishers.
4. de Marneffe et al. 2021: Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, Daniel Zeman (2021): *Universal Dependencies.* In: Computational Linguistics, ISSN 1530-9312, vol. 47, no. 2, pp. 255-308.
5. Osenova and Simov 2015: P. Osenova and Kiril Simov. *Universalizing BulTreeBank: a Linguistic Tale about Glocalization*. In: Proceedings of BSNLP 2015, Hissar, Bulgaria, pp. 81–89.

Линкове:

1. Универсални зависимости: https://universaldependencies.org/
2. Парсер за българския език в системата CLaRK: http://bultreebank.org/en/clark/bulgarian-nlp-pipeline-in-clark-system/

Допълнителна:

1 Статии от Семинара по универсални зависимости за 2019 и 2020 г.
https://aclanthology.org/volumes/2020.udw-1/
https://aclanthology.org/W19-80.pdf

2. Статии от Семинара по трибанки и лингвистични теории за 2020 г.
https://aclanthology.org/2020.tlt-1.0.pdf

3. Статии от Семинара по граматики на зависимостите за 2019 г.:
https://aclanthology.org/volumes/W19-77/

Дата: 22.02.2022 г.

Съставил:
проф. д-р Петя Осенова