Word Segmentation Standard in Chinese, Japanese and Korean

Key-Sun Choi
KAIST
Daejeon Korea
kschoi@kaist.ac.kr

Hitoshi Isahara
NICT
Kyoto Japan
isahara@nict.go.jp

Kyoko Kanzaki
NICT
Kyoto Japan
kanzaki@nict.go.jp

Hansaem Kim
National Inst. Korean Lang.
Seoul Korea
thesis00@korea.kr

Seok Mun Pak
Baekseok Univ.
Cheonan Korea
smpark@bu.ac.kr

Maosong Sun
Tsinghua Univ.
Beijing China
sms@tsinghua.edu.cn

Abstract

Word segmentation is a process to divide a sentence into meaningful units called "word unit" [ISO/DIS 24614-1]. What is a word unit is judged by principles for its internal integrity and external use constraints. A word unit's internal structure is bound by principles of lexical integrity, unpredictability and so on in order to represent one syntactically meaningful unit. Principles for external use include language economy and frequency such that word units could be registered in a lexicon or any other storage for practical reduction of processing complexity for the further syntactic processing after word segmentation. Such principles for word segmentation are applied for Chinese, Japanese and Korean, and impacts of the standard are discussed.

1 Introduction

Word segmentation is the process of dividing of sentence into meaningful units. For example, "the White House" consists of three words but designates one concept for the President's residence in USA. "Pork" in English is translated into two words "pig meat" in Chinese, Korean and Japanese. In Japanese and Korean, because an auxiliary verb must be followed by main verb, they will compose one word unit like "tabetai" and "meoggo sipda" (want to eat). So the word unit is defined by a meaningful unit that could be a candidate of lexicon or of any other type of storage (or expanded derived lexicon) that is useful for the further syntactic processing. A word unit is more or less fixed and there is no syntactic interference in the inside of the word unit. In the practical sense, it is useful for the further syntactic parsing because it is not decomposable by syntactic processing and also frequently occurred in corpora.

There are a series of linguistic annotation standards in ISO: MAF (morpho-syntactic annotation framework), SynAF (syntactic annotation framework), and others in ISO/TC37/SC4 [1]. These standards describe annotation methods but not for the meaningful units of word segmentation. In this aspect, MAF and SynAF are to annotate each linguistic layer horizontally in a standardized way for the further interoperability. Word segmentation standard would like to recommend what word units should be candidates to be registered in some storage or lexicon, and what type of word sequences called "word unit" should be recognized before syntactic processing.

In section 2, principles of word segmentation will be introduced based on ISO/CD 24614-1. Section 3 will describe the problems in word segmentation and what should be word units in each language of Chinese, Japanese and Korean. The conclusion will include what could be shared among three languages for word segmentation.

2 Word Segmentation: Framework and Principles

Word unit is a layered pre-syntactical unit. That means that a word unit consists of the smaller word units. But the maximal word unit is frequently occurred in corpora under the constraints that the syntactic processing will not refer the internal structure of the word unit

Basic atoms of word unit are word form, morpheme including bound morpheme, and non-lexical items like punctuation mark, numeric string, foreign character string and others as shown in Figure 1. Usually we say that "word" is lemma or word form. Word form is a form that a lexeme takes when used in a sentence. For example, strings "have", "has", and "having" are word forms of the lexeme HAVE, generally distinguished by the use of capital letters. [ISO/CD 24614-1] Lemma is a conventional form used to represent a lexeme, and lexeme is an abstract unit generally associated with a set of word forms sharing a common meaning.

[1] Refer to http://www.tc37sc4.org/ for documents MAF, SynAF and so on.

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 179–186,
Suntec, Singapore, 6-7 August 2009.

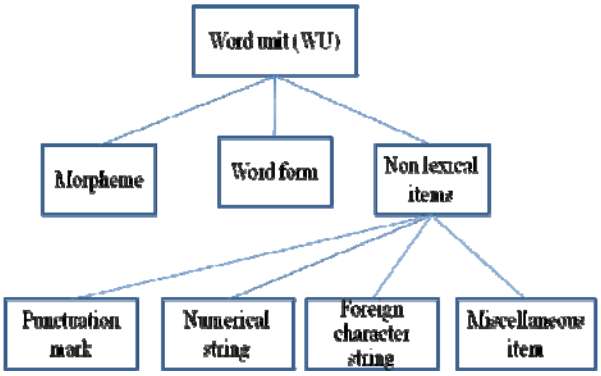

Figure 1. Configuration of Word Unit

BNF of word unit is as follows:

<*word unit*> ::= <*word form*> | <*morpheme*> | <*non lexical items*> | <*word unit*>,

where <*word unit*> is recursively defined because a longer word unit contains smaller word units.

Bunsetsu in Japanese is the longest word unit, which is an example of layered maximized pre-syntactical unit. *Eojeol* in Korean is a spacing unit that consists of one content word (noun, verb, adjective or adverb) plus a sequence of functional elements. Such language-specific word units will be described in section 3.

Principles for word segmentation will set the criteria to validate each word unit, to recognize its internal structure and non-lexical word unit, to be a candidate of lexicon, and other consistency to be necessitated by practical applications for any text in any language. The meta model of word segmentation will be explained in the processing point of view, and then their principles of word units in the following subsections.

2.1 Metamodel of Word Segmentation

A word unit has a practical unit that will be later used for syntactic processing. While the word segmentation is a process to identify the longest word unit and its internal structure such that the word unit is not the object to interfere any syntactic operation, "chunking" is to identify the constituents but does not specify their internal structure. Syntactic constituent has a syntactic role, but the word unit is a subset of syntactic constituent. For example, "blue sky" could be a syntactic constituent but not a word unit. Figure 2 shows the meta model of word segmentation. [ISO CD 24614-1]

2.2 Principles of Word Unit Validation

Principles for validating a word unit can be explained by two perspectives: one is linguistic one and the other is processing-oriented practical perspective.

In ISO 24614-1, principles from a linguistic perspective, there are five principles: principles of (1) bound morpheme, (2) lexical integrity, (3) unpredictability, (4) idiomatization, and (5) unproductivity.

First, bound morpheme is something like "in" of "inefficient". The principle of bound morpheme says that each bound morpheme plus word will make another word. Second, principle of lexical integrity means that any syntactic processing cannot refer the internal structure of word (or word unit). From the principle, we can say that the longest word unit is the maximal meaningful unit before syntactic processing. Third, another criterion to recognize a word is the principle of unpredictability. If we cannot infer the real fact from the word, we consider it as a word unit. For example, we cannot image what is the colour of blackboard because some is green, not black. [ISO 24614-1] The fourth principle is that the idiom should be one word, which could be a subsidiary principle that follows the principle of unpredictability. In the last principle, unproductivity is a stronger definition of word; there is no pattern to be copied to generate another word from this word formation. For example, in "白菜" (white vegetable) in Chinese, there is no other coloured vegetable like "blue vegetable."

Another set of principles is derived from the practical perspective. There are four principles: frequency, Gestalt, prototype and language economy. Two principles of frequency and language economy are to recognize the frequent occurrence in corpora. Gestalt and prototype principles are the terms in cognitive science about what are in our mental lexicon, and what are perceivable words.

Principle of language economy is to say about very practical processing efficiency: "if the inclusion of a word (unit) candidate into the lexicon can decrease the difficulty of linguistic analysis for it, then it is likely to be a word (unit)."

Gestalt principle is to perceive as a whole. "It supports some perceivable phrasal compounds into the lexicon even though they seem to be free combinations of their perceivable constituent parts," [ISO/CD 24614-1] where the phrasal compound is frequently used linguistic unit and its meaning is predictable from its constituent elements. Similarly, principle of prototype pro-

vides a rationale for including some phrasal compounds into lexicon, and phrasal compounds serve as prototypes in a productive word formation pattern, like “apple pie” with the pattern “fruit + pie” into the lexicon.

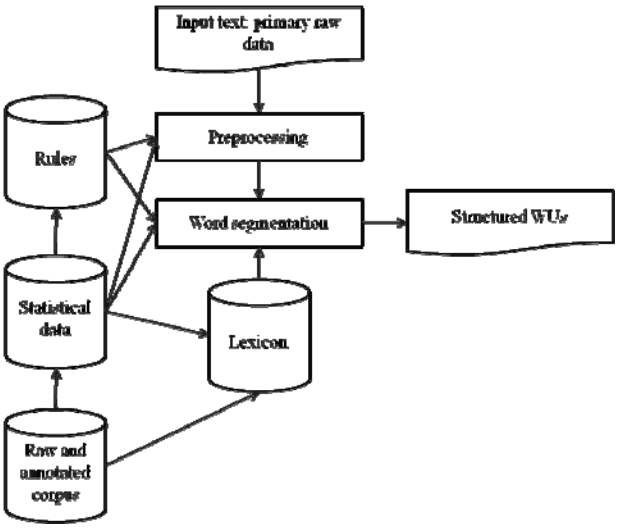

Figure 2. Meta model of word segmentation proess [ISO/CD 24614-1]

2.3 Principles for Word Unit Formation

As a result of word segmentation of sentence, we will get word units. These principles will describe the internal structure of word unit. They have four principles: granularity, scope maximization of affixations, scope maximization of compounding, and segmentation for other strings. In the principle of granularity, a word unit has its internal structure, if necessary for various application of word segmentation.

Principles of scope maximization of affixations and compounding are to recognize the longest word unit as one word unit even if it is composed of stem + affixes or compound of word units. For example, “unhappy” or “happy” is one word unit respectively. “Next generation Internet” is one word unit. The principle of segmentation for other strings is to recognize non-lexical strings as one word unit if it carries some syntactic function, for example, 2009 in “Now this year is 2009.”

3 Word Segmentation for Chinese, Japanese and Korean

If the word is derived from Chinese characters, three languages have common characteristics. If their word in noun consists of two or three Chinese characters, they will be one word unit if they are “tightly combined and steadily used.” Even if it is longer length, it will be a word unit if it is fixed expression or idiom. But if the first character is productive with the following numeral, unit word or measure word, it will be segmented. If the last character is productive in a limited manner, it forms a word unit with the preceding word, for example, “東京都” (Tokyo Metropolis), “8 月” (August) or “加速器” (accelerator). But if it is a productive suffix like plural suffix and noun for locality, it is segmented independently in Chinese word segmentation rule, for example, “朋友|们” (friends), “长江|以北” (north of the Yangtzi River) or “桌子|上” (on the table) in Chinese. They may have different phenomena in each language.

Negation character of verb and adjective is segmented independently in Chinese, but they form one word unit in Japanese. For example, “yasashikunai” (優しく無い, not kind) is one word unit in Japanese, but “不|写” (not to write), “不| 能” (cannot), “没|研究” (did not research) and “未| 完成” (not completed) will be segmented independently in Chinese. In Korean, “chinjeolhaji anhta” (친절하지 않다, not kind) has one space inserted between two eojeols but it could be one word unit. “ji anhta” makes negation of adjectival stem “chinjeolha”.

We will carefully investigate what principles of word units will be applied and what kind of conflicts exists. Because the motivation of word segmentation standard is to recommend what word 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 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 unit.

3.1 Word Segmentation for Chinese

For convenience of description, the specification in this section follows the convention that classifies words into 13 types: noun, verb, adjective, pronoun, numeral, measure word, adverb, preposition, conjunction, auxiliary word, modal word, exclamation word, imitative word.

3.1.1 Noun

There is word unit specification for common nouns as follows:

- Two-character word or compact two-character noun phrase, e.g., 牛肉(beef) 钢铁(steel)

- Noun phrase with compact structure, if violate original meaning after segmentation, e.g., 有功功率 (Active power)
- Phrase consisting of adjective with noun, e.g., 绿叶 (green leave)
- The meaning changed phrase consisting of adjective, e.g., 小媳妇(young wife)
- Prefix with noun, e.g., 阿哥(elder brother) 老鹰 (old eagle) 非金属 (nonmetal) 超声波 (ultrasonic)
- Noun segmentation unit with following postfix (e.g. 家手性员子化长头者), e.g., 科学家 (scientist)
- Noun segmentation unit with prefix and postfix, e.g., 超导性(superconductance)
- Time noun or phrase, e.g., 五月(May), 11 时 42 分 8 秒(forty two minutes and eight seconds past eleven), 前天(the day before yesterday), 初一(First day of a month in the Chinese lunar calendar)

But the followings are independent word units for noun of locality (e.g., 桌子|上 (on the table), 长江|以北 (north of the Yangtzi River)), and the "们" suffix referring to from a plural of front noun (e.g., 朋友 们(friends)) except "人们", "哥儿们"(pals), "爷儿们"(guys), etc. Proper nouns have similar specification.

3.1.2 Verb

The following verb forms will be one word unit as:

- Various forms of reiterative verb, e.g., 看看 (look at), 来来往往(come and go)
- Verb–object structural word, or compact and stably used verb phrase, e.g., 开会(meeting) 跳舞(dancing)
- Verb–complement structural word (two-character), or stably used Verb-complement phrase (two-character), e.g., 提高(improve)
- Adjective with noun word, and compact, and stably used adjective with noun phrase, e.g., 胡闹(make trouble) , 瞎说(talk nonsense)
- Compound directional verb, e.g., 出去(go out) 进来(come in).

But the followings are independent word units:

- "AAB, ABAB" or "A 一 A, A 了 A, A 了一 A", e.g., 研究|研究(have a discuss), 谈|一|谈 (have a good chat)
- Verb delimited by a negative meaning characters, e.g., 不|写(not to write) 不|能(cannot) 没|研究(did not research) 未|完成(not completed)
- "Verb + a negative meaning Chinese character + the same verb" structure, e.g., "说|没|说(say or not say)?"
- Incompact or verb–object structural phrase with many similar structures, e.g., 吃|鱼(Eat fish) 学|滑冰(learn skiing)
- "2with1" or "1with2" structural verb- complement phrase, e.g., 整理|好(clean up), 说|清楚 (speak clearly), 解释|清楚(explain clearly)
- Verb-complement word for phrase, if inserted with "得 or 不", e.g., 打|得|倒 (able to knock down), and compound directional verb of direction, e.g., 出|得|去(able to go out)
- Phrase formed by verb with directional verb, e.g., 寄|来(send), 跑|出|去(run out)
- Combination of independent single verbs without conjunction, e.g., 苫|盖(cover with), 听|说, 读|写 (listen, speaking, read and write)
- Multi-word verb without conjunction, e.g., 调查|研究(investigate and research)

3.1.3 Adjective

One word unit shall be recognized in the following cases:

- Adjective in reiterative form of "AA, AABB, ABB, AAB, A+"里"+AB", e.g., 大大(big), 马里马虎(careless), except the adjectives in reiterative form of "ABAB", e.g., 雪白|雪白 (snowy white)
- Adjective phrase in from of "一 A 一 B，一 A 二 B，半 A 半 B，半 A 不 B，有 A 有 B", e.g., 一心一意(wholeheartedly)
- Two single-character adjectives with word features varied, 长短 (long-short) 深浅 (deep-shallow) 大小(big-small)
- Color adjective word or phrase, e.g., 浅黄(light yellow) 橄榄绿(olive green)

But the followings are segmented as independent word units:

- Adjectives in parataxis form and maintaining original adjective meaning, e.g., 大|小 尺寸 (size), 光荣 |伟大(glory)
- Adjective phrase in positive with negative form to indicate question, e.g., 容易| 不| 容易(easy or not easy), except the brachylogical one like 容不容易(easy or not).

3.1.4 Pronoun

The followings shall be one word unit:

- Single-character pronoun with “们”, e.g.,我们 (we)
- “这、那、哪”with unit word “个” or “些、样、么、里、边”, e.g., 这个(this)
- Interrogative adjective or phrase, e.g., 多少 (how many)

But, the following will be independent word units:

- “这、那、哪” with numeral , unit word or noun word segmentation unit, e.g., 这 |十 天 (these 10 days), 那| 人(that person)
- Pronoun of “各、每、某、本、该、此、全”, etc. shall be segmented from followed measure word or noun, e.g., 各| 国 (each country), 每| 种(each type).

3.1.5 Numeral

The followings will be one word unit:

- Chinese digit word, e.g., 一亿八千零四万七百二十三(180,040,723)
- “分之” percent in fractional number, e.g., 五分之三(third fifth)
- Paratactic numerals indicating approximate number, e.g., 八九 公斤(eight or nine kg)

On the other hand, Independent word unit cases are as follows:

- Numeral shall be segmented from measure word, e.g., 三| 个(three)
- Ordinal number of “第” shall be segmented from followed numeral, e.g., 第 一 (first)
- “多、一些、点儿、一点儿”, used after adjective or verb for indicating approximate number.

3.1.6 Measure word

Reiterative measure word and compound measure word or phrase is a word unit, e.g., 年年 (every year), 人年 man/year.

3.1.7 Adverb

Adverb is one word unit. But “越…越…、又…又…”, etc. acting as conjunction shall be segmented, e.g., 又 香 又 甜(sweet yet savory).

3.1.8 Preposition

It is one word unit. For example, 生于(be born in), and 按照规定(according to the regulations).

3.1.9 Conjunction

Conjunction shall be deemed as segmentation unit.

3.1.10 Auxiliary word

Structural auxiliary word “的、地、得、之” and tense auxiliary word “着、了、过” are one word unit, e.g., 他|的|书 (his book), 看|了 (watched). But the auxiliary word “所” shall be segmented from its followed verb, e.g., 所 想 (what one thinks).

3.1.11 Modal word

It is one word unit, e.g., 你好吗？(How are you?).

3.1.12 Exclamation word

Exclamation word shall be deemed as segmentation unit. For example, “啊，真美！” (How beautiful it is!)

3.1.13 Imitative word

It is one word unit like “当当” (tinkle).

3.2 Word Segmentation for Japanese

For convenience of description, the specification in this section follows the convention that classifies words into mainly 10 types: meishi (noun), doushi (verb), keiyoushi (adjective), rentaishi (adnominal noun: only used in adnominal usage), fukushi (adverb), kandoushi (exclamation), setsuzoushi (conjunction), setsuji (affix), joshi (particle), and jodoushi (auxiliary verb). These parts of speech are divided into more detailed classes in terms of grammatical function.

The longest "word segmentation" corresponds to “Bunsetsu” in Japanese.

3.2.1 Noun

When a noun is a member constituting a sentence, it is usually followed by a particle or auxiliary verb (e.g. “猫が” (neko_ga) which is composed from “Noun + a particle for subject marker”). Also, if a word like an adjective or adnominal noun modifies a noun, then a modifier (adjectives, adnominal noun, adnominal phrases) and a modificand (a noun) are not segmented.

3.2.2 Verb

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 a verb with auxiliary verbs and/or particles is considered as a word segmentation unit (e.g. “歩きました” (aruki_mashi_ta) is composed from “Verb + auxiliary verb for politeness + auxiliary verb for tense”).

3.2.3 Adjective

A Japanese adjective has an inflectional ending. Based on the type of inflectional ending, there

are two kinds of adjectives, "i_keiyoushi" and "na_keiyoushi". However, both are treated as adjectives.

In terms of traditional Japanese linguistics, “keiyoushi” refers to “i_keiyoushi”(e.g. 美しい, utsukushi_i) and “keiyoudoushi”(e.g. 静かな, shizuka_na) refers to “na_keiyoushi.” In terms of inflectional ending of “na_keiyoushi,” it is sometimes said to be considered as “Noun + auxiliary verb (da)”.

The ending of an adjective changes depending on whether it is a negation form, an adverbial form, a base form, an adnominal form, or an assumption form. Japanese adjectives in predicative position are sometimes used with auxiliary verbs and/or particles, and they are considered as a word segmentation unit.

3.2.4 Adnominal noun

An adnominal noun does not have an inflectional ending; it is always used as a modifier. An adnominal noun is considered as one segmentation unit.

3.2.5 Adverb

An adverb does not have an inflectional ending; it is always used as a modifier of a sentence or a verb. It is considered as one segmentation unit.

3.2.6 Conjunction

A conjunction is considered as one segmentation unit.

3.2.7 Exclamation

An exclamation is considered as one segmentation unit.

3.2.8 Affix

A prefix and a suffix used as a constituent of a word should not be segmented as a word unit.

3.2.9 Particle

Particles can be divided into two main types. One is a case particle which serves as a case marker. The other is an auxiliary particle which appears at the end of a phrase or a sentence.

Auxiliary particles represent a mood and a tense.

Particles should not be segmented from a word. A particle is always used with a word like a noun, a verb, or an adjective, and they are considered as one segmentation unit.

3.2.10 Auxiliary verb

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 or a sentence. Some linguist consider “だ” (da), which is Japanese copura, as a specific category such as 判定詞(hanteishi).

An auxiliary verb should not be segmented from a word. An auxiliary verb is always used with a word like a noun, a verb, or an adjective, and is considered as one segmentation unit.

3.2.11 Idiom and proverb

Proverbs, mottos, etc. should be segmented if their original meanings are not violated after segmentation. For example:

Kouin	yano	gotoshi
noun	noun+particle	auxiliary verb
time	arrow	like (flying)

Time flies fast.

3.2.12 Abbreviation

An abbreviation should not be segmented.

3.3 Word Segmentation for Korean

For convenience of description, the specification in this section follows the convention that classifies words into 12 types: noun, verb, adjective, pronoun, numeral, adnominal, adverb, exclamation, particle, auxiliary verb, auxiliary adjective, and copula. The basic parts of speech can be divided into more detailed classes in terms of grammatical function. Classification in this paper is in accord with the top level of MAF.

In addition, we treat some multi-Eojeol units as the word unit for practical purpose. Korean *Eojeol* is a spacing unit that consists of one content word (like noun, verb) and series of functional elements (particles, word endings). Functional elements are not indispensable. Eojeol is similar with *Bunsetsu* from some points, but an Eojeol is recognized by white space in order to enhance the readability that enables to use only Hangul alphabet in the usual writing.

3.3.1 Noun

When a noun is a member constituting a sentence, it is usually followed by a particle. (e.g. “사자_가” (*saja_ga*, ‘a lion is’) which is composed from “Noun + a particle for subject marker”). Noun subsumes common noun, proper noun, and bound noun.

If there are two frequently concatenated Eojeols that consist of modifier (an adjective or an adnominal) and modificand (a noun), they can be one word unit according to the principle of language economy. Other cases of noun word unit are as follows:

1) Compound noun that consists of two more nouns can be a word unit. For example,

"손목" (*son_mok*, 'wrist') where *son+mok* = 'hand'+'neck'.

2) Noun phrase that denotes just one concept can be a word unit. For example, "예술의 전당" (*yesul_ui jeondang*, 'sanctuary of the arts' that is used for the name of concert hall).

3.3.2 Verb

A Korean verb has over one inflectional endings. The endings of a verb can be changed and attached depending on grammatical function of verb (e.g. "깨/뜨리/시/었/겠/군/요" (break [+emphasis] [+polite] [+past] [+conjectural] final ending [+polite]). Compound verb (verb+verb, noun+verb, adverb+verb) can be a segmentation unit by right. For example, "돌아가다" (*dola-ga-da*, 'pass away') is literally translated into 'go+back' (verb+verb). "빛나다" (*bin-na-da*, 'be shiny') is derived from 'light + come out' (noun+verb). "바로잡다" (*baro-jap-da*, 'correct') is one word unit but it consists of 'rightly+hold' (adverb+verb).

3.3.3 Adjective

A Korean adjective has inflectional endings like verb. For example, in "예쁘/시/었/겠/군/요" (pretty [+polite] [+past] [+conjectural] final ending [+polite]), one adjective has five endings. Compound adjective can be a segmentation unit by right. (e.g. "검붉다" (*geom-buk-da*, 'be blackish red'))

3.3.4 Adnominal

An adnominal is always used as a modifier for noun. Korean adnominal is not treated as noun unlike Japanese one. (e.g. "새 집" (*sae jip*, 'new house')" which consist of "adnominal + noun").

3.3.5 Adverb

An adverb does not have an inflectional ending; it is always used as a modifier of a sentence or a verb. In Korean, adverb includes conjunction. It is considered as one segmentation unit. Compound adverb can be a segmentation unit by right. Examples are "밤낮" (*bam-nat*, 'day and night'), and "곳곳" (*gotgot*, 'everywhere' whose literal translation is 'where where').

3.3.6 Pronoun

A pronoun is considered as one segmentation unit. Typical examples are "나" (*na*, 'I'), "너" (*neo*, 'you'), and "우리" (*uri*, 'we). Suffix of plural "들" (*deul*, '-s') can be attached to some of pronouns in Korean. (e.g. "너희들" (*neohui-deul*, 'you+PLURAL'), "그들" (*geu-deul*, 'they' = 'he/she+PLURAL')).

3.3.7 Numeral

A numeral is considered as one segmentation unit: e.g. "하나" (hana, 'one'), "첫째" (*cheojjae*, 'first'). Also figures are treated as one unit like "2009 년" (the year 2009).

3.3.8 Exclamation

An exclamation is considered as one segmentation unit.

3.3.9 Particle

Korean particles can not be segmented from a word just like Japanese particles. A particle is always used with a word like a noun, a verb, or an adjective, but it is considered as one segmentation unit.

Particles can be divided into two main types. One is a case particle that serves as a case marker. The other is an auxiliary particle that appears at the end of a phrase or a sentence. Auxiliary particle represents a mood and a tense.

3.3.10 Auxiliary verb

A Korean auxiliary verb represents various semantic functions such as a capability, a voice, a tense, an aspect and so on.

Auxiliary verb is only used with a verb plus endings with special word ending depending on the auxiliary verb. For example, "보다" (*boda*, 'try to'), an auxiliary verb has the same inflectional endings but it should follow a main verb with a connective ending "어" (*eo*) or "고" ('*go*'). Consider "try to eat" in English where "eat" is a main verb, and "try" is an auxiliary verb with specialized connective "to". In this case, we need two Korean Eojeols that corresponds to "eat + to" and "try". Because "to" is a functional element that is attached after main verb "eat", it constitutes one Eojeol. It causes a conflict between Eojeol and word unit. That means every Eojeol cannot be a word unit. What are the word units and Eojeols in this case? There are two choices: (1) "eat+to" and "try", (2) "eat"+ "to try". According to the definition of Eojeol, (1) is correct for two concatenated Eojeols. But if the syntactic processing is preferable, (2) is more likely to be a candidate of word units.

3.3.11 Auxiliary adjective

Unlike Japanese, there is auxiliary adjective in Korean. Function and usage of it are very similar to auxiliary verb. Auxiliary adjective is considered as one segmentation unit.

Auxiliary verb can be used with a verb or adjective plus endings with special word ending depending on the auxiliary adjective. For example, in "먹고 싶다" (*meokgo sipda*, 'want to eat'), *sipda* is an auxiliary adjective whose meaning is 'want' while *meok* is a main verb 'want' and *go* corresponds to 'to'; so *meokgo* is 'to eat'.

3.3.12 Copula

A copula is always used for changing the function of noun. After attaching the copula, noun can be used like verb. It can be segmented for processing.

3.3.13 Idiom and proverb

Proverbs, mottos, etc. should be segmented if their original meanings are not violated after segmentation like Chinese and Japanese.

3.3.14 Ending

Ending is attached to the root of verb and adjective. It means honorific, tense, aspect, modal, etc.

There are two endings: prefinal ending and final ending. They are functional elements to represent honorific, past, or future functions in prefinal position, and declarative (*-da*) or concessive (*-na*)-functions in final ending. Ending is not a segmentation unit in Korean. It is just a unit for inflection.

3.3.15 Affix

A prefix and a suffix used as a constituent of a word should not be segmented as a word unit.

4 Conclusion

Word segmentation standard is to recommend what type of word sequences should be identified as one word unit in order to process the syntactic parsing. Principles of word segmentation want to provide the measure of such word units. But principles of frequency and language economy are based on a statistical measure, which will be decided by some practical purpose.

Word segmentation in each language is somewhat different according to already made word segmentation regulation, even violating one or more principles of word segmentation. In the future, we have to discuss the more synchronized word unit concept because we now live in a multi-lingual environment. For example, consider figure 3. Its English literal translation is "white vegetable and pig meat", where "white vegetable" (白菜) is an unproductive word pattern and forms one word unit without component word units, and "pig meat" in Chinese means one English word "pork". So "pig meat" (猪肉) is the longest word unit in this case. But in Japanese and Korean, "pig meat" in Chinese characters cannot be two word units, because each component character is not used independently.

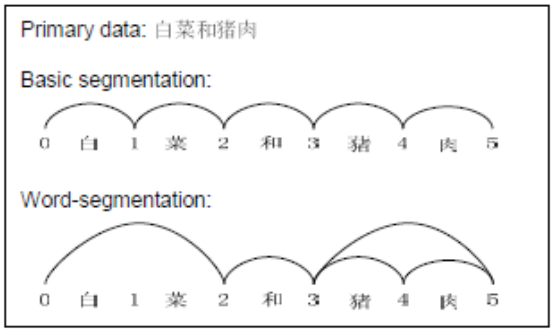

Figure 3. Basic segmentation and word segmentation [ISO/CD 24614-1]

What could be shared among three languages for word segmentation? The common things are not so much among CJK. The Chinese character derived nouns are sharable for its word unit structure, but not the whole. On the other hand, there are many common things between Korean and Japanese. Some Korean word endings and Japanese auxiliary verbs have the same functions. It will be an interesting study to compare for the processing purpose.

The future work will include the role of word unit in machine translation. If the corresponding word sequences have one word unit in one language, it is one symptom to recognize one word unit in other languages. It could be "principle of multi-lingual alignment."

The concept of "word unit" is to broaden the view about what could be registered in lexicon of natural language processing purpose, without much linguistic representation. In the result, we would like to promote such language resource sharing in public, not just dictionaries of words in usual manner but of word units.

Acknowledgement

This work has been supported by ISO/TC37, KATS and Ministry of Knowledge Economy (ROKorea), CNIS and SAC (China), JISC (Japan) and CSK (DPRK) with special contribution of Jeniffer DeCamp (ANSI) and Kiyong Lee.

References

ISO CD24614-1, Language Resource Management – Word segmentation of written texts for monolingual and multilingual information processing – Part 1: Basic concepts and general principles, 2009.

ISO WD24614-2, – Part 2: Word segmentation for Chinese, Japanese and Korean, 2009.