INTERNATIONAL STANDARD

ISO 24614-2

First edition
2011-09-01

Language resource management — Word segmentation of written texts —

Part 2: Word segmentation for Chinese, Japanese and Korean

Gestion des ressources langagières — Segmentation des mots dans les textes écrits —

Partie 2: Segmentation des mots pour le chinois, le japonais et le coréen

Reference number
ISO 24614-2:2011(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

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 儿(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**

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 24614-2 was prepared by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

ISO 24614 consists of the following parts, under the general title *Language resource management — Word segmentation of written texts*:

— *Part 1: Basic concepts and general principles*

— *Part 2: Word segmentation for Chinese, Japanese and Korean*

Introduction

This part of ISO 24614 focuses on word segmentation in Chinese, Japanese and Korean written texts. As far as typography is concerned, there is no white space between words in Chinese, Japanese or pre-modern Korean texts. This makes it hard to segment a text into words, unless there is a consistent way of identifying word segmentation units for those languages. On the other hand, in modern-day Korean text, word forms or verbal stems that are agglutinated with grammatical affixes, called ‘eojeol’ or ‘malmadi’, are separated by white space as in English written texts. Hence, it is much easier to identify words or other word segmentation units in a Korean text. Nevertheless, a large number of words in Korean as well as in Japanese are borrowed or derived from Chinese words; their internal structures are also based on the word formation principles of Chinese. As a consequence, general rules for identifying word segmentation units (WSUs) in Chinese, especially internal WSUs embedded in larger WSUs, are also applicable to some extent to the processing of Japanese and Korean texts.

The use of characters does not play a real role in identifying WSUs in a text. Many Korean words can be written either in Chinese or in Korean characters, but the same principles of analysing Chinese-derived words and identifying sub-WSUs of those words apply. A newspaper published in Beijing is written in simplified Chinese characters, while a Hong Kong newspaper may be written in traditional Chinese characters. Here again, the same principles of identifying WSUs apply to both newspapers.

This part of ISO 24614 first sets out the general rules for identifying WSUs in Chinese, Japanese and Korean, then addresses the specific rules for each language.

Language resource management — Word segmentation of written texts —

Part 2: Word segmentation for Chinese, Japanese and Korean

1 Scope

The basic concepts and general principles of word segmentation as defined in ISO 24614-1 apply to Chinese, Japanese and Korean. Text needs to be segmented into tokens, words, phrases or some other types of smaller textual units in order to perform certain computational applications on language resources, such as natural language processing, information retrieval (IR) and machine translation (MT). This part of ISO 24614 is restricted to the segmentation of a text into words or other word segmentation units (WSUs). This task is distinct from morphological or syntactic analysis *per se*, although it greatly depends on morphosyntactic analysis. It is also different from the task of laying out a framework for constructing a lexicon and identifying its lexical entries, namely lemmas and lexemes. The frameworks for the latter tasks are provided by ISO 24611, ISO 24613 and ISO 24615.

The main objective of this part of ISO 24614 is to specify rules for delineating WSUs for Chinese, Japanese and Korean. Some rules are common to all three languages, though each language also has its own distinct rules for identifying WSUs. The common features are discussed in Clause 5, then the distinct rules are laid out in Clause 6 for Chinese, Clause 7 for Japanese and Clause 8 for Korean.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 24611, *Language resource management — Morpho-syntactic annotation framework*

ISO 24613:2008, *Language resource management — Lexical markup framework (LMF)*

ISO 24614-1:2010, *Language resource management — Word segmentation of written texts — Part 1: Basic concepts and general principles*