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

Basic Information:

Course Title: Language Resources and Technologies in Humanities Research
Lecturer: Prof. Svetla Koeva
Phone: +35928722302
Email: svetla@dcl.bas.bg
Total Teaching Hours: 30

Annotation (up to 150 words)

The course *Language Resources and Technologies in Humanities Research* introduces doctoral students to foundational theoretical research and applications, ranging from large-scale corpora and annotated treebanks to large language models and speech recognition and generation tools. The material is presented accessibly and requires no prior background in mathematics or computer science.

Students will develop a solid understanding of language resources and technologies as both theoretical and applied instruments for humanities inquiry. They will gain practical familiarity with the key resources and methods that define the field, including lexical-semantic databases and wordnets, FrameNet resources, knowledge graphs and ontologies, and named entity recognition tools, and will learn how these are used in linguistic, literary, historical, and cultural research. The course also covers practical workflows in automatic terminology extraction and computer-assisted translation (CAT tools), equipping students with skills directly applicable to multilingual and digital humanities projects.

Course content (brief description by topics or modules)

#	Topic	Hours
1	**Corpora** – Large collections of written or spoken texts used for linguistic, literary, and cultural analysis; these may be general-purpose or domain-specific (e.g. historical, legal, literary).	3
2	**Annotated treebanks** – Corpora enriched with morphological, syntactic, and semantic annotation, enabling systematic analysis of grammatical structures across languages and time periods.	3
3	**Lexical-semantic databases and wordnets** – Structured repositories of lexical knowledge that encode semantic relations such as synonymy, hyponymy, and meronymy (e.g. Princeton WordNet, BulNet)**.**	3
4	**FrameNet resources** – Lexical-semantic databases based on frame semantics, documenting the relationships between lexical units and the conceptual structures (frames) they evoke; widely used for semantic role labelling and contextual meaning analysis.	3
5	**Knowledge graphs and ontologies** – Structured representations of entities and their relationships, enabling semantic integration and cross-referencing of cultural, historical, and biographical data.	3
6	**Named entity recognition (NER) tools** – Systems that automatically identify and classify references to entities such as persons, places, organisations, and dates, supporting large-scale text analysis in historical and literary research.	3

7	**Automatic terminology extraction** – Methods and tools for identifying domain-specific terms from corpora, supporting knowledge organisation, digital humanities research, and translation workflows.	3
8	**CAT tools (Computer-Assisted Translation tools)** – Software environments that support human translation through features such as translation memory, terminology management, alignment tools, and workflow automation, improving consistency and efficiency in multilingual contexts.	3
9	**Large language models (LLMs)** – Pretrained neural models used for tasks such as text generation, summarisation, question answering, and semantic search across large textual datasets.	3
10	**Speech corpora and transcription tools** – Collections of recorded speech paired with transcriptions, essential for the study of dialects, oral history, sociolinguistics, and endangered languages.	3
Total		**30**

Teaching and assessment methods

The course is delivered through lectures and hands-on practical sessions. Lectures introduce the theoretical foundations and key concepts of each topic, while practical sessions allow students to apply computational tools and methods to real linguistic data, consolidating their understanding through direct engagement with the material.

Assessment is by coursework. Students must complete an individual project in which they apply one or more of the methods covered in the course to a problem of their choice, ideally related to their doctoral research. The project is accompanied by a written report demonstrating the student's understanding of the relevant methods, their implementation, and a critical discussion of the results.

Competencies acquired as a result of training (3–5 points)

Upon successful completion of the course, students will:

Be able to identify, evaluate, and select language resources, including corpora, annotated treebanks, lexical-semantic databases, wordnets, and FrameNet resources, for use in humanities research contexts such as literary, historical, and cultural analysis.

Be able to apply core computational techniques, including named entity recognition, automatic terminology extraction, and knowledge graph construction, to real-world research tasks in linguistics and the humanities.

Develop an informed and critical perspective on large language models and their applications, including text generation, semantic search, and question answering, and be equipped to assess their potential and limitations within academic research.

Be prepared to integrate computational and digital methods into their own research agendas, bridging the gap between linguistics, digital humanities, and data-driven approaches to the study of language, culture, and history.

Literature:

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., & Clematide, 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.

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.

Additional information (optional) (e.g., special requirements, laboratory equipment, prior knowledge)

There are no special requirements.

The course is primarily designed for linguists, but is equally open to researchers from other humanities disciplines, as well as mathematicians and computer scientists seeking to broaden their perspective on language as a computational and cultural phenomenon.