L2/11-108

ISO/FDIS 24614-2: Language resource management -- Word segmentation of written texts -- Part 2: Word segmentation for Chinese, Japanese and Korean

NOTE: Permission is granted by the American National Standards Institute to reproduce this International Standard for the purpose of review and comment related to the preparation of a U.S. position, provided this notice is included. All other rights are reserved.

DRAFT INTERNATIONAL STANDARD ISO/DIS 24614-2

ISO/TC **37**/SC **4** Secretariat: **KATS**

Voting begins on: **2010-05-26**

Voting terminates on: **2010-10-26**

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION • МЕЖДУНАРОДНАЯ ОРГАНИЗАЦИЯ ПО СТАНДАРТИЗАЦИИ • ORGANISATION INTERNATIONALE DE NORMALISATION

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

ICS 01.140.10

In accordance with the provisions of Council Resolution 15/1993 this document is circulated in the English language only.

Conformément aux dispositions de la Résolution du Conseil 15/1993, ce document est distribué en version anglaise seulement.

To expedite distribution, this document is circulated as received from the committee secretariat. ISO Central Secretariat work of editing and text composition will be undertaken at publication stage.

Pour accélérer la distribution, le présent document est distribué tel qu'il est parvenu du secrétariat du comité. Le travail de rédaction et de composition de texte sera effectué au Secrétariat central de l'ISO au stade de publication.

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENT AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

© International Organization for Standardization, 2010

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

Copyright notice

This ISO document is a Draft International Standard and is copyright-protected by ISO. Except as permitted under the applicable laws of the user's country, neither this ISO draft nor any extract from it may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, photocopying, recording or otherwise, without prior written permission being secured.

Requests for permission to reproduce should be addressed to either ISO at the address below or ISO's member body in the country of the requester.

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

Reproduction may be subject to royalty payments or a licensing agreement.

Violators may be prosecuted.

© ISO 2010 – All rights reserved

Contents

Page

© ISO 2010 – All rights reserved

 © ISO 2010 – All rights reserved

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*.

This second/third/... edition cancels and replaces the first/second/... edition (), [clause(s) / subclause(s) / table(s) / figure(s) / annex(es)] of which [has / have] been technically revised.

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*

© ISO 2010 – All rights reserved

Introduction

Word segmentation challenges technology of natural language processing when it concerns written text with no words boundaries like Chinese, Japanese, and pre-modern Korean texts. Such problem doesn't exist in texts like English text where words are separated by a space.

Part 2 focuses on word segmentation for Chinese, Japanese, and Korean. In regard to typography, both Chinese and Japanese texts don't display any space between their different written forms while Korean texts contain some fragments (oejeols) separated by a space. Due to the fact these three languages share similarities in words composed of Chinese characters, general rules for identifying "word segmentation units" (WSU) in Chinese text can also be applied to the processing for Japanese and Korean to some extent.

In Part 2, the general rules for identifying WSU in Chinese, Japanese, and Korean will be described; then will follow the specific rules for each of these three languages.

© ISO 2010 – All rights reserved

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 for word segmentation defined in Part 1 are applied for Chinese, Japanese and Korean (CJK). The objective of the word segmentation is to suit the requirements for the computational applications of language resources, for the natural language processing, and for other specific applications such as IR (information retrieval) and MT (machine translation). Part 2 is restricted to a particular task delineated by word segmentation, which is distinct from morphological or syntactic analysis per se, although word segmentation greatly depends on morpho-syntactic analysis. The main task of Part 2 is to define word segmentation unit for Chinese, Japanese and Korean. Although they are related to each other at the lexical level, each of these three languages has distinct structural differences and these differences have to be reflected on the definition of word segmentation and its practical guidelines. Due to the fact that these three languages share similarities in words composed of Chinese characters, general rules for identifying word segmentation units (WSU) in Chinese text can also be applied to the processing for Japanese and Korean to some extent.

2 Normative references

The following referenced documents are indispensable for the application of Part 2. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO DIS 24611, *Language resource management — morphosyntactic anotation framework*

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

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

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO DIS 24611, ISO 24613:2008, and ISO 24614-1, in addition to the following definitions, apply.

3.1
phrase
group of words forming a conceptual unit and being a component of a sentence that carries a grammatical function

3.2
bunsetsu
phrase (3.1) in Japanese text without internal modifying relations

© ISO 2010 – All rights reserved

EXAMPLE The sentence “私は学校へ 早く 行きました(I went to school early)。” consists of four bunsetsu: 私は (watashiwa), 学校へ(gakkoue), 早く(hayaku) 行きました(ikimashita). “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)” which is an auxiliary verb for a politeness, and “た(ta)” is an auxiliary verb for a past tense. The sentence contains four bunsetsu.

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

3.3
eojeol
malmadi
phrase (3.1) in Korean text without internal modifying relations separated by a space

EXAMPLE A sentence “나는 학교에 일찍 갔다(I went to school early)” consists of four eojeols: “나는(naneun)”, “학교에(hakgyoe)”, “일찍(iljjik)”, and “갔다(gatta)”. “나(I)” is a pronoun, “는” is a grammatical affix, “학교(hakgyo; noun; school)” is a noun, “에” is a grammatical affix, “일찍(early)” is an adverb, “가(go)” is a verbal stem followed by two grammatical affixes: “았” and “다” .

NOTE1 An eojeol normally consists of a noun plus its particle(s) or a verb plus its ending(s), auxiliary verb(s), and particle(s) as shown in the example above.

NOTE2 An eojeol is also called “word phrase“. Eojeol (word phrase) consists of one or more word forms. Auxiliary words can concatenate to word unit standing in front. E.g. "살아있다(to keep alive)" is composed of two word form; 살아(to live) and 있다(keep).

3.4
particle
part of speech (known as *joshi* in Japanese) to perform a semantic, grammatical and/or discursive function.

NOTE1 Japanese particles cannot be used independently; they follow a word, a clause or even a sentence. They mainly function as a marker of a case, as a connective, or as a conveyor of some trivial meaning. Like a suffix they are sometimes attached to a word, but they don’t have any inflectional ending. They also differ from the suffix by being considered as a part-of-speech.

EXAMPLE1 The particle “ね(ne)” in “寒いね？ (It is very cold, isn’t it?)” is corresponding to “isn’t it ?”

NOTE2 The combination of a word followed by a particle is a “bunsetsu”.

3.5
ending
inflectional part of verb, adjective and auxiliary verb in Japanese

NOTE1 A verb, adjective and auxiliary verb have inflections at the end of them, defined as ending. For example, as the ending of a verb, there are a negation form, an adverbial form, a base form, an adnominal form, an assumption form, or an imperative form.

NOTE2 Inflections that are localized at the ended part of a verb, an adjective or an auxiliary verb are defined as “ending”. The ending of verbs can be a negative form, and adverbial form, a base form, an adnominal form, an assumption form or an imperative form.

3.6
measure word
part-of-speech in Chinese to define, along with numbers, the quantity of a given object, or to identify specific objects with demonstrative pronouns such as "this" and "that".

NOTE1 While 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.

© ISO 2010 – All rights reserved

NOTE2 There is a set of "verbal measure words" used for counting the number of times an action occurs, rather than counting a 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), “次”(ci) functions as a verbal measure word to modify the verb "去 (qu) ".

4 Overview

4.1 Introduction

This clause will first review the concept of word segmentation unit (WSU) which was introduced in Part 1. Then, some features shared by Chinese, Japanese and Korean (CJK) will be discussed (4.3).

4.2 Review of the concept of word segmentation unit

Word segmentation is the process of dividing a text into meaningful units called word segmentation units. Each word segmentation unit defines one concept, for example “the White House” consists of three words but designates one concept known as the President’s residence in USA. In other words, word count and concept are not similar and may differ between languages. 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 돼지 고기 (doeji-gogi).

A unit that carries a meaning useful for any syntactic processing can be defined as a word segmentation unit (WSU). It could be an entry of a lexicon or of any other type of words storage as far as such entry matches with for the syntactic processing required in natural language processing. In other words, the WSU’s dimension is more or less fixed; but syntactic interferences between compounds inside a WSU are not allowed. Such extensive/opened definition is useful for the further syntactic processing because some WSU that frequently occurred in corpora are not systematically decomposable by a syntactic processing.

4.3 Common Features among Chinese, Japanese, and Korean

There are two basic features common to Chinese, Japanese, and Korean that originate from a common cultural heritage in the Far Eastern part of Asia. First, Chinese characters have been used and are still used in this part of the world to some differences in the degree of its use. China totally depends on Hanji, while Japanese also uses Kana characters. On the other hand, Korean hardly uses Hanji characters, but has its own writing system. Second, many of the Chinese-originated words or phrases are used both in Japanese and Korean such as "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 kept in these languages or transliterated into the characters of the Korean writing system as in the case of Korean.

Because of this historical background, some principles of the Chinese word segmentation apply to Chinese derived words in Japanese and Korean to a large extent. If the word is derived from Chinese characters, three languages have common properties. If their word in noun consists of two or more Chinese characters, they will be one word segmentation unit if they are “tightly combined and steadily used” according to principles of Part 1. For example, “each country” in English is not a word segmentation unit as its translation “各|国”. If the last character is productive in a limited manner, it forms a word segmentation unit with the preceding word, for example, “東京都(Tokyo Metropolis)”, “8 月(August)” or “加速器(accelerator)”.

Because the motivation of word segmentation standard is to recommend what word segmentation units should be registered in a type of lexicon where it is not the lexicon in linguistics but any kind of practical indexed container for word segmentation units, it has two possibly conflicting principles. For example, principles of unproductivity, frequency, and granularity could cause conflicts because they have different perspectives to define a word segmentation unit.

The Chinese character derived nouns are sharable for its word segmentation unit structure for three languages, but not the whole. On the other hand, there are common things between Korean and Japanese. Some Korean word endings and Japanese auxiliary verbs have the same functions. Word segmentation in each language is somewhat different according to already made word segmentation regulation, even violating

© ISO 2010 – All rights reserved

one or more principles of word segmentation. It will be a starting point to recommend the more synchronized word segmentation unit concept in a multi-lingual environment. The concept of “word segmentation unit” is to broaden the view about what could be registered in lexicon of natural language processing purpose, without much linguistic representation.

NOTE This standard adopts a notation which uses the underline to indicate the presence of a WSU under consideration.

5 General rules for identifying WSU in Chinese, Japanese, and Korean text

5.1 Lexical Items

Each lexical item is a WSU.

NOTE Most of the items given below are lexical items.

5.2 Derivationally formed items

Each derivationally formed item is treated as a single WSU.

EXAMPLE1 (Chinese):

科学 家	物理 学 家
ke xue jia	*wu li xue jia*
science -er	physics -ology -er
noun suffix	noun suffix suffix
scientist	physicist

EXAMPLE2 (Japanese):

非−常勤	音楽−家
hi joukin	ongaku ka
Noun (prefix-noun)	Noun(noun-suffix)
non full-time	music professional-person
part-time working	musician

NOTE This example “音楽家, ongakuka (musician)” consisting of a noun and a suffix, “音楽,ongaku (musician)” and “家, ka, (professional person),” is a WSU. And also these two components are WSU.

EXAMPLE3 (Korean):

음악−가	헛−돌−다
eumak-ga	heosdolda
noun(prefix-noun)	verb (prefix+verb)
music artist	nothing + spin
musician	spin with no traction

NOTE Derivative affixes themselves are not treated as WSU.

5.3 Word compound

Each word compound is treated as a single WSU.

EXAMPLE1 (Chinese):

白菜
Baicai

© ISO 2010 – All rights reserved

Noun
white vegetable
Chinese cabbage

EXAMPLE2 (Japanese):

海外旅行
kaigai ryokou
noun noun
abroad travel
traveling abroad

NOTE This example, ”海外旅行,” consists of two nouns, “海外 kaigai, (abroad)” and “旅行, ryokou, (traveling),” which are WSUs. And whole “海外旅行” is also a WSU.

EXAMPLE3 (Korean):

손-목	바로-잡다
Sonmok	*baro_jabda*
Noun	adverb_verb
hand – neck	rightly + hold
Wrist	correct

5.4 Phrasal compound

Each phrasal compound is treated as a single WSU.

EXAMPLE1 (Chinese):

猪 肉	发电 厂
zhu rou	fa dian chang
pig meat	to generate electricity plant
noun noun	verb noun
pork	power plant

NOTE Phrasal compounds, frequently used in Chinese text and, mainly consisting of two or three characters, are WSU.

EXAMPLE2 (Japanese):

豚 肉
buta niku
pig meat
noun noun
pork

EXAMPLE3 (Korean):

돼지고기
Doeji gogi
noun noun
pig meat
Pork

© ISO 2010 – All rights reserved

5.5 Idioms

Idioms are WSU.

EXAMPLE1 (Chinese):

胸有成竹	欣欣向荣
xiong you cheng zhu	*xin xin xiang rong*
have a well-thought-out plan	Prosperous

NOTE Most idioms in Chinese consist of four characters.

EXAMPLE2 (Japanese):

腹が立つ
haraga_atatsu
phrase(Noun_particle+verb)
stomach occur
feel angry

EXAMPLE3 (Korean):

수박겉핥기	함흥차사 (咸興差使)
subakgeothalgi	*hamheungchasa*
noun	Noun
half knowledge	lost messenger

5.6 Fixed expressions

Fixed expressions such as proverbs and mottos are segmented as WSU.

EXAMPLE1 (Chinese):

对不起	春夏秋冬	由此可见
dui bu qi	*chun xia qiu dong*	*you ci ke jian*
sorry	spring summer autumn winter	this shows

不管	三	七	二	十	一
bu guan	*san*	*qi*	*er*	*shi*	*yi*
no matter	three	seven	two	ten	one

no mater what happens

失败	是	成功	之	母
shi bai	*shi*	*cheng gong*	*zhi*	*mu*
Failure	is	success	of	mother

Failure is the mother of success.

EXAMPLE2 (Japanese):

時は	金なり
Toki_wa	Kane_nari
Noun _particle	Noun_auxirialy verb
Time_topic marker	Money_copula

Time is money

© ISO 2010 – All rights reserved

NOTE This example has two bunsetsu, “時は(toki_wa)” and “金なり(kane_nari).”

EXAMPLE3 (Korean):

울며 겨자 먹기	한 마디-로 말해
ulmyeo gyeoja meokgi	*han madiro malhae*
verb_noun_verb	adnoun_noun_verb
cry_mustard_eateat	one word_with talk
no choice	in a word (speaking briefly)

5.7 Abbreviations

Abbreviations are treated as WSU.

EXAMPLE1 (Chinese):

科技	工农业
ke ji	*gong nong ye*
science and technology	industry and agriculture

NOTE Abbreviations in Chinese text mainly consist of two, three or four characters.

EXAMPLE2 (Japanese): 特急(tokkyuu; noun: super express)→特別(super) + 急行(express)

EXAMPLE3 (Korean): 의대[醫大](*uidae*,noun; medical university) : 의학(*uihak*; medicin) + 대학(*daehak*; college)

5.8 Transliterated loanwords

Transliterated loanwords are treated as WSU.

EXAMPLE1 (Chinese):

吉普	巧克力
ji pu	*qiao ke li*
Jeep	chocolate

EXAMPLE2 (Japanese): ジープ (jeep) チョコレート(chocolate)

NOTE Transliterated loan words in Japanese are normally written in kata kana.

EXAMPLE3 (Korean): 피아노 (piano) 바이올린(violin)

5.9 String of foreign or special characters

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

EXAMPLE Chomsky, F16, X-Ray, 1298, +, CO2 …

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

5.10 Component of a WSU

Some components of a WSU can be WSU.

NOTE Some WSU have an internal structure which organizes several WSU hierarchically. Such a structure can be manipulated at different granularity levels in the process of word segmentation according to the need of various

© ISO 2010 – All rights reserved

applications. 猪肉 in Chinese, for instance, can be treated as a single WSU for MT that translates it into a single word "pork" in English, whereas it can be treated as two WSU for IR that looks for two different ontological entities, pig and meat.

EXAMPLE1 (Chinese):

chocolate: WSU(巧克力)
pork: WSU(WSU(猪) WSU(肉))
physicist: (WSU(WSU(WSU(物理) WSU(学)) 家(WSU))
Mao Zedong: WSU(WSU(毛) WSU(泽东))

EXAMPLE2 (Japanese):

豚	肉
buta	*niku*
pig	meat
noun	noun
pork	

EXAMPLE3 (Korean):

돼지고기	
doeji	*gogi*
noun	noun
pig	meat
Pork	

6 Specific rules for identifying WSU in Chinese text

6.1 Lexical items followed by a nonsyllabic character 儿(r)

Lexical items followed by a nonsyllabic character 儿(r) are treated as single WSU.

NOTE This nonsyllabic character 儿(r) is often attached to nouns and sometimes verbs.

EXAMPLE

花儿	玩儿	悄悄儿
huar	*wanr*	*qiaoqiaor*
flower r	play r	quietly r
noun r	verb r	adverb r
Flower	Play	Quietly

6.2 Lexical items

6.2.1 Noun

A noun is a WSU, being subcategorized into a common noun and a proper noun.

6.2.1.1 Noun preceded by an adjective

The nominal expression of the form "adjective + noun", namely a noun preceded by an adjective, is segmented into two WSU, unless the meaning of the expression is not the sum of its parts.

EXAMPLE1 (for two WSU)

小 床

© ISO 2010 – All rights reserved

	xiao chuang
	small bed
	adjective, noun
	small bed

EXAMPLE2 (for single WSU)

	小 媳妇
	xiao xi wu
	small wife
	adjective, noun
	young wife

6.2.1.2 The localizer word

A localizer word (shows a direction or position) is treated as a WSU.

EXAMPLE

桌子上	长江以北
zuo zi shang	*chang jiang yi bei*
table above	the Yangtzi River the north
noun, localizer word	noun, localizer word
on the table	to the north of the Yangtzi River

6.2.1.3 The plural suffix “们” (*men*; -s)

The plural suffix “们” (*men*; -s) is treated as a WSU.

EXAMPLE

朋友	们
peng you men	
friend –s;	
noun –s	
friends	

NOTE In the following cases the plural suffix is not treated as a WSU.

人们	哥儿们	爷儿们
ren men	*ger men*	*yier men*
People	pals	guys

6.2.1.4 Time expressions

Time expressions are treated using the following rules:

6.2.1.4.1 January-December and Monday-Sunday are WSU.

EXAMPLE

五月	元月	3 月	星期 日	礼拜 三
wu yue	*yuan yue*	*3 yue*	*xing qi ri*	*li bai san*
five month	first month	3 month	Week + day	week three
May	January	March	Sunday	Wednesday

6.2.1.4.2 The time measure words “Year, day, hour, minute, second” are WSU.

© ISO 2010 – All rights reserved

EXAMPLE

1988 年 3 月 15 日	11 时 42 分 8 秒
1988 nian 3 yue 15 ri	*11 shi 42 fen 8 miao*
1988 year 3 month 15 day	11 hour 42 minute 8 second
March 15th,1998	forty two minute and eight second past eleven

6.2.1.4.3 The results of “前、后、上、下、大前、大后” (before last, after next, last, next, before before last, after after next) each combined directly with a time noun or a time measure word are WSU.

EXAMPLE

前天	后年	上星期	下月	大前天	大后年
qian tian	*hou nian*	*shang xingqi*	*xia yue*	*da qian tian*	*da hou nian*
before last, day	*after next, year*	last, week	next month	before before last, day	after after next, year
the day before yesterday	*the year after next*	last week	next month	three days ago	three years later

6.2.1.4.4 The time nouns “初一”(First day of a month in the Chinese lunar calendar) to “初十”(Tenth day of a month in the Chinese lunar calendar) are WSU.

6.2.1.5 Proper noun

6.2.1.5.1 Personal name and title

6.2.1.5.1.1 The full personal names of Han nationalities are WSU. Such a WSU can be further segmented into two WSU, i.e. the surname and the last name.

EXAMPLE

张 胜利	欧阳 志华
zhang sheng li	*ou yang zhi hua*
surname, given name	surname, given name
Zhang Shengli	Ouyang Zhihua

6.2.1.5.1.2 The full personal names of other nationalities or foreign countries are WSU each of which may have an internal structure in accordance with their own historical origins.

EXAMPLE

牛顿	小林 多喜二
niu dun	*xiao lin duo xi er*
Newton	Kobayashi Takiji

6.2.1.5.1.3 The expression structured as “surname + title” is segmented into two WSU.

EXAMPLE

张 教授	王 部长	李 师傅
zhang jiao shou	*wang bu zhang*	*li shi fu*
surname professor	*surname miniter*	*surname master*

© ISO 2010 – All rights reserved

professor Zhang minister Wang master Li

6.2.1.5.1.4 The expressions “one-character honorific title + surname” or “surname + one-character title” are WSU.

EXAMPLE

老张	陈总
lao zhang	*chen zong*
one-character honorific title	surname, one-character title
surname; old Zhang	manager Chen

6.2.1.5.1.5 The titles for kinship regarding rankings are WSU each with an internal structure.

EXAMPLE

三　叔	大　女儿
san shu	*da nv er*
three uncle	big daughter
the third younger uncle	the eldest daughter

6.2.1.5.2 Place name and nationality name

6.2.1.5.2.1 "族、省、市、州、县、乡、区、江、河、山 (nationality, province, city, prefecture, county, town, district, river, mountain)" are segmented into WSU independent of the proper names (e.g., nationality names or place names) that precede them.

NOTE In the case the preceding proper name is a single character, these Chinese character strings are not treated as WSU.

EXAMPLE

汉族 the Han nationality	哈萨克 族 the Kazakstan nationality
北京 市 Beijing Municipality	浙江 省(Zhejiang Province)
正定 县(Zhengding County)	忻县(Qi County)

6.2.1.5.2.2 Proper nouns that can bear more than one meaning are not treated as WSU.

EXAMPLE 牡丹江(Mudan River) 横断山(Hengduan Mountains)

6.2.1.5.2.3 Chinese names of street, road, village, town, ocean, or sea are treated as WSU.

EXAMPLE 长安街(Chang’an Avenue) 学院路(Xueyuan Road) 周口店(Zhoukoudian)

刘家村 (Liujiacun Village) 大西洋(Atlantic ocean) 地中海(Mediterranean Sea)

6.2.1.5.3 Other type of proper names

— Full country names are treated as single WSU.

EXAMPLE 中华人民共和国(People's Republic of China) 大不列颠及北爱尔兰联合王国(United Kingdom)

© ISO 2010 – All rights reserved

— Full names of organizations, agencies or institutions are segmented in accordance with their word segmentation structures.

EXAMPLE 联合国 教科文 组织(United Nations Educational, Scientific, and Cultural Organization)

中国 共产党(Communist party of China)

■ Trade mark, produce type or product series are segmented from the common nouns that precede them.

EXAMPLE 永久 牌(Yongjiu Brand) 中华 烟(Zhonghua Cigarette) 牡丹Ⅱ型(Peony II)

6.2.2 Verb

6.2.2.1 Various forms of reiterative verbs

a) A single-character-reiterated verb is treated as one WSU.

EXAMPLE 看看(look at) 动动(move)

b) A two-character-reiterated verb in the form of “AABB” is treated as one WSU.

EXAMPLE 来来往往(come and go) 拉拉扯扯(drag)

c) A verb reiterated in the form of “AAB, ABAB” is segmented into WSU.

EXAMPLE 说说 看(try to say) 研究 研究(to have a discussion)

d) A verb reiterated in the form of "A+'一'+A", "A+'了'+A", or "A+'了一'+ A" is segmented into WSU.

EXAMPLE 谈 一 谈 (have a good chat) 想 一 想(think carefully)

读 一 读(to read) 想 了 想(think it over)

想 了 一 想(think it over)

6.2.2.2 Verb delimited by a negative meaning Chinese character

The negative meaning Chinese character before a verb is treated as WSU.

EXAMPLE 不 写(not to write) 不 能(cannot) 没 研究(not to do research) 未 完成(having not yet been completed)

6.2.2.3 "Verb + a negative meaning Chinese character + the same verb" structure

A lexical structure that represents a question is segmented into WSUs.

EXAMPLE 说 不 说(say or not say)? 看 不 看(see or not see)? 相信 不 相信(believe or not believe)?

NOTE Yet the brachylogical form shall be one WSU, for example: 相不相信(believe or not)

6.2.2.4 Verb–object structure and verb collocations

A word of the form verb-object or a verb phrase that is compact and stabilized in use is not segmented.

EXAMPLE 开会(meeting) 跳舞(dancing)

© ISO 2010 – All rights reserved

解决吃饭问题(to resolve the problem of meals)

孩子该念书了(it's time for the child to go to school)

Verb phrases of the form verb–object or many other similar forms which are not compact are segmented.

EXAMPLE 吃 鱼(Eat fish) 学 滑冰(learn skiing)

写 信 (write a letter); (写 文章(write an article); 写 论文(write a thesis);写 书(write a book); …

A word or phrase of the form verb–object that is inserted in other expressions is segmented.

EXAMPLE 吃 两 顿 饭(have two meals) 跳 新疆 舞(to dance "Xinjiang dance")

6.2.2.5 Verb–complement word structure

A single-character verb, adjective or adverb of the form verb–complement is treated as one WSU.

EXAMPLE 打倒(to knock down (*often* politically) 提高(improve) 加长(lengthen) 做好(do well in)

NOTE1 The two-character verb or the two-character adjective or adverb in such forms is treated as a WSU. Then such a structure has two WSUs.

EXAMPLE 整理 好(clean up) 说 清楚(speak clearly) 解释 清楚(explain clearly)

NOTE2 If the Chinese character "得 or 不" is in between the such a word structure, the structure is broken and WSUs can be indentified, for example: 打 得 倒 (able to knock down) 提 不 高(unable to improve)

6.2.2.6 Adverb delimited verb

Adjectives with a noun or a noun phrase that are compact and stabilized in use are not segmented.

EXAMPLE 胡闹(make trouble) 瞎说(talk nonsense) 死记(learn by rote)

早 来(come early) 晚 走(go late) 重 说(retell)

Compound directional verbs are each treated as single WSU.

EXAMPLE 出去(go out) 进来(come in)

However, compound directional verbs into which characters like "得 or 不" inserted are segmented.

EXAMPLE 出 得 去(able to go out) 进 不 来(unable to come in)

Verbal phrases that are formed with a directional verb are segmented.

EXAMPLE 寄 来(send to) 跑 出 去(run out)

6.2.2.7 Combination of independent single verbs

Combinations of independent single verbs without a conjunction are segmented.

EXAMPLE 苫 盖(cover with) 听 说 读 写(listen, speaking, read and write)

Multi-word verbs without a conjunction are segmented.

EXAMPLE 调查 研究(investigate and research) 宣传 鼓动(publicity and instigation)

© ISO 2010 – All rights reserved

6.2.3 Adjective

6.2.3.1 Reiteratively combined adjective

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

EXAMPLE 大大(big) 高高(tall)

高高兴兴(happy) 匆匆忙忙(busy)

绿油油(fresh green) 红彤彤(bright red)

蒙蒙亮(daybreak) 马里马虎(careless)

However, adjectives with a reiterative form “ABAB” are segmented.

EXAMPLE 雪白 雪白(snowy white) 滚圆 滚圆(fat and round)

6.2.3.2 Adjective phrase

Adjective phrases with a 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.

EXAMPLE 一心一意(wholeheartedly) 一清二楚(as plain as daylight)

半明半暗(partly bright partly dark) 半生不熟(half-cooked)

有条有理(orderly)

6.2.3.3 Adjective in parataxis form

Adjectives in a parataxis form are segmented in accordance with the following rules:

a. Two single-character adjectives with word features varied are not segmented.

EXAMPLE 长短(long-short) 深浅(deep-shallow) 大小(big-small)

b. Adjectives in a parataxis form that maintain their original adjective meaning are segmented.

EXAMPLE 大小尺寸(size) 光荣 伟大(glory)

6.2.3.4 Adjective delimited noun for colors

A color adjective word or phrase is not segmented.

EXAMPLE 浅黄(light yellow) 橄榄绿(olive green)

6.2.3.5 Adjective phrase

Adjective phrases in a positive and negative form to indicate a question are segmented.

EXAMPLE 容易 不 容易(easy or not easy)

Yet the brachylogical phrase is not be segmented.

EXAMPLE 容不容易(easy or not)

© ISO 2010 – All rights reserved

6.2.4 Pronoun

a) Single-character pronouns with “们” are considered as WSU.

EXAMPLE 我们 (we) 你们(you) 它们(they) 他们(they)

b) “这、那、哪” with unit word “个” or “些、样、么、里、边” is considered as one WSU.

EXAMPLE 这个(this) 这么(thus) 这边(here)

那些(those) 那样(then) 那里(there)

哪个(which) 哪里(where) 哪些(which)

c) “这、那、哪” with numeral , unit word or noun word segmentation unit is segmented.

EXAMPLE 这 十 天(these 10 days) 那 人(that person) 那 种(that kind)

d) Interrogative adjectives or phrases are each considered as a WSU.

EXAMPLE 多少(how many) 怎样(what about)

为什么(why) 什么(what)

e) Pronouns such as “各、每、某、本、该、此、全”are segmented from a measure word or noun that follows any of them.

EXAMPLE 各 国 (each country) 每 种(each type)

某 工厂(a certain factory) 本 部门(this department)

该 单位(this unit) 此 人(this people)

全 校(whole school)

6.2.5 Numeral

a) A numeral is segmented from a measure word.

EXAMPLE 三 个(three) 一 种(one type)

b) Chinese digit words are treated as WSU.

EXAMPLE 一亿八千零四万七百二十三(180,040,723)

c) The ordinal prefix “第” is segmented from a numeral that follows it.

EXAMPLE 第 一 (first) 第 四(the fourth) 第 五 十 三(the fifty-third)

d) “分之” percent in a fractional number is treated as a WSU.

EXAMPLE 五 分之三(third fifth) 百分之二(2/100) 万分之五(5/10000)

e) Paratactic numerals indicating approximate numbers are treated as WSU.

EXAMPLE 八九 公斤(eight or nine kg.) 十 七八 岁(seventeen or eighteen years old)

© ISO 2010 – All rights reserved

f) “多、一些、点儿、一点儿”, used after adjectives or verbs for indicating approximate numbers, are segmented.

EXAMPLE 两 点 多(past two o’clock) 一 千 多 人(more than one thousand person)

十 来 家(about ten) 十 几 个(over ten)

g) “些、一些、点儿、一点儿” used after adjectives or verbs for indicating approximate numbers, are segmented.

EXAMPLE 大 些(bigger) 懂 一些(know some)

快 点儿(Quickly) 快 一点儿(more Quickly)

h) “近、约、数” etc., used before numerals or numerical digits for indicating approximate numbers are segmented.

EXAMPLE 近 千 人(near one thousand person) 约 三 百(about three hundred) 数 万(ten thousands)

成百(hundreds of) 数千(thousands of)

6.2.6 Measure word

a) Reiterative measure words are not be segmented.

EXAMPLE 年年(every year) 天天(every day) 个个(each) 家家户户(every household)

b) Compound measure words or phrases are treated as WSU.

EXAMPLE 人年 man/year 人次(man/time) 架次(sortie) 吨公里(t/km)

6.2.7 Adverb

a) Adverbs are treated as WSU.

EXAMPLE 很好(very good) 都来了(every one came here)

刚走(have just gone) 互相协助(help each other)

b) The following phrases that are used frequently as adverbs are treated as WSU:

EXAMPLE 越来越(more and more) 不得不(have to) 不能不(cannot but)

“越…越…、又…又…” and other phrases which function as a conjunction are segmented.

越 走 越 远(to go farther and farther) 又 香 又 甜(sweet yet savory)

6.2.8 Preposition

Prepositions are treated as WSU.

EXAMPLE 生于(be born in) 走向胜利(up to success) 按照规定(according to the regulations)

6.2.9 Conjunction

Conjunctions are treated as WSU.

© ISO 2010 – All rights reserved

EXAMPLE 工人和农民(worker and farmer) 光荣而伟大(glorious and grand)

6.2.10 Auxiliary word

a) Structural auxiliary words such as“的、地、得、之” are treated as WSU.

EXAMPLE 他的书 (his book) 慢慢地走(walk slowly) 说得快(speak fast)

美丽的城市(beautiful city) 中国的大熊猫(Chinese panda) 成功之路(road to success)

b) Tense auxiliary words such as “着()、了、过” are treated as WSU.

EXAMPLE 看着(be watching) 看了(watched) 看过(have watched)

c) The auxiliary word “所” is segmented from a verb that follows it.

EXAMPLE 所 想(what one thinks) 所 认识(what one knows)

6.2.11 Modal word

Modal words are treated as WSU.

EXAMPLE 你好吗？(How are you?)

你好吧！(Is everything OK?)

6.2.12 Exclamation word

Exclamation words are treated as WSU.

EXAMPLE 啊，真美！(How beautiful it is !)

唉呀，他走了！(He has gone!)

6.2.13 Imitative word

Imitative words are treated as WSU.

EXAMPLE 嘟(Du) 当当(tinkle) 轰隆隆(rumble)

7 Specific rules for identifying WSU in Japanese text

7.1 Bunsetsu

Each bunsetsu is a WSU.

NOTE As a component of "bunsetsu", there are mainly nine parts of speech. 名詞(*meishi*; noun), 動詞(*doushi*; verb), 形容詞・形容動詞(*keiyoushi, keiyoudoushi;* adjective), 連体詞(*rentaishi;* adnominal noun [only used in adnominal usage]), 副詞(*fukushi;* adverb), 感動詞(*kandoushi*; exclamation), 接続詞(*setsuzoushi;* conjunction), 助詞(*joshi*; particle), and 助動詞(*jodoushi*; auxiliary verb). These parts of speech are basis for identifying word segmentation units. Examples are provided in 6.2.

© ISO 2010 – All rights reserved

7.2 Lexical items

7.2.1 General rule

A string of characters that can be categorized as belonging to a part of speech is a WSU.

7.2.2 Noun

A noun is a WSU, being subcategorized into common nouns, proper nouns, pronouns, interrogative nouns, and numerals.

NOTE1 When a noun is a component constituting a sentence, it is usually followed by a particle or auxiliary verb, but there are exceptions. In some cases, one word becomes one sentence. For example, as a question, “なぜ(*naze*?; why?)”, as an answer, “りんご(*ringo* ; apple)”, “3 (san; three)” and so on.

NOTE2 Also, if a word like an adjective or an adnoun modifies a noun, a modifier (adjective, adnoun, and adnominal phrase) and a modificant (a noun) are segmented.

NOTE3 A bunsetsu that consists of a noun followed by a particle is considered as a single WSU and a noun followed by an auxiliary verb is also considered as a WSU. For becoming a component of a bunsetsu, Simple nouns, derivational nouns, and compound nouns can be applied to this rule. Also, every kind of nouns like common nouns, proper nouns, numerals, and so on can be applied to this rule for becoming a component of a bunsetsu.

EXAMPLE1 gakkou (school) and iku(go) are WSUs.

NOTE A particle and an auxiliary verb always follow a noun, a verb, an adjective, and some other categories. A particle and an auxiliary verb are not used independently, but they are each regarded as a part of speech in Japanese grammar. They are thus treated as WSU.

EXAMPLE2 gakkou_e (to school) consists of a noun and a particle

EXAMPLE3 The verb ikimashita (went_polite form) consists of a verb and two auxiliary verbs: one expresses politeness and the one the past tense.

EXAMPLE4 Noun followed by Particle for a case marker

私は	トマトを	買った。
watashi_wa	tomato _wo	kat_ta
Pronoun_particle	Noun_particle[object]	Verb_auxiliary verb
I	tomato	Bought

I bought a tomato.

EXAMPLE5 Noun followed by Auxiliary verb

私の	好きな	花は	桜です。
watashi_no	sukina	hana_wa	sakura_desu
Noun_particle	Adjective	Noun_particle	Noun_auxiliary verb [polite]
My	favorit	flower	is cherry blossoms

My favorit flower is cherry blossoms.

7.2.2.1 Common noun

7.2.2.1.1 Simple noun

Simple nouns like 桜(sakura, cherry blossoms), 靴(kutsu, shoes) , 学校(school) and “犬(inu, dog)” are WSUs.

© ISO 2010 – All rights reserved

7.2.2.1.2 Derivative noun

Derivative nouns with derivative affixes are each treated as a WSU.

EXAMPLE1 A noun with a prefix

不-参加
hu-sanka
noun(prefix-noun)
non-participant

EXAMPLE2 A noun with a suffix

賃貸-料
chintai-ryou
noun(prefix-noun)
lent al – fee

7.2.2.1.3 Compound noun

Compound nouns are each treated as a WSU.

EXEMPLE noun plus noun

頭-皮
touhi
noun
head-skin
scalp

7.2.2.1.4 Word combination

A word combination that is treated as a WSU may also be segmented for some practical need**:** prefix + noun, noun + suffix, noun + noun.

7.2.2.2 Proper noun

7.2.2.2.1 Japanese name and surname

Surnames (family names) and given names (first or personal names) are separated, but treated as single WSU.

EXAMPLE 鈴木一郎: (suzuki, surname) + (Ichiro, given name)

7.2.2.2.2 Person's name with following titles

Personal names or surnames that are followed by some titles or affixes are segmented as two WSU.

EXAMPLE1

田中	教授
tanaka	kyouju
proper noun	noun
one of surname	professor
prof. tanaka	

© ISO 2010 – All rights reserved

7.2.2.2.3 Other names

Names that refer to a country, a nation or a language, or toponyms in general are treated as single WSU.

EXAMPLE 富士山 (fujisan; proper noun; Mt. Baekdu)

Full names of an organization, agency, institution are treated as single WSU.

EXAMPLE 国際標準化機構 (kokusaihyoujunkakikou; International organization for Standardization)

7.2.2.3 Pronoun

7.2.2.3.1 Personal pronoun

7.2.2.3.1.1 General personal pronoun

General personal pronouns are treated as single WSU.

EXAMPLE 私 (watashi, I) , あなた (anata, you), 彼(kare, he) , 彼女(kanojo, she)

7.2.2.3.1.2 A personal pronoun with a suffix

A personal pronoun with a suffix is treated as a single WSU.

EXAMPLE あなた_たち(anata_tachi, you), 彼_ら(kare_ra)

7.2.2.3.2 Demonstrative pronoun

A pronoun is treated as a WSU.

EXAMPLE それ(sore,it) これ(kore,that) あれ(are,that)

A pronoun with a suffix is treated as a single WSU.

EXAMPLE それら(sore_ra, they), これら(kore_ra, these), あれら(are_ra, those)

A pronoun that refers to a place is treated as a single WSU.

EXAMPLE そこ（soko, there）, ここ(koko, here), あちら(achira, there), こちら(kocjira,here)

7.2.2.3.2.1 Compounding of pronouns

A compound pronoun is treated as a single WSU.

EXAMPLE あちこち(achikochi, here and there), あちらこちら(achirakochira, here and there)

7.2.2.4 Interrogative

An interrogative word is treated as a single WSU.

EXAMPLE1 どれ(dore, which), 何(nani, what), いつ(itsu, when), 誰(dare, who), どこ(doko, where), いくつ(ikutsu, how many), どう（dou, how）

© ISO 2010 – All rights reserved

NOTE Some interrogative nouns cannot be followed by case particles. However, interrogative nouns can be combined with auxiliary verbs in predicative position

EXAMPLE2

*どうは / *が / *を
*dou_wa / *dou_ga / *dou_wo
*noun[interrogative]_particle[topic]/ [subject]/[object]
*how _topic/subject/object marker
*how is

EXAMPLE3

天気は	どうですか
tenki_wa	dou_desu _ka
noun_particle	*noun[interrogative]_auxiliary verb/_auxiliary verb
whether_topic marker	*how _polite_question
how is the weather?	

7.2.2.5 Numeral/measure noun

A numeral noun is treated as a single WSU.

EXAMPLE1

ケーキを	三分の一に	分けた。
keeki_wo	sanbun'noichi_ni	wake_ta
noun_particle	noun[numeral]_particle	verb_auxiliary verb
a cake	three pieces	devided
divided a cake into three pieces.		

EXAMPLE2

休憩は	5 分間です。
kyuukei_wa	gofunkan_desu
noun_particle	noun[numeral]_auxiliary verb[copula, polite]
a break	is for 5minitues
a break is for 5minutes.	

EXAMPLE3 第一位(dai_ichi_i, No.1), 3 番目(san_ban_me, third)

NOTE Some numeral nouns are sometimes used as an adverb without a particle.

EXAMPLE

鉛筆を	4 本	準備しなさい。
enpitsu_wo	yon_hon	junbishinasai
noun_particle	noun[measure]_	verb
a pencil	4	Prepare
Prepare 4 pencils.		

© ISO 2010 – All rights reserved

7.2.3 Verb

Verbs are WSU, being subcategorized into main verbs, compound verbs, "*suru*" (do) verbs and subsidiary verbs.

NOTE1 A Japanese verb has an inflectional ending. The ending of a verb changes depending on whether it is a negation form, an adverbial form, a base form, an adnominal form, an assumption form, or an imperative form. Japanese verbs are often used with auxiliary verbs and/or particles, and they are considered as a word segmentation unit.

NOTE2 Endings are WSUA Japanese verb and adjective have an inflectional ending indicating conjugation form. There are six ending forms in Japanese; a negation form, an adverbial form, a base form, an adnominal form, an assumption form, and an imperative form. An ending is not treated as a WSU.

EXAMPLE 入らない(hairanai, not enter), 入_ら(hai_ra, "enter" and an ending in negation form), ない (nai, "not" auxiliary verb)

7.2.3.1 Single verb and compound verb

EXAMPLE

私は	毎朝	牛乳を	飲む。
watashi_wa	maiasa	gyuunyu_wo	nomu
noun_particle	adverb	noun_particle	verb
i	every morning	milk	drink

I drink milk every morning.

7.2.3.2 Verb composed from a noun and "*suru*"(do)

NOTE An action noun becomes a verb by adding a verb "suru (do)" to the end of an action noun, and is sometimes called "*Sahendoushi*." "*Sahendoushi"* is considered as one segmentation unit.

EXAMPLE

私は	英語を	勉強する。
watashi_wa	eigo_wo	benkyou+suru
noun_particle	noun_particle	verb [noun+do]
i	english	do study

I study English.

7.2.3.3 Verb with a subsidiary verb

A verb with a subsidiary verb is treated as a single WSU.

NOTE A function of a subsidiary verb is to complement the meaning of a main verb, such as "話している (hanashi+te+iru; being speaking)". Subsidiary verbs are not suffixes. They form verbs by being agglutinated to main verbs.

EXAMPLE

彼は	マンガを	読み過ぎる。
kare_wa	manga_wo	yomisugiru
noun_particle	noun_particle	verb[verb + subsidiary]
he	comics	read (them) too much

He reads comics too much.

7.2.3.4 Verb with an auxiliary verb and a particle

A verb with an auxiliary verb or a particle or with both of them is treated as a single WSU.

© ISO 2010 – All rights reserved

EXAMPLE1 A verb with an auxiliary verb

彼は	試験に	合格するだろう。
kare_wa	shiken_ni	goukakusuru_darou
noun_particle	noun_particle	verb_auxiliary verb[expectation]
he	the examination	will pass

He will pass the examination.

EXAMPLE2 A verb with an auxiliary verb and a particle

彼は	試験に	合格するだろうね。
kare_wa	shiken_ni	goukakusuru_darou_ne
noun_particle	noun_particle	verb_auxiliary verb_particle[mood]
he	the examination	will pass, don't you think so?

He will pass the examination. don't you think so?

7.2.4 Adjective

Adjectives are treated as WSU, being subcategorized into simple adjectives, derivative adjectives and compound adjectives.

NOTE Japanese adjectives have an inflectional ending that defines two categories of adjectives respectively known as "I"-type adjectives and "na"-type adjectives. Generally "I"-type adjectives and "na"-type adjectives are considered as one WSU. Though despite the Japanese School grammar that describes "na"-type adjectives like "Noun+ auxiliary verb(da)" that corresponds to two WSU, "na"-type adjectives are counted as one WSU.

7.2.4.1 Simple adjective

EXAMPLE 黒い(kuroi, black), 静かな(shizukana, quiet)

7.2.4.2 Derivative adjective

EXAMPLE 薄暗い(usugurai, dusky)都会的な(tokaitekina, urbane) 国際的な(kokusaitekina, international)

7.2.4.3 Compound adjective

EXAMPLE 青白い(ao_jiroi, pale)

7.2.5 Adnoun

Adnouns are treated as WSU.

NOTE An adnoun does not have an inflectional ending, while it functions as a modifier.

EXAMPLE 1

あらゆる	国
arayuru	kuni
adnoun	noun
every	country

every country

EXAMPLE 2

この	国

© ISO 2010 – All rights reserved

kono	kuni
adnoun	noun
this	country
this country	

7.2.6 Adverb

Adverbs are treated as WSU.

NOTE An adverb has no inflectional ending. it modifies a verb, an adjective, and even a sentence.

EXAMPLE1

やっと	来た。
yatto	ki_ta
adverb	verb_auxiliary verb
at last	Came
At last (someone) came.	

EXAMPLE2

幸運にも	雨が	降る。
kouunnimo	ame_ga	Furu
adverb	noun-particle	verb
fortunately	rain	will come
Fortunately it will rain.		

7.2.7 Conjunction

Conjunctions are treated as WSU.

EXAMPLE

そして、	彼は	笑った。
soshite	kare_wa	warat_ta
conjunction	noun_particle	verb_auxiliary verb
then	he	laughed
Then he laughed.		

7.2.8 Exclamation

Exclamations are treated as WSU.

EXAMPLE

あっ!
A!
Exclamation
Oops!
Oops!

© ISO 2010 – All rights reserved

7.2.9 Particle

Particles are treated as WSU.

NOTE In Japanese, there are seven subcategories, as illustrated below:

1) ”格助詞; kakujoshi” is a maker for a case. (が; ga; subject marker, を; wo; objective marker, に:ni; dative marker, and so on)

2) “係助詞; kakarijoshi” is a maker for a correlation with another phrase. (さえ; sae; even, しか; shika; only and so on)

3) “並立助詞; heiritsujoshi” is a marker for a coordination. (と; to; and, か; ka; or, and so on)

4) “接続助詞;setsuzokujoshi” is a marker for a conjunction between phrases. (ので; node; because, とき; toki; when, and so on)

5) “副助詞;fukujoshi” is a marker for a an attachment of some meanings. (くらい; kurai; about, まで; made; too)

6) “終助詞; shuujoshi” is a marker for representing a mood and a question of a speaker. It is always used at the end of a sentence. (ね; ne; don’t you think so?, か; ka; question)

7) “準体助詞;juntaijoshi” is a marker for a normalization of a phrase. (の; no; thing, こと; koto; thing)

EXAMPLE 1 particles for a case marker

私は(watashi_wa; I), 私を(watashi_wo; me), 私の(watashi_no; my), 私へ(watashi_e; to me) , 私と(watashi_to; with),

私に(watashi_ni; me, for me)

EXAMPLE 2 A particle for a conjunction

行けば(ike_ba; if you go) , 行くので(iku_node; because (someone) goes)

EXAMPLE 3 a particle for adding something of a meaning

私さえ（watashi_sae; even I）, 私も(watashi_mo; I (go together), too)

EXAMPLE 4 particles for representing a mood and a question

行きますね？
iki_masu_ne?
verb_auxiliary verb_particle[mood]
go, don’t you?
(You go there), don’t you?

7.2.10 Auxiliary Verb

Auxiliary verbs are treated as WSU.

NOTE Auxiliary verbs represent various semantic functions such as a capability, a voice, a tense, an aspect and so on. An auxiliary verb appears at the end of a phrase, a clause and a sentence. An auxiliary verb is a part of speech but should not be segmented. An auxiliary verb is used with a noun, a verb and an adjective at the end of a phrase, a clause and a sentence.

EXAMPLE

© ISO 2010 – All rights reserved

雨が	降りそうなので、	家に	いるでしょう。
ame_ga	furi_souna_node	ie_ni	i_masu
noun_particle	Verb_auxirialy verb[guess]_particle[conjunction]	noun_particle	verb_auxiliary verb [prospect, polite]
It	because(it) seems to rain	at home	(I) will be

Because it seems to rain, I will be at home.

8 Specific rules for identifying WSU in Korean text

8.1 Eojeol

Each eojeol is a WSU.

EXAMPLE

나는	학교로	갑니다
naneun	*hakgyoro*	*gabnida*
pronoun+GA	noun+GA	verb+GA
I	school	go

I go to school.

NOTE1 This sentence consists of three eojeols; 나는, 학교로, and 갑니다.

NOTE2 Each eojeol can be further segmented to smaller WSU. For example, the first eojeol “나는” is segmented to two WSU, “나” and “는“, where “나” is a WSU as a noun and “는“ is a WSU as a grammatical affix, as is specified in 8.2 and 8.3.

NOTE3 White space helps segmenting text into eojeols.

8.2 Lexical items

8.2.1 General rule

A string of characters that can be categorized as belonging to a part of speech is a WSU.

EXAMPLE

사과-를	먹-었-다.
sagwa_reul	*meok_eoss_da.*
Noun+GA	verb+GA+GA
apple+ [object]	eat+[past]+[final GA]

Ate apple.

NOTE1 The eojeol 사과를 consists of a noun and a grammatical affix. By 8.2.2 the noun 사과 is a WSU.

NOTE2 The grammatical affix 를 is also treated as a WSU by 8.3.

NOTE3 The parts of speech in Korean consist of noun, verb, adjective, adverb, adnoun, numeral, pronoun, exclamation. Examples are given in 8.2.2 and in the following.

8.2.2 Noun

A noun is treated as a WSU, being subcategorized into a common noun, a proper noun, and a bound noun.

EXAMPLE 1 (common noun)

© ISO 2010 – All rights reserved

소녀가	사과를	먹었다.
sonyeo_ga	*sagwa _reul*	*meogeotta*
noun_GA[subjective]	noun_GA[object]	verb_GA[past]_final GA
Girl	Apple	Ate

A girl ate an apple.

EXAMPLE 2 (proper noun)

국제표준화기구
gukjepyojunhwagigu
proper noun
International Organization for Standardization

EXAMPLE 3 (bound noun)

좋은	것
joeun	*geot*
Adjective	bound noun
Good	thing

good thing

8.2.3 Pronoun

A pronoun is treated as a WSU, being subcategorized into a personal pronoun, a demonstrative pronoun, and an interrogative pronoun.

EXAMPLE 1 personal pronoun

나-는	자기-를	소개하-지	않-았-다.
na-neun	*jagi-reul*	*sogaeha-ji*	*anh-ass-da*
pronoun_GA	pronoun_GA	Verb_GA	auxiliary verb_GA_GA
I	Myself	Introduce	not [past]

I did not introduce myself.

EXAMPLE 2 demonstrative pronoun

저기
jeogi
pronoun
There

EXAMPLE 3 interrogative pronoun

무엇
mueot
pronoun
what

8.2.4 Numeral

A numeral is treated as a WSU, being subcategorized into a quantifier numeral and an ordinal numeral.

EXAMPLE 1 quantifier numeral

© ISO 2010 – All rights reserved

하나
hana
numeral
one

EXAMPLE 2 ordinal numeral

둘째
duljjae
numeral
second

8.2.5 Verb

A verb is treated as a WSU, being subcategorized into a main verb and an auxiliary verb.

EXAMPLE 1

보-았-군-요
boatggunyo
verb+GA+GA+GA
see [past] [final] [polite]
You might saw (something).

EXAMPLE 2

먹-어	보-다
meogeo	*boda*
main verb+GA	auxiliary verb+GA
eat [conjunctive]	try
try to eat	

8.2.6 Adjective

An adjective is treated as a WSU, being subcategorized into a main adjective and an auxiliary adjective.

NOTE Korean adjectives behave like verbs, thus being agglutinated with grammatical affixes referring to tense, mood, etc.

EXAMPLE 1

검-군-요
geomgunyo
adjective+GA+GA
black [final] [polite]
It is black, isn't it?

EXAMPLE 2

새-하얗다
saehayata
Prefix_adjective+GA
very_white [final]
snowy

© ISO 2010 – All rights reserved

EXAMPLE 3

마시-고	싶다
masigo	*siptta*
Verb+GA	auxiliary adjective
Drink [conjunctive]	want
want to drink	

8.2.7 Adnoun

An adnoun is treated as a WSU.

NOTE Korean adnouns are like adjectives or determiners in western languages.

EXAMPLE

새	책
sae	*chaek*
Adnoun	noun
New	book
a new book	

8.2.8 Adverb

An adverb is treated as a WSU, being subcategorized into a degree adverb, a sentential adverb, and a conjunctive adverb.

EXAMPLE 1

매우	바쁘다
maeu	*babbeuda*
Adverb	verb
Very	busy
very busy	

EXAMPLE 2

다행히	비-가	온다.
dahaenghi	*biga*	*onda*
Adverb	noun-GA	verb+GA
Fortunately	rain	come [present]
Fortunately it rains.		

EXAMPLE 3

경제	및	문화
gyeongje	*mit*	*munhwa*
Noun	adverb	noun
Economy	and	culture
economy and culture		

8.2.9 Exclamation

An exclamation is treated as a WSU.

EXAMPLE

© ISO 2010 – All rights reserved

아!
A!
exclamation
Oops!

8.2.10 Grammatical affix

A grammatical affix is treated as a WSU, being subcategorized into a nominal grammatical affix, a verbal grammatical affix, an auxiliary grammatical affix, and a converting grammatical affix.

EXAMPLE 1

내	가
nae	*ga*
pronoun	grammatical affix
I	[subject]
I	

EXAMPLE 2

가	시	겠	습니다
ga	*si*	*get*	*seumnida*
Verb	GA	GA	final GA
Go	[polite]	[conjectural]	
Might have gone			

EXAMPLE 3

나-는	소설-만	읽-는다.
na-neun	*soseol-man*	*ik-neunda*
pronoun-GA	noun-auxiliary GA	Verb-GA
as for me	only novel	Read

As for me, I read only novels.

EXAMPLE 4

건강하기를	바란다
geongangha_gi_reul	*bara_nda*
adjective+GA+GA	verb+GA
Health	wish
wish health	

© ISO 2010 – All rights reserved

Annex A
(informative)

A comparative table for parts of speech in Chinese, Japanese, and Korean

POS	Chinese	Japanese	Korean
Noun	○(名词)	○ (名詞)	○(명사 名詞)
Verb	○（动词）	○（動詞）	○（동사 動詞）
Adjective	○（形容词）	○（形容詞 and 形容動詞）	○（형용사 形容詞）
Numeral	○（数词）	Subcategory of Noun （名詞[数詞]）	○(수사 數詞)
Adverb	○（副词）	○（副詞）	○（부사 副詞）
Exclamation	○（叹词）	○（感動詞）	○（감탄사 感歎詞）/（감동사 感動詞）
Pronoun	○（代词）	Subcategory of Noun （名詞[代名詞]）	○（대명사 代名詞）
Auxiliary word	○（助词）	×	×
Measure word	○（量词）	Noun or Adverb（名詞/副詞 [序数詞]）	（명사 名詞/부사 副詞 [序數詞]）
Modal word	○（语气词）	×	×
Imitative word	○（拟声词）	Part of Adverb （擬態語・擬音語）	（擬態語・擬音語）/（擬態語・擬音語）
Preposition	○（介词）	×	×
Conjunction	○（连词）	○（接続詞）	○(접속부사 接続副詞)
Particle	×	○（助詞）	○（조사 助詞）/ 토
Adnoun	×	○（連体詞）	○(관형사 冠形詞)
Auxiliary verb	Subcategory of Verb（能愿动词）	○ (助動詞）	Subcategory of Verb (보조동사 補助動詞) Subcategory of Adjective (보조형용사 補助 形容詞)
Differentiating word	○（区别词）	×	×

© ISO 2010 – All rights reserved