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

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*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 24611, ISO 24613 and ISO 24614-1 and the following apply.

3.1
adnoun
ADN
non-conjugating word that modifies a noun

NOTE Adnouns modify nouns, as adverbs modify verbs.

EXAMPLE 1 <Japanese>

a. あらゆる 国
arayuru kuni
ADN N
'every country'

b. 好きな 花
suki+na hana
ADNst+SX N
'favourite flower'

EXAMPLE 2 <Korean>

a. 새 옷
sae ot
ADN noun
'new clothes'

b. 빨간 옷
bbalga+n ot
ADJst+GX N
'red clothes'

3.2
bunsetsu
<Japanese text> **phrase** (3.8) without internal modifying relations

EXAMPLE The sentence 私は学校へ早く行きました(I went to school early) consists of four bunsetsus: 私は(watashiwa), 学校へ (gakkoue), 早く(hayaku) and 行きました(ikimashita) in which

私(watashi)	is a pronoun,
は(wa)	is a particle,
学校(gakkou)	is a noun,
へ(e)	is a particle,
早く(hayaku)	is an adjective in adverbial usage,
行き(iki)	is a verbal stem followed by
まし(mashi)	is an auxiliary verb denoting politeness, and
た(ta)	is an auxiliary verb indicating the past tense.

NOTE A bunsetsu normally consists of a noun plus its particle(s) or a verb plus its ending(s), auxiliary verb(s) or particle(s) as shown in the example above.

3.3
ending
<Japanese text> agglutinative affix of a verb or adjective

NOTE Verbs and adjectives end with agglutinative forms, called “endings”. These endings may be a negative form, an adverbial form, a base form, an adnominal form, an assumption form or an imperative form.

3.4
eojeol
malmaldi
<Korean text> word or its variant word form agglutinated with grammatical affixes

NOTE 1 White space (space between characters) helps to segment text into eojeols.

EXAMPLE

내가 사과를 먹었다
nae+ga sagwa+reul meok+eot+da
pronoun+GX noun+GX Vst+GX+GX
‘I’+SBJ ‘apple’+OBJ ‘eat’+PST+DCL = ‘I ate (an) apple’

NOTE 2 This sentence consists of three eojeols: 내가, 사과를 and 먹었다, each of which is separated by white space. The acronyms GX, SBJ, OBJ, PST and DCL in the example above stand for grammatical affix, subject, object, past tense and declarative sentential type, respectively. The pronoun 내 is a variant form of the pronoun 나 referring to the speaker. 먹었다 is an eojeol and at the same time is a word form agglutinated with two grammatical affixes 었 and 다 to a verb stem 먹.

3.5
lexical item
entry in a lexicon that is a lexeme or one of its variant forms

NOTE Headed by a lemma, each lexical item may be either a free-standing word (or one of its variant word forms) or a bound (non-free-standing) form such as stems and affixes. See ISO 24614-1:2010 for the definitions of lexeme, lemma and lexicon.

3.6
measure word
<Chinese text> part of speech defining, along with numbers, the quantity of a given object, or identifying specific objects with demonstrative pronouns such as “this” and “that”

NOTE 1 Whereas English speakers say “one person” or “this person”, Chinese speakers say respectively 一个人(yi ge ren; numeral + measure word + noun; one person) or 这个人 (zhe ge ren; demonstrative pronoun + measure word + person; this person), where 个 (ge) is a measure word.

NOTE 2 A set of “verbal measure words” is used to count the number of times an action occurs, rather than the number of items. For example, in the sentence 我去过三次北京(wo qu guo san ci Beijing; pronoun + verb + auxiliary word + numeral + measure word + proper noun; I have been to Beijing three times), the 次(ci) functions as a measure word to combine with a numeral 三 to derive the adverb 三次 (sanci) that modifies the verb 去(qu).

3.7
particle
<Japanese text> grammatical affix agglutinated mostly to nominal forms but sometimes to other free-standing **lexical items** (3.5)

NOTE The grammatical category particle can be treated as a part of speech.

EXAMPLE The noun phrase 学校へ(gakkoue) is analysed into a noun 学校 (gakkou) and a particle へ(e). The verb phrase 寒いね(samuine, ‘It is very cold, isn't it?’) is analysed into a verb 寒い (samui) and a particle ね(ne) which corresponds to the tag ‘isn't it ?’.

3.8
phrase
group of words that perform a grammatical function and that form a conceptual unit within a sentence

4 Overview

4.1 Introduction

This clause first introduces a markup convention for word segmentation units, then reviews the concept of word segmentation unit (WSU) which was introduced in ISO 24614-1. Some features shared by Chinese, Japanese and Korean are discussed in 4.3. A comparative table of parts of speech is given in Annex A.

4.2 Markup convention

The following clauses contain a very large number of examples of WSUs. A simple way of representing WSUs is introduced here.

NOTE This markup convention is introduced here just for the sake of simple illustration in this part of ISO 24614.

First, a stand-off annotation is adopted; this allows primary data to be kept intact from markup notations. (More information on linguisitic annotation can be found in ISO 24612.) Exceptions are made concerning this requirement when primary data in Chinese or some other language are not provided with a romanized version or when the identification of syllables is not easy.

Second, a citation format consisting of four lines is adopted.

— Line 1 introduces primary text fragment in its original script.

— Line 2 represents annotated fragment in romanized form.

— Line 3 assigns morpho-syntactic descriptions.

— Line 4 provides an English equivalent.

The following symbols are used (optionally) for marking up primary data in a romanized form:

1) the dot '.' for syllable boundaries, if ambiguity arises;

2) the sign '+' (plus) for boundaries between a word and an affix;

3) the underscore '_' for word segmentation units (WSUs);

4) the square brackets '[ ]' for WSUs, if ambiguity arises;

5) the parentheses '( )' for non-WSUs;

6) the symbol ':=' represents combined resultant information.

EXAMPLE <Korean>

헛돌다
[(heot)+dol]_da
prefix + verbStem_GX := verb
'in vain' 'spin' := 'to spin in vain'

The verb 헛돌다 consists of a prefix 헛 (heot) and a verb 돌다 (dol+da). The verb 돌다 is then analysed into a stem 돌 (dol) and a verbal grammatical affix 다 (da). The example 헛돌다 as a whole is a WSU, while its subpart 돌다 is also a WSU embedded in it. In Korean, as will be discussed, the prefix 헛 (heot) is not treated as a WSU.

For computational purposes, namely character encoding schemes, the characters in Chinese, Japanese and Korean are all treated as being syllabic. Hence, each of the characters in primary data can easily be identified with its corresponding romanized string of alphabetic characters. The word given in the above example, for instance, consists of three syllable characters: 헛, 돌 and 다, while its corresponding romanized version 'heotdolda' also consists of three syllables, 'heot', 'dol' and 'da'. As a result, we can identify the first syllable 'heot' in the romanized version as corresponding to the first syllable 헛 in the Korean example and also the following sequence of the two syllables in the romanized version as corresponding to the sequence of two syllable characters 돌다. This indicates that there is no need to mark up primary data, but to keep them intact in order to show how they are analysed into WSUs or other morphological units.

In Japanese, however, a single Chinese character may be pronounced as more than one syllable: for example, the noun consisting of one character 桜 is pronounced as a three-syllable string 'sakura'. In such a case, primary data are marked up.

4.3 Review of the concept of word segmentation unit

Word segmentation is the process of dividing a text into meaningful units called word segmentation units (WSUs). Each WSU corresponds to a single concept: for example, 'the White House' consists of three words but designates a single concept referring to the residence of the US President. It follows that 'the White House' corresponds to one WSU. In other words, word count and concept do not necessarily correspond, and may differ from one language to the next. The single English word 'pork' is translated by two words that mean 'pig meat' in Chinese 猪肉 (zhu_rou), in Japanese 豚肉 (buta_niku), and in Korean 돼지고기 (doeji_gogi). So although English uses only one word and the other languages use two, in all four languages there is just one WSU.

A unit that carries a meaning that is useful for any linguistic processing can be defined as a WSU. A WSU can be an entry in a lexicon or any other type of lexical resource insofar as such an entry is leveraged in some natural language processing application. In other words, the WSU's dimension is more or less fixed, but linguistic interferences between compounds inside a WSU are not allowed. Such an extensive, open definition of WSU is useful for further linguistic processing because some WSUs that frequently occur in corpora are not systematically decomposable by syntactic or any other linguistic processing.

4.4 Features common to Chinese, Japanese and Korean

Two basic features that are common to Chinese, Japanese and Korean derive from a common cultural heritage in Far East Asia.

— Firstly, Chinese characters have long been used, and continue to be used, in this part of the world notwithstanding some differences in their degree of use: Chinese utilizes all of these characters, while Japanese also uses them in addition to Kana characters. On the other hand, Korean has its own writing system, but sometimes uses Chinese characters, especially for scholarly purposes in humanities such as classical studies.

— Secondly, many words and phrases of Chinese origin are used in both Japanese and Korean; they include 四面楚歌 and 第二次世界大战. Note, however, that the non-simplified or original shapes of Chinese characters are retained in these languages; in the case of Korean, they may simply be written using the characters of the Korean writing system. The phrase 四面楚歌 written in Chinese characters, for instance, is written as 사면초가 in Korean.

Because of this historical background, some principles of Chinese word segmentation apply significantly to Chinese-derived words found in Japanese and Korean. If the word is derived from Chinese characters, the three languages have common properties. If the word is a noun and consists of two or more Chinese characters, it will constitute a single WSU as long as the characters are "tightly combined and steadily used" in accordance with the principles set out in ISO 24614-1; for example, 'each country' in English is not a single

WSU, but two WSUs unlike its Chinese equivalent 各国. However, if the final character is productive in a limited manner, it forms a single WSU with the preceding word; for example, 東京都 (Tokyo Metropolitan), 8 月 (August) and 加速器(accelerator) are single WSUs without being analysed into two WSUs, say 東京 and 都.

Because the motivation for a word segmentation standard is to recommend which WSUs should be listed in a given type of lexicon (i.e. not a linguistics lexicon but any kind of practical, indexed container of WSUs), there may be conflicting principles; for example, principles of non-productivity, frequency and granularity could trigger conflicts because they are marked by different perspectives for defining WSUs.

Nouns derived from Chinese characters may be shared for the purposes of establishing the WSU structure of the three languages, but not in every respect. However, Korean and Japanese do have certain features in common; for example, some Korean verbal affixes and Japanese auxiliary verbs perform the same functions. Word segmentation in each language varies in line with existing word segmentation rules, and sometimes even breaches one or more principles of word segmentation. This will be a starting point for recommending a more synchronized concept of “word segmentation unit” (WSU) in a multilingual environment. The aim of the concept of “word segmentation unit” is to broaden our view about what could be contained in a lexicon used for natural language processing purposes, and with little linguistic representation.

5 General rules for identifying WSUs in Chinese, Japanese and Korean

5.1 Words

All words and their variant word forms are WSUs.

5.2 Derivationally formed words

All derivationally formed words or their variant word forms are treated as WSUs.

Derivational affixes (AX), which can be either prefixes or suffixes, are also treated as WSUs in Chinese, but not in Japanese or Korean.

EXAMPLE 1 <Chinese>

Two examples of WSUs, 科学家 ‘scientist’ and 物理学家 ‘physicist’, are shown below.

a. 科学家
kexue_jia
N AX
‘science’ ‘expert’ := ‘scientist’
b. 物理学家
[wuli_xue]_jia
N AX AX
‘physics’ ‘discipline’ ‘expert’ := ‘physicist’

The first Chinese example consists of two WSUs, 科学 ‘science’ and 家 ‘expert’, while the second consists of four WSUs, 物理学 (wuli.xue) ‘physics’, 物理 (wuli) ‘physics’, 学 (xue) ‘science’ and 家 (jia) ‘expert’.

EXAMPLE 2 <Japanese>

a. 非常勤
(hi)_jouikin
AX N := noun
‘non-’ ‘full-time working’ := ‘part-time work’

b. 音楽家
ongaku_(ga)
N AX := noun
'music' 'professional person' := 'musician'

Both of the Japanese examples are derived nouns. The noun 非常勤 is derived by adding the prefix 非 to the noun 常勤. The derived noun as a whole is a WSU and so is the noun 常勤, while the prefix 非 by itself is not a WSU. Likewise, the noun 音楽家 is also a derived noun that consists of a noun 音楽 and a suffix 家. Here again, the derived noun 音楽家 and its component noun 音楽 are treated as WSUs, but the suffix 家 is not a WSU.

EXAMPLE 3 <Korean>

a. 음악가
eumak+(ga)
N _ AX := noun
'music' 'professional' := 'musician'

b. 헛돌다
[(heot)+dol]_da
AX_V := verb
'false' 'spin' := 'to spin without any result'

The Korean examples contain four WSUs: 음악가, 헛돌다, while 음악 and 돌다 are also treated as WSUs if they occur independently in a text. Neither the suffix 가 nor the prefix 헛 in isolation is treated as a WSU. No derivational affixes are treated as WSUs in Korean, mainly because they are not words.

5.3 Word compounds

All word compounds are treated as single WSUs.

NOTE The term "word compound" is defined in ISO 24614-1:2010, 2.28. Unlike a phrasal compound, the meaning of a word compound is only partially predictable from the meanings of its constituent words.

EXAMPLE 1 <Chinese>

白菜
baikcai
noun
'white' 'vegetable' := 'Chinese cabbage'

The word 白菜 above is a noun, consisting of two words 白(baik) 'white' and 菜 (cai) 'vegetable'. It is a compound noun and is treated as a single WSU, for the meaning of its component words is only partially preserved in the process of compounding. The noun 白菜 refers to a kind of vegetable that may in fact not be white, but green with green leaves.

EXAMPLE 2 <Japanese>

海外旅行
kaigai_ryokou
noun noun := noun
'abroad' 'travel' := 'traveling abroad'

The WSU 海外旅行 as a compound noun consists of two nouns 海外(kaigai) and 旅行(ryokou); both of them are also WSUs.

EXAMPLE 3 <Korean>

a. 손목
son_ mok
noun noun := noun
'hand' 'neck' := 'wrist'

b. 바로잡다
baro_jabda
adverb verb := verb
‘rightly’ ‘hold’ := ‘to correct’

Here both of the compounds, 손목 and 바로잡다, are WSUs, as are their component words, 손 ‘hand’, 목 ‘neck’, 바로 ‘rightly’ and 잡다 ‘to hold’ if they occur independently in a text.

5.4 Phrasal compounds

All phrasal compounds are treated as single WSUs.

NOTE The term “phrasal compound” is defined in ISO 24614-1:2010, 2.20. Unlike a word compound, the meaning of a phrasal compound is predictable from the meaning of each of its constituents.

EXAMPLE 1 <Chinese>

a. 猪肉
zhu_rou
noun noun := noun
‘pig’ ‘meat’ := ‘pork’

b. 发电厂
fadian_chang
verb noun := noun
‘to generate electricity’ ‘plant’ := ‘power plant’

The phrasal compounds 猪肉 and 发电厂 are WSUs and so are their components, 猪 ‘pig’, 肉 ‘meat’, 发电 ‘to generate electricity’ and 厂 ‘place’.

EXAMPLE 2 <Japanese>

豚肉
buta_niku
noun noun := noun
‘pig’ ‘meat’ := ‘pork’

EXAMPLE 3 <Korean>

돼지고기
doeji_gogi
noun noun := noun
‘pig’ ‘meat’ := ‘pork’

In both Japanese and Korean, these compounds are treated as WSUs and so are their component words.

5.5 Idioms

Idioms are treated as single WSUs.

NOTE An idiom is a kind of multiword expression (MWE) defined in ISO 24614-1:2010, 2.19. Like some other types of MWE, the parts of speech of idioms vary, e.g. from noun to verb.

EXAMPLE 1 <Chinese>

a. 胸有成竹
xion you cheng zhu
‘to have a well-thought-out plan’

b. 欣欣向荣
xin xin xiang rong
‘to be prosperous’

NOTE Most idioms in Chinese are four-character phrases.

EXAMPLE 2 <Japanese>

腹が立つ
[hara _(ga)] _atatsu
noun particle verb := idiom
‘stomach’ nominative ‘occur’ := ‘I am upset’

The string of words 腹が立つ of a sentential form is a WSU because it is an idiom. Their components, except for the particle が (ga), are also WSUs.

EXAMPLE 3 <Korean>

a. 수박겉핥기
subak_[geot_halgi]
noun noun verb := idiom
‘watermelon’ ‘surface’ ‘licking’ := ‘superficial knowledge’

b. 함흥차사 (咸興差使)
hamheung _chasa
proper noun noun := idiom
‘Hamhung’ ‘messenger’ := ‘no news’

The string of characters 수박겉핥기 in a. is an idiom and thus treated as a WSU. Its components, 수박 (subak), 겉 (geot) and 핥기 (halgi) are also WSUs.

The character string 함흥 as a proper name can be treated as a WSU, as can the character string 차사 which refers to a messenger.

5.6 Fixed expressions

Fixed expressions such as proverbs and mottos are treated as single WSUs.

EXAMPLE 1 <Chinese>

a. 对不起
dui bu qi
‘sorry’

b. 春夏秋冬
chun xia qiu dong
‘spring summer autumn winter’

c. 由此可见
you ci ke jian
‘this shows’

d. 不管三七二十 一
bu quan san qi er shi yi
‘no matter three seven two ten one’

e. 失败是成功之母
shibai shi chengong zhi mu
‘Failure is the mother of success’