

Course descriptions

Course name: Basics of language technology	Credits: 6
Class type: lecture/lab, hours per week: 3/2	
Type of the exam: oral exam	
Semester:	
Prerequisites (if exist): Theory of Digital Computation	
Course description:	
<p>Characters, coding standards. ABCs and order. Representing the form and content of the text. Main languages. Text content representation. TEI. Coding dictionaries and corpora.</p> <p>Research issues related to the computer representation of natural languages. Utilizing formal languages for the research of natural languages. Qualitative examination of language models: accuracy and coverage, over- and undergeneration. Techniques for modeling natural languages. Encyclopedias. Corpus linguistics. Disambiguation. Finite automaton, finite state transducer. RTN, ATN. Morphology. Linguistic foundations. FSA. Bi-level morphology. Special morphological problems of the Hungarian language. Unification-based morphological model: investigating the possibility of unification and feature structure. Development of morphological systems. Minimum grammar, adding by analogy. Application of morphological analysing programs: spell-check, (conjugation) thesaurus, searching by finding the root of the word. Grammars of natural languages. Important linguistic phenomena. Application of corpus linguistics in syntactic research. Treebank. Syntactic analysis algorithms: symbolic (finite: RTN/ATN, context-free) and probability algorithms (disambiguation of analysis). Feature structures and unification: sample-based and unification process of syntactic analysis (built on formal grammar or lexicon). Transition between the two. Knowledge representation: formal semantics, taxonomies, ontology - NLP-encyclopedia. Word-sense disambiguation: the word-sense model.</p> <p>Issues of translation. Relationship of different languages. Linguistic phenomena from the point of view of translation. Supporting human translation: intelligent dictionaries, tools assisting understanding. Terminology databases, administrative tools.</p> <p>Parallel corpora. Synchronization of texts. Translation memories. Similarity search. Machine translation. Dictionaries for machine translation. Representation of dictionaries. Translating procedures: syntactic/lexical, interlingual, transfer. Statistic and rule-based algorithms. Further issues of processing natural languages: modeling pragmatics and discourse. Meeting points of speech processing and language technology.</p>	
Required reading:	
<p>Dale, R., H. Moisl, H. Somers (eds.) Handbook of Natural Language Processing. Marel Dekker (2000); Jurafsky, D. & J. H. Martin. Speech and Language Processing. Prentice Hall (2000); Wilks, Y.A., B.M. Slator & L.M.Guthrie. Electric Words (Dictionaries, Computers and Meanings). MIT Press (1996).</p>	
Recommended reading:	
<p>Allen, J. F.: Natural Language Understanding, The Benjamin/Cummings Publishing Company, Menlo Park, California (1988); Hutchins, W. J., H. L. Somers, An Introduction to Machine Translation, Academic Press, San Diego (1992).</p>	
Lecturer (name, position, degree): Dr. Gábor Prószéky, professor, DSc.	
Additional lecturers, if exist (name, position, degree): Dr. Gábor Prószéky, professor, DSc	