TITLE: **Computational linguistics**

LECTURERS:
Prof.dr.sc. Mario Essert, *Faculty of Mechanical Engineering and Naval Architecture, Zagreb,* messert@fsb.hr
Dr.sc. Kristina Štrkalj Despot, *Institute of Croatian Language and Linguistics, Zagreb,* kdespot@ihjj.hr

*Abstract*

According to Roland Hausser[1] there are three basic approaches to natural language:

(i) *traditional grammar* – uses the method of informal classification and description based on tradition and experience,
(ii) *theoretical linguistics* – uses the method of mathematical logic to describe natural languages by means of formal rule systems intended to derive all and only the well-formed expressions of a language,
(iii) *computational linguistics* – combines the methods of traditional grammar and theoretical linguistics with the method of effectively verifying explicit hypotheses by implementing formal grammars as efficient computer programs and testing them automatically on realistic amounts of real data.

Despite their different methods, goals, and applications, the three variants of language science described divide the field into the same components, namely phonology, morphology, lexicon, syntax, semantics, and the additional field of pragmatics. Formal language theory (ii and iii) works with mathematical methods which treat the empirical contents of grammatical analysis and the functioning of communication as neutrally as possible – a language become a set of finite word sequences (*free monoid*).

On the other side, since natural language processing (NLP) is a subfield of artificial intelligence (AI) and traditional linguistics, it is concerned with the structure of text and algorithms that extract meaningful information from text. A well-known and effective technique is the vector space model that represents documents as a matrix of n × m dimensions (Salton, Wong & Yang, 1975). A distance metric can then be used as a function of the matrix to compute similarity between documents. This kind of machine learning algorithms favor statistical approaches and is related with fields like text mining, text categorization and information retrieval. This vector space model is fundamental to many tasks in natural language processing and machine learning, from search queries to classification and clustering.

This lecture will present the theoretical and practical steps in the process of extracting information from raw data (e.g., finding words and sentences in a string of characters), identifying and generating word types (e.g., nouns, verbs, adjectives), recognizing the logical laws in sentences, and deriving meaning from the relations between words. Finally, some programs with vector space model in linguistics will be shown.

[1] Foundations of Computational Linguistics, Human-Computer Communication in Natural Language, Third Edition, Springer-Verlag, 2014.