ISO 24614-2:2011-09 (E)

Language resource management - Word segmentation of written texts - Part 2: Word segmentation for Chinese, Japanese and Korean

Contents

Page

Foreword v

Introduction vi

1 Scope 1

2 Normative references 1

3 Terms and definitions 2

4 Overview 4
4.1 Introduction 4
4.2 Markup convention 4
4.3 Review of the concept of word segmentation unit 5
4.4 Features common to Chinese, Japanese and Korean 5

5 General rules for identifying WSUs in Chinese, Japanese and Korean 6
5.1 Words 6
5.2 Derivationally formed words 6
5.3 Word compounds 7
5.4 Phrasal compounds 8
5.5 Idioms 8
5.6 Fixed expressions 9
5.7 Abbreviations 10
5.8 Transliterated loanwords 10
5.9 Strings of foreign or special characters 11
5.10 Components of a WSU 11

6 Specific rules for identifying WSUs in Chinese 12
6.1 Lexical items followed by the suffix (r) 12
6.2 Lexical items 12
6.2.1 Nouns 12
6.2.2 Verbs 17
6.2.3 Adjectives 20
6.2.4 Pronouns 22
6.2.5 Numerals 23
6.2.6 Measure words 25
6.2.7 Adverbs 25
6.2.8 Prepositions 26
6.2.9 Conjunctions 26
6.2.10 Auxiliary words 26
6.2.11 Modal words 27
6.2.12 Exclamations 27
6.2.13 Imitative words 27

7 Specific rules for identifying WSUs in Japanese text 27
7.1 Bunsetsus 27
7.2 Lexical items 27
7.2.1 General rule 27
7.2.2 Nouns 28
7.2.3 Verbs 32

7.2.4 Adjectives 33
7.2.5 Adnouns 34
7.2.6 Adverbs 34
7.2.7 Conjunctions 35
7.2.8 Exclamations 35
7.2.9 Particles 35
7.2.10 Auxiliary verbs 35

8 Specific rules for identifying WSUs in Korean text 36
8.1 Eojeols 36
8.2 Lexical items 36
8.2.1 General rule 36
8.2.2 Nouns 37
8.2.3 Pronouns 38
8.2.4 Numerals 39
8.2.5 Verbs 39
8.2.6 Adjectives 39
8.2.7 Adnouns 40
8.2.8 Adverbs 40
8.2.9 Exclamations 40
8.3 Grammatical affixes 40

Annex A (informative) Comparative table of parts of speech in Chinese, Japanese and Korean 42

Bibliography 43