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'

EXAMPLE 2 <Japanese>

時は金なり
toki +wa _kane +nari
noun particle noun auxiliary := sentence
‘time’ topic-marker ‘money’ copula := ‘Time is money’

EXAMPLE 3 <Korean>

a. 천리 길도 한 걸음부터
cheol+li gil+do han georum+buteo
measure noun+affix adnoun noun suffix
‘thousand li’ ‘road’ ‘one’ ‘walk’ ‘from’ := ‘step by step’

b. 다시 말하여
dasi_malha+yeo
adv verb + GX
‘again’ ‘speaking’ := ‘In other words’

The idioms as a whole are treated as single WSUs, while their components are also treated as WSUs. Note that the occurrences of white space between these expressions indicate that each constituent word is a WSU as well as an eojeol.

5.7 Abbreviations

Abbreviations are treated as single WSUs.

EXAMPLE 1 <Chinese>

a. 科技
keji
‘science and technology’

b. 工农业
gongnongye
‘industry and agriculture’

NOTE Abbreviations in Chinese texts usually consist of two, three or four characters.

EXAMPLE 2 <Japanese>

特急 (tokkyuu, noun: super express): 特別 (super) + 急行 (express)

EXAMPLE 3 <Korean>

의대[醫大] (uidae, noun, ‘medical college’ : 의과 (uigwa; ‘medical area’) + 대학 (daehak; ‘college’)

5.8 Transliterated loanwords

Transliterated loanwords, namely those foreign words that keep pronunciations in their original languages, are treated as single WSUs.

EXAMPLE 1 <Chinese>

a. 吉普
jipu
‘Jeep’

b. 巧克力
giaokeli
'chocolate'

EXAMPLE 2 <Japanese>

a. ジープ (jeep)

b. チョコレート (chocolate)

Transliterated loanwords in Japanese are normally written in kata kana.

EXAMPLE 3 <Korean>

a. 피아노 (piano)

b. 바이올린 (violin)

5.9 Strings of foreign or special characters

A string of foreign or special characters such as foreign language characters, Arabic numerals, and mathematical and chemical symbols are treated as WSUs.

EXAMPLE 1

a. Chomsky

b. F16

c. X-Ray

d. 1298

e. +

f. CO_2

These strings may be mixed with Chinese, Japanese or Korean characters in the text.

EXAMPLE 2

<Korean text> Chomsky 는 언어학자이다.. 'Chomsky is a linguist.'

5.10 Components of a WSU

Some components of a WSU can themselves be WSUs.

As has been discussed, many WSUs have an internal structure that organizes several WSUs hierarchically. Structures of this type can be manipulated at different granularity levels in the process of word segmentation according to the requirements of various applications. For example, 猪肉 in Chinese can be treated as a single WSU for machine translation (MT), which translates it into a single word 'pork' in English, whereas it can be treated as two WSUs for the purposes of information retrieval (IR), which looks for two different ontological entities ('pig' and 'meat').

6 Specific rules for identifying WSUs in Chinese

6.1 Lexical items followed by the suffix儿(r)

Lexical items followed by the suffix 儿, a retroflex r, which has a discourse function for marking prominence in an informal speech or text, are treated as single WSUs. The suffix (SX) 儿 is often attached to nouns (N), and sometimes to verbs (V) or adverbs (ADV) to make these items displayed more vividly in informal speech or writing of Mandarin Chinese. The suffix has no meaning in itself except for such a discourse function.

EXAMPLE

a. 花儿
hua_(r)
N SX := noun
'flower' emphasis := 'flower'

b. 玩儿
wan_(r)
V SX := verb
'to play' emphasis := 'to play'

c. 悄悄儿
qiaogiao_(r)
ADV SX := adverb
'quietly' emphasis := 'quietly'

The noun 花儿 as a whole, for instance, is a WSU, as is the first character 花. The suffix 儿, however, is not a WSU. It is thus put into parentheses.

6.2 Lexical items

6.2.1 Nouns

6.2.1.1 General

Nouns (N) are WSUs. They are subcategorized into common nouns (CN) and proper nouns (PN).

6.2.1.2 Common nouns

A common noun (CN) modified by an adjective (ADJ) is segmented into two WSUs, while the meaning of the modifying adjective contributes to the meaning of the modified noun as a whole.

EXAMPLE 1

a. 小床
xiao_chuang
ADJ CN := noun phrase
'small' 'bed' := 'small bed'

b. 小媳妇
xiao.xiwu
ADJ CN := noun
'small wife' := '(young) wife'

Example 1 a. is segmented into two WSUs, 小 (xiao) and 床 (chuang), for the meaning of the adjective 小 'small' is preserved, whereas example 1 b. is treated as a single WSU because the meaning of the adjective 小 'small' is not preserved. The wife referred to may be a big woman.

Common nouns (CN) that are combined with an expression referring to a direction (DIR) or location (LOC) are treated as two WSUs.

EXAMPLE 2

a. 桌子上
zuozi_shang
N LOC
'table' 'above' := 'on the table'

b. 长江以北
N LOC
'Yangtzi River' 'north' := 'north of the Yangtzi River'

Each of these character strings constitutes two WSUs.

Common nouns with the plural suffix (SX) 们 (men) are treated as two WSUs. The character string 朋友们 'friends', for instance, is segmented into two WSUs, 朋友 (noun) and 们 (plural suffix).

EXAMPLE 3

朋友们
penyou_men
CN_SX := noun
'friend'plural := 'friends'

However, the following nouns are treated as single WSUs because the character 们 is understood to be an inseparable part.

EXAMPLE 4

a. 人们 renmen, 'people'
b. 哥儿们 germen, 'pals'
c. 爷儿们 yiermen, 'guys'

Each of the names of months and days as a whole is a single WSU.

EXAMPLE 5

a. 元月
yuanyue
'first' 'month' := 'January'

b. 3 月
sanyue
'three month' := 'March'

c. 五月
wuyue
'five month' := 'May'

d. 礼拜三
libaisan
'week three' := 'Wednesday'

e. 星期日
xinqiri
'week day' := 'Sunday'

Each temporal expression referring to a year, day, hour, minute or second is also treated as a single WSU.

EXAMPLE 6

1988 年 3 月 15 日
1998nian_sanyue_15 ri
‘year 1998’ ‘March’ ‘day fifteen’ := ‘15 March 1998’

11时42 分 8 秒
11shi_42fen_8miao
‘11 o'clock’ ‘42 minutes’ ‘8 seconds’ := ‘11:42:08 sec’

Temporal expressions that are combined with affixes such as 上 (last), 下 (next), 前 (before last), 后 (after next), 大前 (before before last), and 大后 (after after next) are single WSUs.

EXAMPLE 7

a. 上星期 shangxingqi ‘last week’

b. 下月 xiayue ‘next month’

c. 前天 qiantian ‘the day before yesterday’

d. 后年 hounian ‘the year after next’

e. 大前天 daqiantian ‘the day before the day before yesterday, three days ago’

f. 大后年 dahounian ‘the year after the year after next, three years later’

The temporal nouns referring to the first ten days of each month, starting from 初一 (the first day) to 初十 (the tenth day) in the Chinese lunar calendar, are WSUs.

6.2.1.3 Proper nouns

Personal names that consist of a surname (family name) and a given name are WSUs. Each of these WSUs can be further segmented into two WSUs, corresponding to a surname and a given name.

EXAMPLE 1

a. 张胜利
zhang_shengli
surname given name
‘Zhang Shengli’

b. 欧阳志华
ouyang_zhihua
surname given name
‘Ouyang Zhihua’

NOTE In Chinese as well as in Japanese and Korean, surnames are written first before given names.

The surname with a title is segmented into two WSUs.

EXAMPLE 2

a. 张教授
zhang _ jiaoshou
surname title
‘Zhang’ ‘professor’ := ‘Professor Zhang’

b. 王部长
wang_buzhang
surname title
'Wang Minister' := 'Minister Wang'

c. 李师傅
li_shifu
surname title
'Li' 'master' := 'Master Li'

Whether a title comes before or after, surnames with a title are, however, treated as single WSUs, if the title consists of only one character.

EXAMPLE 3

a. 老张
laozhang
title surname
'Venerable Zhang'

b. 陈总
chenzong
surname title
'Manager Chen'

The kinship terms with titles are single WSUs. They each have an internal structure.

EXAMPLE 4

a. 三叔
sanshu
'three uncle' := 'the third younger uncle'

b. 大女儿
danyer
'big daughter' := 'the eldest daughter'

The geopolitical expressions such as 族(nationality), 省(province), 市(city), 州(prefecture), 县 (county), 乡 (twon), 区(district), 江(river), 河(bigger river) and 山(mountain) that are suffixed to proper names are treated as separate WSUs, unless these proper names are single character names (e.g. 汉族 the Han nationality or 忻县 Qi County).

EXAMPLE 5

a. 哈萨克族 the Kazakstan nationality

b. 北京市 Beijing Municipality

c. 浙江省 Zhejiang Province

d. 正定县 Zhengding County

These names are each segmented into two WSUs: Example 5 a., for instance, is segmented into 哈萨克 'Kazakstan' and 族 'nationality'.

Proper names with multiple references are not segmented, but treated as single WSUs.

EXAMPLE 6

a. 牡丹江 Mudanjiang

b. 横断山 Hengduan Mountains

The place name 牡丹江 refers to the Mundan river in Helongjian, a prefecture-level city in Helongjian, or even Songjian province, formerly Mundanjian province in China. The name 横断山 refers to not a single mountain, but a range of mountains.

Suffixal expressions referring to streets, roads, villages, towns, oceans and seas are not segmented. Thus their full names are treated as single WSUs.

EXAMPLE 7

a. 长安街 'Chang'an Avenue'

b. 学院路 'Xueyuan Road'

c. 周口店 'Zhoukoudian'

d. 刘家村 'Liujiacun Village'

e. 大西洋 'the Atlantic Ocean'

f. 地中海 'the Mediterranean Sea'

Country names in a non-abbreviated full form are not segmented, but treated as single WSUs.

EXAMPLE 8

a. 中华人民共和国 'People's Republic of China'

b. 大不列颠及北爱尔兰联合王国 'United Kingdom'

Names of organizations, agencies or institutions may be segmented into more than one WSU, depending on their constituent structures.

EXAMPLE 9

a. 联合国_教科文_组织
United Nations - Educational, Scientific and Cultural - Organization)

b. 中国_共产党
China - Communist Party

The names of trademarks, product types and product series that are suffixed to their names are treated as separate WSUs.

EXAMPLE 10

a. 永久_牌 'Yongjiu Brand' (trademark)

b. 中华_烟 'Zhonghua Cigarette' (product type)

c. 牡丹_Ⅱ型 'Peony II' (product series)

6.2.2 Verbs

6.2.2.1 Various forms of reiterative verbs

Verbs that are formed by reiterating one or two identical characters are not segmented, but treated as single WSUs.

EXAMPLE 1

a. 看看 'look at'

b. 动动 'move'

c. 来来往往 'come and go'

d. 拉拉扯扯 'drag'

Verbs reiterated in the form of "AAB, ABAB" are each segmented into two WSUs: $_{WSU}$AAB into $_{WSU}$AA and $_{WSU}$B and $_{WSU}$ABAB into $_{WSU}$AB and $_{WSU}$AB.

EXAMPLE 2

a. 说说_看 'try to say'

b. 研究_研究 'to have a discussion'

An emphatic expression such as 一, 了 or 了 一, may be inserted into verbs that are formed by iterating one identical character. These verbs are segmented into three WSUs.

EXAMPLE 3

a. 谈_一_谈 'have a good chat'

b. 想_一_想 'think carefully'

c. 读_一_读 'to read'

d. 想 - 了 - 想 'think it over'

e. 想_了 一_想 'think it over'

6.2.2.2 Verbal prefixes with a negative meaning

Negative expressions such as 不 'not' that are prefixed to verbs are treated as WSUs themselves.

EXAMPLE

a. 不_写 'not to write'

b. 不_能 'cannot, impossible'

c. 没_研究 'not to do research'

d. 未_完成 'incomplete'

6.2.2.3 Alternative question forms

Alternative ('either - or') questions may be formed by conjoining a verb with its negative form, as illustrated below. These question forms are segmented into more than one WSU.

EXAMPLE

a. 说 _[不_说] ‘say or not say?’

b. 看_[不_看] ‘see or not see?’

c. 相信_[不_相信] ‘believe or not believe?’

NOTE However, the forms such as 相不相信 ‘believe or not’ are not segmented into more than one WSU, but treated as single WSUs.

6.2.2.4 Verb-object structures and verb collocations

A string of words in the form of verb-object is segmented into at least two WSUs, one corresponding to a verb and another to its object.

EXAMPLE 1

a. 吃_鱼 ‘to eat fish’

b. 学_滑冰 ‘to learn skiing’

c. 写_信 ‘to write a letter’

d. 写_文章 ‘to write an article’

e. 写_论文 ‘to write a thesis’

f. 写_书 ‘write a book’

There are, however, strings of words in the form of verb-object that are commonly used in a compact form. These forms are considered as words and thus treated as single WSUs.

EXAMPLE 2

a. 开会 ‘meeting’

b. 跳舞 ‘dancing’

c. 解决吃饭问题 ‘to resolve the problem of meals’

d. 孩子该念书了 ‘it's time for the child to go to school’

Verb-object forms, whether a phrase or a word, are segmented into two separate WSUs, if some expression is inserted into those forms.

EXAMPLE 3

a. 吃_两顿_饭 ‘have two meals’

b. 跳_新疆_舞 ‘to dance the Xinjiang dance’

In the second example, for instance, the place name 新疆 (Xinjiang) is inserted between the verbal word form 跳舞 ‘dancing’. Hence, this verbal form is segmented into two WSUs, 跳 ‘to dance’ and 舞 ‘a dance’ as well as into the inserted form 新疆, a WSU in itself.

6.2.2.5 Verb–complement word structures

A single-character verb, adjective or adverb that is followed by a complement is not segmented, but treated as a single WSU.

EXAMPLE 1

a. 打倒 ‘to knock down often in a political sense’

b. 提高 ‘to improve’

c. 加长 ‘to lengthen’

d. 做好 ‘to do well’

A two-character verb, adjective or adverb which is followed by a complement is segmented into two WSUs: one corresponding to the verb, adjective or adverb and another to the complement.

EXAMPLE 2

a. 整理_好 ‘clean up well’

b. 说_清楚 ‘speak clearly’

c. 解释_清楚 ‘explain clearly’

If a character such as 得 ‘able’ or 不 ‘not’ is inserted into a one-character verb followed by a complement, then such a form is segmented into three WSUs.

EXAMPLE 3

a. 打_得_倒 ‘able to knock down’

b. 提_不_高 ‘unable to improve’

6.2.2.6 Adverb-delimited verbs

Commonly used adjectives that are composed with a noun or a noun phrase in a compact manner are treated as single WSUs.

EXAMPLE 1

a. 胡闹 ‘make trouble’

b. 瞎说 ‘talk nonsense’

c. 死记 ‘learn by rote’

d. 早来 ‘come early’

e. 晚走 ‘go late’

f. 重说 ‘retell’

Compound directional verbs are treated as single WSUs.

EXAMPLE 2

a. 出去 ‘go out’

b. 进来 ‘come in’

Compound directional verbs that include characters like “得” or “不” are, however, are segmented into three WSUs.

EXAMPLE 3

a. 出_得_去 ‘able to go out’

b. 进_不_来 ‘unable to come in’

Verbal phrases formed with a directional verb such as 来 ‘come’ or 出去 ‘go out’ are segmented into two WSUs.

EXAMPLE 4

a. 寄_来 ‘send to’

b. 跑_出去 ‘run out’

6.2.2.7 Sequences of independent single verbs

Independent one-character verbs in a sequence or coordinate form without a conjunction are each segmented into a separate WSU.

EXAMPLE 1

a. 苫_盖 ‘to cover and cover’

b. 听_说_读_写 ‘to listen, speak, read and write’

Multi-character verbs in a sequence without a conjunction are each segmented into a separate WSU.

EXAMPLE 2

a. 调查_研究 ‘to investigate and do research’

b. 宣传_鼓动 ‘to publicize and instigate’

6.2.3 Adjectives

6.2.3.1 Reiteratively combined adjectives

Adjectives with a reiterative form “AA”, “AABB”, “ABB”, “AAB” or “A+里+AB” are treated as single WSUs.

EXAMPLE 1

a. 大大 ‘big’

b. 高高 ‘tall’

c. 高高兴兴 ‘happy’

d. 匆匆忙忙 ‘busy’

e. 绿油油 ‘fresh green’

f. 红彤彤 ‘bright red’

g. 蒙蒙亮 ‘daybreak’

h. 马**里**马虎 ‘careless’

Adjectives in the reiterative form “ABAB” are, however, segmented into two WSUs.

EXAMPLE 2

a. 雪白_雪白 ‘snowy white’

b. 滚圆_滚圆 ‘fat and round’

6.2.3.2 Adjectival phrases

Adjectival phrases in the form “一 A 一 B”,“一 A 二 B”, “半 A 半 B”, “半 A 不 B” or “有 A 有 B” are not segmented, but treated as single WSUs.

EXAMPLE

a. 一心一意 ‘whole-heartedly’

b. 一清二楚 ‘as plain as daylight’

c. 半明半暗 ‘partly bright partly dark’

d. 半生不熟 ‘half-cooked’

e. 有条有理 ‘orderly’

6.2.3.3 Adjectives in a sequential form without a conjunction

Two single-character adjectives in a sequence with contrasting features are not segmented, but treated as single WSUs.

EXAMPLE 1

a. 长短 ‘long and short’

b. 深浅 ‘deep and shallow’

c. 大小 ‘big and small’

Adjectives in a sequence with their original meaning are each segmented into a separate WSU.

EXAMPLE 2

a. 大_小_尺寸 ‘big and small in size’

b. 光荣_伟_大 ‘glory great and big’

6.2.3.4 Adjective-delimited nouns for colours

Adjectives and phrases denoting colour are not segmented, but treated as single WSUs.

EXAMPLE

a. 浅黄 ‘light yellow’

b. 橄榄绿 ‘olive green’

6.2.3.5 Adjectival phrases

Adjectival phrases combining a positive and negative form to indicate a question are segmented into three WSUs.

EXAMPLE 1

容易_不_容易 ‘easy or not easy?’

Such adjectival phrases are, however, not segmented, but treated as single WSUs, if they are partially in an elliptical form.

EXAMPLE 2

容不容易 ‘easy or not (easy)?’

6.2.4 Pronouns

Single-character pronouns containing the plural form 们 are treated as single WSUs.

EXAMPLE 1

a. 我们 ‘we’

b. 你们 ‘you’

c. 它们 ‘they’

d. 他们 ‘they’

Pronouns that are formed with a characters such as 这 ‘proximate’, 那 ‘remote’ or 哪 ‘interrogative’ and then followed by a unit word such as 个 ‘item’, 些 ‘selection’, 样 ‘temporal’, 么 ‘reason’, 里 ‘place’ or 边 ‘place’ are treated as single WSUs.

EXAMPLE 2

a. 这个 ‘this’ 这么 ‘thus’ 这边 ‘here’

b. 那些 ‘those’ 那样 ‘then’ 那里 ‘there’

c. 哪个 ‘which’ 哪里 ‘where’ 哪些 ‘which’

Pronouns that are formed with a character 这 ‘proximate’, 那 ‘remote’ or 哪 ‘interrogative’ but are followed by a numeral, unit word or noun are segmented into two separate WSUs.

EXAMPLE 3

a. 这_十 天 ‘these 10 days’

b. 那_人 ‘that person’

c. 那_种 ‘that kind’

Interrogative adjectives or phrases are all treated as single WSUs.

EXAMPLE 4

a. 多少 ‘how many’

b. 怎样 ‘what about’

c. 为什么 ‘why’

d. 什么 ‘what’

Pronouns such as 各 ‘each’, 每 ‘each’, 某 ‘a certain’, 本 ‘proximate’, 该 ‘proximity’, 此 ‘proximity’ and 全 ‘whole’ are segmented from a following measure word or noun, thus constituting two separate WSUs.

EXAMPLE 5

a. 各_国 ‘each country’

b. 每_种 ‘each type’

c. 某_工厂 ‘a certain factory’

d. 本_部门 ‘this department’

e. 该_单位 ‘this unit’

f. 此_人 ‘this people’

g. 全_校 ‘whole school’

6.2.5 Numerals

Numerals are segmented from measure words.

EXAMPLE 1

a. 三_个 ‘three (item)’

b. 一_种 ‘one type’

Numerals are treated as single WSUs.

EXAMPLE 2

一亿八千零四万七百二十三 ‘180,040,723’

The ordinal prefix 第 is segmented as an independent WSU from a numeral that follows it.

EXAMPLE 3

a. 第_一 ‘the first’

b. 第_四 ‘the fourth’

c. 第_五十三 ‘the 53rd’

The term 分之 ‘part of’ that is used with fractional numerals is treated as a separate WSU.

EXAMPLE 4

a. 五_分之 - 三 ‘3 over 5’

b. 百_分之 - 二 ‘2/100’

c. 万_分之 - 五 ‘5/10000’

Sequences of numerals indicating approximate numbers are treated as separate WSUs.

EXAMPLE 5

a. 八九_公斤 ‘8 or 9 kg’

b. 十七八_岁 ‘17 or 18 years old’

Characters such as 多, 来 or 几 that occur after numerals for indicating approximate numbers are segmented as separate WSUs.

EXAMPLE 6

a. 两点_多 ‘past two o'clock’

b. 一千_ 多_人 ‘more than one thousand people’

c. 十_来_家 ‘about ten’

d. 十_几_个 ‘over ten’

Characters such as 些, 一些, 点儿 or 一点儿 that occur after adjectives or verbs to indicate approximate numbers are segmented as separate WSUs.

EXAMPLE 7

a. 大_些 ‘bigger’

b. 懂_一些 ‘know some’

c. 快_点儿 ‘quickly’

d. 快_一点儿 ‘more quickly’

Characters such as 近, 约, 成 or 数 that occur before numerals or numerical digits to indicate approximate numbers are segmented as separate WSUs.

EXAMPLE 8

a. 近_千人 ‘nearly one thousand people’

b. 约_三百 ‘about three hundred’

c. 成_百 ‘hundreds of’

d. 数_万 ‘ten thousand’

e. 数_千 ‘thousands of’

6.2.6 Measure words

Reiterative measure words are not segmented.

EXAMPLE 1

a. 年年 ‘every year’

b. 天天 ‘every day’

c. 个个 ‘each’

d. 家家户户 ‘every household’

Compound measure words or phrases are treated as WSUs.

EXAMPLE 2

a. 人年 ‘man per year’

b. 人次 ‘man per time’

c. 架次 ‘sortie, flying out’

d. 吨公里 ‘ton per kilometre’

6.2.7 Adverbs

Adverbs are treated as separate WSUs.

EXAMPLE 1

a. 很_好 ‘very well’

b. 都_来_了 ‘every has come’

c. 刚_走 ‘just gone’

d. 互相_协助 ‘help each other’

Commonly used adverbial phrases are treated as single WSUs.

EXAMPLE 2

a. 越来越 ‘more and more’

b. 不得不 ‘necessarily’

c. 不能不 ‘cannot but’

Phrases of the forms such as 越…越… and 又…又… as well as other similar phrases in a sequential or coordinate form are also treated as single WSUs.

EXAMPLE 3

a. 越走越远 ‘to go further and further’

b. 又香又甜 ‘savoury and sweet’

6.2.8 Prepositions

Prepositions are treated as separate WSUs.

EXAMPLE

a. 生_ ‘born in’

b. 走向_胜利 ‘up to, towards, heading for success’

c. 按照_规定 ‘according to regulations’

6.2.9 Conjunctions

Conjunctions are treated as separate WSUs.

EXAMPLE

a. 工人_和_农民 ‘worker and farmer’

b. 光荣_而_伟大 ‘glorious and grand’

6.2.10 Auxiliary words

Auxiliary markers such as 的 (adjective marker), 地 (adverb marker), 得(marker for effect after a verb) and 之 (possessive marker) are treated as WSUs.

EXAMPLE 1

a. 他_的_书 ‘his book’

b. 美丽_的_城市 ‘beautiful city’

c. 中国_的_大熊猫 ‘Chinese panda’

d. 慢慢_地_走 ‘walk slowly’

e. 说 _得_快 ‘speak fast’

f. 成功_之_路 ‘road to success’

Tense or aspect markers 着 (present continuous, progressive), 了(past or perfective, completed action marker) and 过 (past or perfective, experienced action marker) are treated as separate WSUs.

EXAMPLE 2

a. 看_着 ‘to be watching’

b. 看_了 ‘watched’

c. 看_过 ‘have watched’

The relative construction marker 所 ‘what, that which’ is treated as a separate WSU, irrespective of the verb that follows it.

EXAMPLE 3

a. 所_想 ‘what one thinks’

b. 所_认识 ‘what one recognizes’

6.2.11 Modal words

Modality or mood markers such as 吗 (ma, question marker) or 吧 (ba, marker for polite suggestion) are treated as separate WSUs.

EXAMPLE

a. 你好_吗 (ma)? ‘How are you?’

b. 你好_吧 (ba)? ‘Is everything OK?’

6.2.12 Exclamations

Exclamation markers such as 啊 (for surprise or approval) or 唉呀 (aiya, for realization or agreement) are treated as WSUs.

EXAMPLE

a. 啊_真美! ‘How beautiful it is!’

b. 唉呀_他走了! ‘He has gone!’

6.2.13 Imitative words

Expressions that imitate sounds are treated as WSUs.

EXAMPLE

a. 嘟 ‘toot, honk’

b. 当当 ‘tinkle’

c. 轰隆隆 ‘rumble’

7 Specific rules for identifying WSUs in Japanese text

7.1 Bunsetsus

All bunsetsus are WSUs.

7.2 Lexical items

7.2.1 General rule

There are exactly nine parts of speech in Japanese:

— noun (名詞 meishi);

— verb (動詞 doushi);

— adjective with two subcategories, adjective proper (形容詞 keiyoushi) and adjectival verb (形容動詞 keiyoudoushi);

— adnoun (連体詞 rentaishi);

— adverb (副詞 fukushi);

— exclamation (感動詞 kandoushi);

— conjunction (接続詞 setsuzoushi);

— particle (助詞 joshi); and

— auxiliary verb (助動詞 jodoushi).

These parts of speech form the basis for identifying WSUs.

A string of characters that belongs to one of these parts of speech, except for the particle and auxiliary verb parts of speech, is treated as a WSU.

7.2.2 Nouns

7.2.2.1 Nouns in general

Nouns are WSUs.

EXAMPLE 1

a. 学校, gakkou, ‘school’

b. 開始, kaishi, ‘starting, beginning’

Used as part of a sentence, nouns (N) often agglutinate with particles (P) or auxiliary verbs (AV), thus forming bunsetus.

EXAMPLE 2

a. 学校へ
gakkou+e
N P
‘school’ ‘to’ := ‘to school’

b. 桜+です
sakura+desu
N AV
‘cherry blossom’ Copula := ‘ be cherry blossom’

Each of the examples (2 a. and 2 b.) is a bunsetu, and is therefore a WSU. Their constituent nouns, 学校 (gakkou) and 桜 (sakura), are also busetus, for they are nouns for themselves.

Nouns are subcategorized into

— common nouns (CN) or abstract nouns (AN),

— proper nouns (PN),

— pronouns (PRN), and

— measure nouns (MN).

7.2.2.2 Common nouns and abstract nouns

Whether common or abstract, all nouns are WSUs. Examples of simple common nouns with no internal morphological structure are: 桜 (sakura) 'cherry blossoms', 靴 (kutsu) 'shoes', 学校 (gakkou) 'school' and 犬 (inu) 'dog'. As these are nouns, they are all treated as single WSUs. Examples of complex abstract nouns formed with affixes (AX) are shown in Example 1. They are also treated as single WSUs.

EXAMPLE 1

a. 不参加
hu+sanka
AX AN
'non' 'participation' := 'non-participation, absence'

b. 賃貸料
chintai+ryou
AN AX
'rent' 'fee' := 'rental fee'

Unlike simple nouns, each noun that forms a complex noun with one or more affixes is also a WSU; the component noun 参加 (sanka) 'participation', for instance, is a WSU.

Compounds are each treated as single WSUs. Here, each of the nouns that constitute each compound noun is also a WSU.

EXAMPLE 2

a. 頭皮
tou_hi
N N
'head' 'skin' := 'scalp'

b. 髪_飾り
kami_kazari
N N
'hair' 'dressing item' := 'hair dressing item'

In this example, the compound noun 頭皮 (touhi) 'scalp' as a whole is a WSU and so are its component nouns, 頭 (tou) 'head' and 皮 (hi) 'skin'.

7.2.2.3 Proper nouns

Proper nouns (PN) are treated as single WSUs. If they are personal names, each of which consists of a surname (family name) and a given name, then they may also be segmented into two WSUs.

EXAMPLE 1

鈴木_一郎
Suzuki_Ichiro
surname given
'Ichiro Suzuki'

Surnames with titles are segmented into two WSUs.

EXAMPLE 2

田中_ 教授
tanaka_kyouju
PN CN
'Tanaka' 'professor' = 'Professor Tanaka'

Names that refer to a country, a nation, a language, an organization, etc., are treated as single WSUs.

EXAMPLE 3

a. 富士山
Fuji+san
'Fuji' 'mountain' = 'Mt. Fuji'

b. 東京都
tokyo+to
'Tokyo Prefecture'

c. 国際+ 標準化機構
kokusai+hyoujunka+kikou
'International' 'standardization' 'organization'
:= 'International Organization for Standardization'

7.2.2.4 Pronouns

Pronouns (PRN) are all treated as WSUs. Examples of personal pronouns that are treated as WSUs are shown below.

EXAMPLE 1

a. 私 (watashi) 'first person singular'

b. あなた (anata) 'second person singular'

c. 彼 (kare) 'third person singular, masculine'

d. 彼女 (kanojo) 'third person singular, feminine'

Plural personal pronouns are marked with the plural suffix たち(tachi) for the second person or the plural suffix ら (ra) for the third person. Each of these plural pronouns is treated as a single WSU.

EXAMPLE 2

a.あなたたち (anata+tachi) 'second person plural'

b. 彼+ら (kare+ra) 'third person plural'

The following example shows demonstrative pronouns, both singular and plural, that are again treated as WSUs.

EXAMPLE 3

a. それ (sore) 'it'

b. これ (kore) 'this'

c. あれ (are) 'that'

d. それら (sore+ra) 'they'

e. これら (kore+ra) 'these'

f. あれら (are+ra) 'those'

Locative pronouns that are treated as WSUs are shown in Example 4.

EXAMPLE 4

a. そこ (soko) ‘there’

b. ここ (koko) ‘here’

c. こちら (kocjira) ‘here, over here’

d. あちら (achira) ‘there, over there’

There are also locative compound pronouns which are again treated as single WSUs.

EXAMPLE 5

a. あちこち (achi+kochi) ‘here and there’

b. あちらこちら (achira+kochira) ‘here and there’

Interrogative pronouns that are treated as WSUs are shown in Example 6.

EXAMPLE 6

a. どれ (dore) ‘which’

b. 何 (nani) ‘what’

c. いつ (itsu) ‘when’

e. 誰 (dare) ‘who’

f. どこ (doko) ‘where’

g. いくつ (ikutsu) ‘how many’

h. どう (dou) ‘how’

7.2.2.5 Measure nouns

Measure nouns (MN), each consisting of a numeral (NUM) which is often followed by a unit expression (U), are treated as single WSUs.

EXAMPLE 1

a. 2つ
futa.tsu
‘two (pieces, items)’

b. 三分 as in 三分の一 (sanbunno_ichi) ‘one third, one out of three parts’
san+bun
NUM U
‘three’ ‘parts’ := ‘three parts, dividing into three parts’

c. 5 分間
go+fun+gan
‘five’ ‘minutes’ ‘duration’ := ‘for five minutes’

d. 第一位
dai+ichi+i
‘first’ ‘place’ := ‘the first place’

e. 3 番目
san+banme
‘three’ ‘turn’ := ‘the third turn, place’

f. 4 本 as in 鉛筆 4 本 (enpitsu yonhon) ‘four pencils’
yon+hon
‘four’ ‘items’ := ‘four items’

7.2.3 Verbs

Verbs are WSUs. Each verb in Japanese consists of one or more stems followed by a nonnull sequence of suffixes that are called ‘endings’. These endings are of various forms: base, imperative, assumptive, negative, adverbial or adnominal forms. Neither verb stems nor endings alone constitute WSUs.

Japanese verbs are subcategorized into

— simple verbs,

— compound verbs,

— ‘*suru*’ (do)-type verbs, and

— ubsidiary verbs.

Simple verbs are verbs which consist of a single stem followed by one or more endings, whereas compound verbs may have two or more stems followed by endings. Examples follow.

EXAMPLE 1

a. 飲む (no+mu) ‘to drink’

b. 見+送+る(mi+oku+ru) ‘to see and send off’

Example 1 a. is a simple verb, consisting of a stem 飲 (no) and a base form ending む (mu). Example 1 b., on the other hand, consists of two verb stems 見 (mi) and 送 (oku) followed by a base form ending る (ru).

Some verbs called *‘sahendoushi’* each consist of an action noun (N) followed by a verb ‘su+ru’ which means ‘do’. These verbs are treated as single WSUs.

EXAMPLE 2

a. 勉強する
benkyou+su+ru
N Vst E
‘study’ ‘do’ Base form := ‘to study’

b. 合格する
goukaku+su+ru
N Vst E
‘passing’ ‘do’ Base form := ‘to pass (an exam)’

Verbs that consist of stems (Vst) and endings (E) often combine with auxiliary verbs (AV) or particles (P). Example 3 a. shows a verb form with an ending ら(ra) that combines with a negative auxiliary verb ない (nai) to form a negative verb.

EXAMPLE 3

a. 入らない
hai+ra+nai
Vst E AV
'enter' 'not' := 'not enter'

b. 飲むだろう
no+mu+darou
Vst E AV
'drink' 'expected' := 'to be expected to drink'

c. 飲むだろうね
no+mu+darou+ne
Vst E AV P
'drink' 'expected' tagQuestion := 'to be expected to drink, isn't it?'

d. 合格するだろうね
goukaku+su+ru+darou+ne
N Vst E AV P
'passing' 'do' 'expected' := '(He) will pass (an exam), won't he?'

Here, each of these verbs is treated as a single WSU, for auxiliary verbs as well as endings and particles are not treated as independent WSUs.

Subsidiary verbs are verbs that are supplemented with aspectual or modal meanings. See Example 4.

EXAMPLE 4

a. 話し+て+いる
hanashi+te+ir+u
Vst E E E
'speak' Conj Progressive := 'be speaking',
where the ending て (te) is a conjunctive ending needed to bridge the verb stem with the progressive ending

b. 読み過ぎる
yomi+sugi+ru
Vst E E
'read' 'excessive' := 'overead, read too much'

These verbs are treated as single WSUs. The subsidiary expressions いる (i+ru) and 過ぎる(su+gi+ru) are not auxiliary verbs, but treated as endings, only supplementing the meaning of the verb stems that precede them.

7.2.4 Adjectives

Adjectives are treated as WSUs. Like verbs, Japanese adjectives have inflectional endings.

There are two typical endings, い (i) and な (na), that classify adjectives to two major types:

— i-type adjectives;

— na-type adjectives.

EXAMPLE 1

a. 黒い
kuro+i
'black'

b. 静か+な
shizuka+na
'quiet'

Morphological structures subcategorize adjectives into three classes:

1) simple adjectives;

2) derived adjectives;

3) compound adjectives.

As in Example 1, simple adjectives each consist of a single adjective stem (Ast) followed by a sequence of endings. On the other hand, the stems of derived or compound adjectives are more complex; the stems of derived adjectives are often nouns with adjectival affixes (AX) and endings (E), while those of compound adjectives have two or more adjectival stems (Ast).

EXAMPLE 2

a. 薄+暗+い
usu+gura+i
AX N E
'thin' 'darkness' := 'dusky'

b. 都会+的+な
tokai+teki+na
N AX E
'town' := 'urbane'

c. 青+白+い
ao+jiro+i
N N E
'blue' 'white' := 'pale'

Despite differences in their morphological composition, all of these adjectives are treated as single WSUs.

7.2.5 Adnouns

Adnouns are treated as WSUs. Just as adverbs modify verbs, adnouns modify nouns. Unlike adjectives that also modify nouns, adnouns do not have inflectional endings. In this respect, adnouns in Japanese as well as in Korean resemble determiners in English.

EXAMPLE

a. あらゆる (arayuru) 'every, all sorts of'

b. この (kono) 'this'

7.2.6 Adverbs

Adverbs are treated as WSUs. Adverbs have no inflectional endings. They modify verbs, adjectives and even sentences.

EXAMPLE

a. やっと (yatto) 'at last'

b. 幸運にも (kouun.nimo) 'fortunately'

c. スイスイ (suisui) 'smoothly'