Annotation on PhD Thesis

Title: Automation of the process of computational linguistic resources creation
Author: Mircea Petic
Place of defence: Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova, Chisinau
Date of defence: 22 January 2012
Speciality: 01.05.01 – Theoretical foundation of computer science; programming

The thesis was elaborated at the Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova, Chisinau, in 2011. The thesis is written in Romanian and contains introduction, three chapters, general conclusions and recommendations, bibliography of 200 titles, 14 appendices, 133 pages of the main text, 15 figures, and 44 tables. The results are published in 27 scientific papers.

Keywords: computational linguistic resources, derivational algorithm, affix, prefix, suffix, words segmentation, vocalic/consonantal alternations, automatic derivative generation, generative derivational mechanisms.

The study in this thesis concerns an actual research area related to automation of the process of computational linguistic resources creation, namely, by automatic generation of the derived words that are absent in computational linguistic resources.

The purpose is to study the mechanisms and to elaborate algorithms for automatic generation of the derived words for these resources completion.

The research objectives are: evaluation of the existent methods in the automation of the derivational process; study of the structure particularities of computational linguistic resources available for research; establishing the quantitative and qualitative characteristics of the derived words; elaboration of the algorithms for automatic recognition of the derived words; establishing the mechanisms and elaboration of algorithms for automatic generation of the derived words.

Novelty and scientific originality. This work contributes to complete research in the field of natural language processing by development of mathematical models and algorithms to solve the problem of automatic derivatives generation. The results of the study represent a realization of a new methodology of studying the issues in computational derivational morphology, related to the algorithmization of certain linguistic mechanisms, such as affixes substitution, derivatives projection, derivational constraints and formal derivational rules.

Important scientific problem solved in the field of research. The problem of automatic generation of derivatives for some classes of Romanian words was solved, which contributes to the facilitation "human – computer" communication in natural language by creation of the computational lexicons, which are the basis of various applications of this field.

Theoretical significance and applied value of the thesis. A statistical method for Romanian affixes uncertainty evaluation based on the notion of entropy was proposed. The mathematical formal descriptions of the derivatives word formation mechanisms were elaborated which served to development of algorithms for automatic generation of the derivatives. During the research the important results were obtained, which permitted to elaborate algorithm for automatic generation of derivatives which can facilitate computational linguistic resources completion. The research results present interest for lexicographic practice, in the process of dictionary elaboration and lexicographic treatment of the derivatives. Also, the results of the investigation can serve as a methodical support in activity of the specialists in both computer science and linguistics.

Implementation of scientific results. An extension of RRTLN database was developed which allowed a correct extraction of about 15.000 derivatives without having a special program of word segmentation in morphemes (41 of prefixes, about 420 of suffixes, over 8 thousand of roots/stems). The established mechanisms, which permitted the elaboration of algorithms and corresponding programs, led to generation a significant number of derivatives with different affixes, 8839 with 11 prefixes, and 2352 with 24 suffixes which will help in Romanian language computational linguistic resources essential enrichment.