



ЦЕНТЪР ЗА ОБУЧЕНИЕ – БАН

1000 София
ул. „Сердика“ № 4
<http://edu.bas.bg>

email: tdc-phd@cu.bas.bg
тел.: 02 987 31 67
02 979 52 60

Основна информация:

Име на курса: Езикови ресурси и технологии в хуманитарните изследвания

Лектор: .проф. д-р Светла Коева

Телефон: +35929722302

Имейл: svetla@dcl.bas.bg

Хорариум: 30 часа

Анотация (до 150 думи):

Курсът запознава докторантите с фундаментални теоретични изследвания и приложения, свързани с езиковите ресурси и технологии – от корпуси и анотирани корпуси до големи езикови модели и инструменти за анализ и синтез на реч. Материалът е представен достъпно и не изисква предварителни знания по математика или компютърни науки.

Докторантите ще получат теоретични и приложни знания за езикови ресурси и технологии, приложими в хуманитарните изследвания. Те ще придобият практически умения за работа с основни езикови ресурси: лексикално-семантични бази от данни и Уърднет, ресурси от типа Фреймнет, графи за представяне на знание и онтологии, инструменти за разпознаване на именувани обекти, компютърно подпомогнат превод, и ще научат как могат да ги използват в различни хуманитарни дисциплини.

Тематично съдържание на курса (кратко описание по теми или модули):

#	Тема	Часове
1	Корпуси – големи колекции от писмени или аудиотекстове, използвани за лингвистичен, литературен или друг тип анализ; текстовете могат да бъдат с общо предназначение или специфични за дадена област (например история, право, литература).	2
2	Анотирани корпуси – корпуси, обогатени с морфологична, синтактична и семантична анотация, позволяваща систематичен анализ на граматически структури в различни езици и времеви периоди.	2
3	Лексикално-семантични бази от данни и Уърднет – структурирано лексикално знание, представящо семантични отношения като синонимия, хипонимия и меронимия (например Принстънският уърднет, Българският уърднет и др.).	2
4	Ресурси от типа на Фреймнет – лексикално-семантични бази от данни, базирани на семантиката на концептуални фреймове, представящи връзките между лексикалните единици и концептуалните структури, които те предизвикват.	2



ЦЕНТЪР ЗА ОБУЧЕНИЕ – БАН

1000 София
ул. „Сердика“ № 4
<http://edu.bas.bg>

email: tdc-phd@cu.bas.bg
тел.: 02 987 31 67
02 979 52 60

5	Графи на знание и онтологии – структурирано представяне на обекти и техните връзки, позволяващо семантична интеграция и връзки между културно-исторически и биографични данни.	2
6	Инструменти за разпознаване на именувани обекти – автоматично идентифицират и класифицират препратки към лица, места, организации и дати, подпомагайки текстовия анализ в журналистически, исторически и литературни изследвания.	2
7	Автоматично извличане на термини – методи и инструменти за идентифициране на специфични термини за дадена област, подпомагащи изследванията в хуманитарните науки и превода.	2
8	Инструменти за компютърно подпомогнат превод – подпомагат експертния превод чрез функции като преводна памет, управление на терминологията, инструменти за подравняване и автоматизация на процесите.	2
9	Големи езикови модели – предварително обучени невронни модели, използвани за задачи като генериране на текст, автоматично резюмиране, отговаряне на въпроси и семантично търсене в големи колекции от данни.	2
10	Корпуси с устна реч и инструменти за анализ и синтез на реч – записи на реч, съчетани с транскрипции, приложими за изучаване на диалекти, устна комуникация и науки като социолингвистика.	2
Общо		30

Форми на обучение и оценяване:

Курсът включва лекции и практически упражнения. Лекциите представят теоретичните основи и ключовите понятия на всяка тема, а упражненията дават възможност на докторантите да прилагат инструменти и методи за обработка на езикови данни, като затвърждават знанията си чрез непосредствена работа с материала.

Оценяването се осъществява чрез курсова работа. Докторантите разработват индивидуален проект, в който прилагат един или няколко от разгледаните в курса методи за решаване на самостоятелно избран проблем, свързан по възможност с тяхното дисертационно изследване. Проектът се придружава от писмена курсова работа, в която докторантът представя използваните методи, начина на тяхното прилагане и критичен анализ на получените резултати.

Компетентности, придобити в резултат на обучението (3-5 точки):



ЦЕНТЪР ЗА ОБУЧЕНИЕ – БАН

1000 София
ул. „Сердика“ № 4
<http://edu.bas.bg>

email: tdc-phd@cu.bas.bg
тел.: 02 987 31 67
02 979 52 60

След успешно завършване на курса студентите ще могат:

Да подбират и оценяват езикови ресурси, включително корпуси, анотирани корпуси, лексикално-семантични бази от данни, уърднети, ресурси като Фреймнет и други, за приложение в хуманитарните науки.

Да прилагат основни техники (като разпознаване на именувани обекти, автоматично извличане на термини и изграждане на онтологии) към реални изследователски задачи в хуманитарните науки.

Да изработят критично мнение за големите езикови модели и техните приложения, включително генериране на текст, семантично търсене и отговаряне на въпроси, и да са подготвени да оценят потенциала и ограниченията им в научните изследвания.

Да интегрират компютърнолингвистични методи в собствените си изследователски програми, преодолявайки разликата между лингвистиката, дигиталните хуманитарни науки и подходите, основани на данни, за да се усъвършенства и подпомогне изучаването на езика, културата и историята.

Литература:

. Almeman, F. Y., Schockaert, S., & Espinosa Anke, L. WordNet under scrutiny: Dictionary examples in the era of large language models. In *Proceedings of LREC-COLING 2024*. European Language Resources Association, Torino, 2024, pp. 17683–17695.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33. Curran Associates, Red Hook, NY, 2020, pp. 1877–1901.

Carta, S., Giuliani, A., Piano, L., Podda, A. S., Pompianu, L., & Tiddia, S. G. Knowledge graph construction: Extraction, learning, and evaluation. *Applied Sciences*, 15(7), Article 3727. MDPI, Basel, 2025.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. Universal Dependencies. *Computational Linguistics*, 47(2) MIT Press, Cambridge, MA, 2021, pp. 255–308.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics, Minneapolis, 2019, pp. 4171–4186.

Ehrlinger, L., & Wöß, W. Towards a definition of knowledge graphs. In *Proceedings of the Workshop on Semantic Web Technologies for Intelligent Information Systems (SEMANTICS)*. CEUR Workshop Proceedings, Potsdam, 2016.

Ehrmann, M., Romanello, M., & Clemenide, S. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2), Article 40. ACM Press, New York, 2023.

Fellbaum, C. (Ed.). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

Fillmore, C. J., & Baker, C. F. Frame semantics for text understanding. In *Proceedings of the WordNet and Other Lexical Resources Workshop*. Association for Computational Linguistics, Pittsburgh, 2001.

Jurafsky, D., & Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed. draft). Prentice Hall, Upper Saddle River, NJ, 2026.

Koeva, S. (ed.). *Universality of semantic frames and language-specific Bulgarian data*. Language Science Press, 2025.



ЦЕНТЪР ЗА ОБУЧЕНИЕ – БАН

1000 София
ул. „Сердика“ № 4
<http://edu.bas.bg>

email: tdc-phd@cu.bas.bg
тел.: 02 987 31 67
02 979 52 60

-
- MacWhinney, B. *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Lawrence Erlbaum Associates, Mahwah, NJ, 2000.
- McEnery, T., & Wilson, A. *Corpus Linguistics: An Introduction* (2nd ed.). Edinburgh University Press, Edinburgh, 2001.
- McEnery, T., Xiao, R., & Tono, Y. *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge, London, 2006.
- Miller, G. A. WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39–41. ACM Press, New York, 1995.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Honolulu, 2023, pp. 28492–28518.
- Rigouts Terryn, A., Hoste, V., & Lefever, E. Tagging terms in text: A supervised sequential labelling approach to automatic term extraction. *Terminology*, 28(1), pp. 157–194. John Benjamins, Amsterdam, 2022.
- Rothwell, A., Way, A., & Youdale, R. *Computer-Assisted Literary Translation*. Routledge, London, 2023.
- Schweinberger, M., & Haugh, M. Reproducibility, replicability, and robustness in corpus linguistics. *International Journal of Corpus Linguistics*, 30(2). John Benjamins, Amsterdam, 2025, pp. 119–129.
- Sinclair, J. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford, 1991.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, 30. Curran Associates, Red Hook, NY, 2017, pp. 5998–6008.
- Коева, Св., Д. Благоева (ред.). *Езикови ресурси и технологии за български език*. София: Академично издателство „Проф. “Марин Дринов”, 2014, 310 с. ISBN: 978-954-322-797-6.

Допълнителна информация (по желание) (например: специални изисквания, лабораторно оборудване, предварителни знания):

Няма.

Курсът е предназначен предимно за езиковеди, но може да бъде посещаван от специалисти от други хуманитарни дисциплини, математици и информатици., които искат да разширят своята перспективи.