

V.A.Yatsko

Problems in text preprocessing  
and automatic analysis

(A monograph)

Abakan - 2012

ББК **81.1+** 32.973.202

**Я 936**

Published on the decision of the Resource Board at Katanov State University of Khakasia

Reviewers: PhD, Associate Professor  
**L.V.Tatarinova** PhD,  
Associate Professor  
**S.V.Shvets**

**Yatsko V.A.**

**Я 936**

**Problems in text preprocessing and automatic analysis.** A monograph - Abakan, 2012. - 178 p.

The monograph focuses on algorithms and programs for text preprocessing classified according to different levels of language system. Original approaches for solving the problems of text decomposition, morphological analysis and POS tagging are described. The classification of lexicographic sources, which are used to support text processing systems, is given and methods for their generation are offered.

The monograph reflects the author's experience in creation of linguistic software. It is intended for experts in natural language processing and postgraduate students.

© Viatcheslav Yatsko, 2012

## Contents

Preface.....	4
1. INTRODUCTION .....	6
1.1. Terminological issues.....	6
1.2. Some historical facts .....	11
1.2.1. Machine translation .....	12
1.2.2. Information retrieval.....	16
1.2.3. Summarization .....	26
1.2.4. Text mining .....	34
2. TEXT PREPROCESSING ALGORITHMS.....	42
2.1. Decomposition .....	44
2.1.1. Tokenization and syntactic splitting .....	47
2.1.2. Clause splitting.....	55
2.1.3. Segmentation.....	67
2.2. Morphological analysis.....	76
2.3. Tagging .....	85
2.4. Anaphora resolution.....	94
3. LEXICOGRAPHIC SOURCES.....	108
3.1. Dictionaries, thesauri, and ontologies.....	108
3.2. Methods for dictionary generation.....	119
4. CONCLUSIONS .....	134
REFERENCES.....	138
APPENDIX 1 .....	153
APPENDIX 2.....	173
APPENDIX 3 .....	175