INTERNATIONAL STANDARD

ISO 24614-1

First edition
2010-11-01

Language resource management — Word segmentation of written texts —

Part 1: Basic concepts and general principles

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

Partie 1: Notions fondamentales et principes généraux

Reference number
ISO 24614-1:2010(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 24614-1 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*

Word segmentation for other languages is to form the subject of a future Part 3.

Introduction

Word segmentation is the dividing of text into linguistic units that carry meaning. For example, “the white house” can be divided into three meaningful units, “the,” “white,” and “house”, when it refers to a house that is white; whereas “the White House” corresponds to only one meaningful unit when it refers to the residence of the US President.

For the purposes of ISO 24614, such meaningful linguistic units are called *word segmentation units* (WSU). As demonstrated in the previous example, a WSU can be comprised of more than one word. A WSU can consist of a stem and affixes (e.g. “re+work+ing”). It can be a compound word (e.g. “blackboard”), a proper noun (e.g. “Cape Town”), an idiom (e.g. “It's raining cats and dogs”), or a multiword expression (e.g. “take care of”). For languages that have spaces between words, such as English, segmenting a text into WSU is facilitated by using the spaces as a basis for establishing the boundaries of a WSU, although additional considerations need to be taken into account for handling abbreviations, punctuation and multiword units of meaning, among others. For languages that do not have spaces between words, such as Chinese and Japanese, or for languages that have spaces partially between words, such as Thai and Korean, segmenting a text into WSU requires a different approach.

Furthermore, word segmentation is complex for languages that are characterized by extensive compounding, such as Chinese, and for languages that are characterized by extensive agglutination, such as Japanese, Korean and Hungarian. On the other hand, the fact that Japanese supports multiple scripts is beneficial for word segmentation.

However, white space alone is not sufficient to segment a text. “Apple pie,” for example, is understood as a kind of pie made of apples, so “apple” and “pie” are treated as two distinct WSUs. Alternatively, it can be viewed as a single entity due to its collocational and idiomatic properties, and treated as a single WSU. Segmentation rules can differ between languages, even when applied to equivalent expressions (as discussed in ISO 24614-2).

Elaborating standards for the rules and methods for word segmentation can facilitate innovation and development in areas such as language learning and translation. It could improve language-related technologies, including spell checking, grammar checking, dictionary lookup, terminology management, translation memory, information retrieval, information extraction and machine translation. For instance, by failing to identify “kick the bucket” as a single WSU, translation memory and machine translation technologies would produce a literal rather than idiomatic translation.

This part of ISO 24614 is the first in a series of International Standards targeted at word segmentation in written languages. It focuses on the basic concepts and general principles of word segmentation that apply to languages in general. The subsequent parts will, however, focus on the issues specific to particular languages.

Language resource management — Word segmentation of written texts —

Part 1: Basic concepts and general principles

1 Scope

This part of ISO 24614 presents the basic concepts and general principles of word segmentation, and provides language-independent guidelines to enable written texts to be segmented, in a reliable and reproducible manner, into word segmentation units (WSU).

NOTE 1 In language-related research and industry, the word is a fundamental and necessary concept. It is thus critical to have a universal definition of what comprises a word for the purposes of segmenting a text into words. One cannot simply use rules based only on spaces and punctuation to delimit words. Such rules do not account for situations such as hyphenated compounds, abbreviations, idioms or word-like expressions that contain symbols or numbers. Word segmentation is even more problematic for languages that do not use spaces to separate words, such as Chinese and Japanese, and for agglutinative languages, where some functional word classes are realized as affixes, such as Korean.

The many applications and fields that need to segment texts into words — and thus to which this part of ISO 24614 can be applied — include the following.

Translation

Word count is the principal method for calculating the cost of a translation. Word segmentation is a standard function in translation memory systems and computer-assisted translation (CAT) tools. Word segmentation is performed by term extraction tools, which are sometimes provided in terminology management systems and CAT tools.

Content management

Most content management systems and databases allow for searching by individual words. The content being searched has to be segmented to permit matching with a search word. Furthermore, search functions require knowledge of the boundaries of words.

Speech technologies

Text-to-speech systems generate speech based on words and therefore require word segmentation for lexicon lookup, stress assignment, prosodic pattern assignment, etc.

Computational linguistics

Various natural language processing (NLP) systems must segment text into words in order to carry out their functions. NLP systems include

— morphosyntactic processors,

— syntactic parsers,

— spellcheckers,

— text classification systems, and

— corpus linguistics annotators.

Lexicography

Lexical resources are often evaluated by size, usually by referring to the number of words.

NOTE 2 The size of language resources is an essential benchmark for their management. Quantifying the size of language resources is typically achieved by counting the words. However, because NLP applications use different segmentation methods, each calculates the number of words differently and arrives at a different sum for the same text. A reliable, reproducible, standard measure would allow comparable results. This is not to say that applications may not use their own, application-specific segmentation methods. For example, a speech synthesis application might segment a text into smaller or larger units compared to another application.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

2.1
abbreviation
verbal designation formed by omitting words or letters from a longer form and designating the same concept

[ISO 1087-1:2000]

2.2
affix
bound morpheme (2.5) which may be added to a **stem** (2.22) or a **lexeme** (2.14)

NOTE Affixes can be classified into several sub-types such as prefix, suffix, infix and circumfix. Affixes can be derivational or they can be inflectional or agglutinative.

2.3
agglutination
process of concatenating one or more **affixes** (2.2) to a **stem** (2.22)

[ISO 24613:2008]

2.4
borrowing
process of word formation in which a linguistic expression is adopted from another language, usually when no term exists for the new object or concept

2.5
bound morpheme
morpheme (2.18) that appears only together with one or several other morphemes

[ISO 24613:2008]

EXAMPLE 1 Chinese: 伟 means “great,” but cannot stand by itself as a word in text. Instead, it is used as a constituent element of many words, such as 伟大 (“great”), 伟人 (“giant”), and 雄伟 (“majesty”).

EXAMPLE 2 Korean: the suffix “-e”, which is equivalent to the English preposition “to” — as in “hakkyo-e” (to school) — is a bound morpheme.

2.6
compound
word (2.23) built from two or more **lexemes** (2.14)

NOTE 1 Adapted from ISO 24613:2008, definition 3.10.

NOTE 2 A compound may be endocentric if it has a head (i.e. the fundamental part that contains the basic meaning of the whole compound) and modifiers (which restrict this meaning), or exocentric if it does not have a head. A compound can be long. There are two main sub-types of compound according to their degree of lexicalization: word compound and phrasal compound.

2.7
compounding
word formation in which a new word is formed by adjoining at least two **lexemes** (2.14), in their original forms or with slight transformations

[ISO 24613:2008]

2.8
derivation
change in the form of a **word** (2.23) to create a new **word** (2.23), usually by modifying the **stem** (2.22) or by affixation

[ISO 24613:2008]

2.9
free morpheme
morpheme (2.18) that can be used as a **word** (2.23) by itself

EXAMPLE Given the word “goodness,” “good” is a free morpheme, whereas “-ness” is not. The latter is a bound morpheme.

2.10
homograph
each of two or more **word forms** (2.24) or **words** (2.23) with identical spelling but representing different concepts (semantic homography) or syntactic functions (syntactic homography)

[ISO 1087-2:2000]

2.11
inflection
process in which a **word form** (2.24) is made up by adding an **affix** (2.2) to a **stem** (2.22)

NOTE Inflection is a grammatical rather than lexical process.

2.12
lemma
conventional form chosen to represent a **lexeme** (2.14)

[ISO 24613:2008]

EXAMPLE Given a set of word forms such as “find,” “finds,” “found,” and “finding” in English, the form “find” is chosen as a lemma to represent the group of all these word forms.

2.13
lemmatization
process of determining the **lemma** (2.12) for a given **word form** (2.24) in a context

EXAMPLE Given the word “found” in English, lemmatization results in “find” as its lemma.

NOTE Adapted from ISO 1087-2:2000, definition 2.19 and ISO 30042:2008, definition 3.14.

2.14
lexeme
abstract unit generally associated with a set of forms sharing a common meaning

[ISO 24613:2008]

NOTE 1 A lexeme may be a part of another lexeme, as a consequence of derivation and compounding.

NOTE 2 "Form" is defined in ISO 24613 as "sequence of morphs".

2.15
lexicalization
process of making a linguistic unit function as a word

NOTE Such a linguistic unit can be a single morph, e.g. "laugh," a sequence of morphs, e.g. "apple pie" or even a phrase, such as "kick the bucket", that forms an idiomatic phrase.

2.16
lexicon
list of entries mainly headed by **lemmas** (2.12) with associated information

2.17
morph
surface form represented by a unique **morpheme** (2.18)

EXAMPLE In English, the morphs of the plural morpheme "-s" include "-s", "-en", and "-NULL" (as in "boys", "oxen", and "sheep"), where "–NULL" has no unique surface form. Thus, the word "boys" consists of the two morphs, "boy" and "-s", whereas the morphemes corresponding to the morphs "ox" and "-en" are "ox" and "-s", respectively.

2.18
morpheme
smallest unit of meaning expressed by a sequence of phonemes or a sequence of graphemes

[ISO 24613:2008]

NOTE There are two sub-types of morphemes: free morphemes and bound morphemes.

2.19
multiword expression
MWE
lexeme (2.14) made up of a sequence of lexemes that has properties that are not predictable from the properties of the individual lexemes or their normal mode of combination

[ISO 24613:2008]

NOTE A multiword expression can be a compound [a word compound or phrasal compound, an idiom, a fragment of a sentence or a sentence (e.g. a proverb or familiar quotation)]. It is not always possible to specify the part of speech for the whole MWE span.

2.20
phrasal compound
word (2.23) consisting of two or more **lexemes** (2.14), the meaning of which is predictable from its constituent elements

EXAMPLE "Apple pie" in English is a phrasal compound composed of two lexemes, "apple" and "pie", whose meanings are preserved in the meaning of the compound.

NOTE 1 Idioms use two or more lexical items, but do not compose a phrasal compound.

NOTE 2 A phrasal compound might be thought of as phrases by some linguists. In practice, however, there is not always a clear distinction between a word compound and a phrasal compound, or between a phrasal compound and a phrase, due to the fuzziness of semantic predictability and the degree of lexicalization. Lexico-statistics — word frequency in particular — will play an important role in this respect.

2.21
reduplication
process in which the entire **word** (2.23), or part of it, is repeated

2.22
stem
linguistic unit whose form is smaller than or equal to the form of a single **lexeme** (2.14) and that may be affected by an inflectional, agglutinative, compositional or derivational process

[ISO 24613:2008]

2.23
word
lexeme (2.14) that has, as a minimal property, a part of speech

[ISO 24613:2008]

2.24
word form
morphosyntactical variant of a given **word** (2.23)

[ISO 1087-2:2000]

EXAMPLE In English, the strings “find”, “finds”, “found” and “finding” are word forms of the word “find”.

2.25
word segmentation
process of splitting text into a sequence of **word segmentation units** (2.26)

2.26
word segmentation unit
WSU
word form (2.24) or character string of some other type that is treated as a unit

NOTE A character string that is not a word form may consist of numeric characters, foreign characters, punctuation marks or some other miscellaneous characters such as Chinese radicals, chemical symbols, such as H_2O, or a mixture of Latin and numeric characters, such as F16.

2.27
word structure
internal structure of a **word** (2.23) resulting from the morphological analysis

NOTE In agglutinative languages, such as Korean, Japanese and Turkish, a word may consist of a sequence of morphemes, with a comparatively high morpheme-per-word ratio, where each affix involved (both derivational and inflectional) typically expresses a particular grammatical meaning in a clear, one-to-one way. The structure of a word in these languages can be very sophisticated, with free morphemes and separate affixes as its constituent elements.

2.28
word compound
compound (2.6) whose overall meaning is not totally predictable from its constituent parts

EXAMPLE “Hotdog,” “ice-cream,” “blackboard”.

3 Basic framework for word segmentation

3.1 Essential concepts related to word segmentation

The concepts described in this clause are critical to understanding the principles of word segmentation.

Figure 1 shows the relationship between the abstract entities of “morpheme” and “lexeme” and the concrete entities of “morph,” “word forms,” and “lexicon.” The concrete form of a morpheme is a morph. The concrete form of a lexeme is a word form. A lexicon is mainly composed of lemmas, which are derived from word forms via a process of lemmatization.

NOTE 1 Terms such as “morpheme” and “word” have different meanings in the fields of linguistics and terminology. These and other terms are used as described in Clause 2, according to their linguistic interpretation.

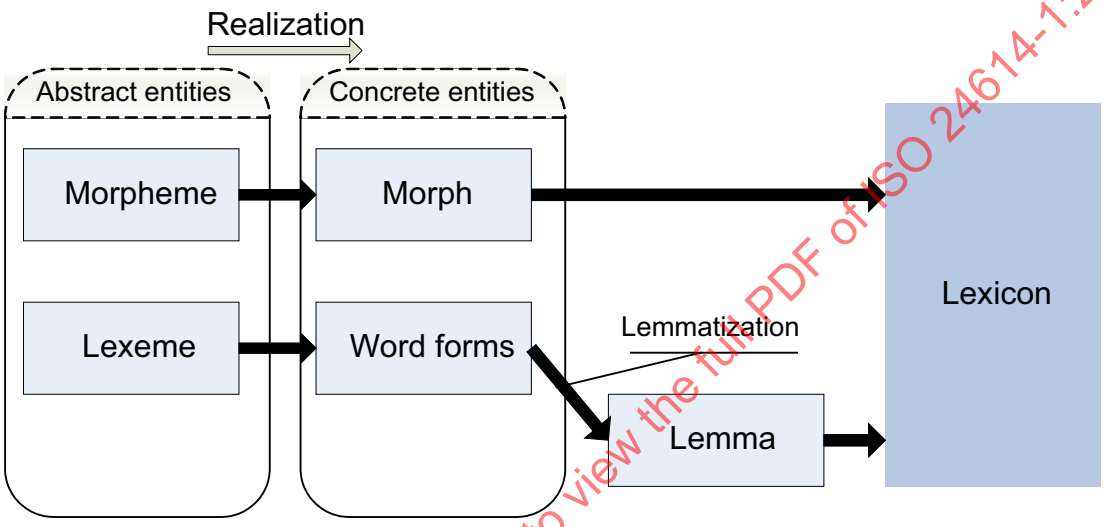

Figure 1 — Relation between abstract and concrete entities in constructing a lexicon

Morphology is the study of the meaningful units of language and how they can be combined to form words. Morphology can be divided into lexical morphology, which is primarily concerned with word formation based on lexemes, and either inflectional morphology or agglutinative morphology (depending on the type of language), which is primarily concerned with word formation based on morphemes. Lexical morphology includes the processes of derivation, compounding, abbreviation, borrowing and reduplication.

NOTE 2 The term “lexical morphology” is used rather than “derivational morphology” since derivation is only one process of word formation.

Inflectional morphology or agglutinative morphology involves two different types of affixation as well as reduplication. Reduplication may result in new word forms, which is why it is also considered a process in lexical morphology. For example, Afrikaans utilizes reduplication to emphasize the meaning of the repeated word, such as “krap” which means “to scratch,” while “krap-krap-krap” means “to scratch vigorously.” For agglutinative languages, where affixes are attached to stems, a particular set of morphological rules are needed in order to perform word segmentation.

NOTE 3 These rules are given in ISO 24614-2.

See Figure 2.

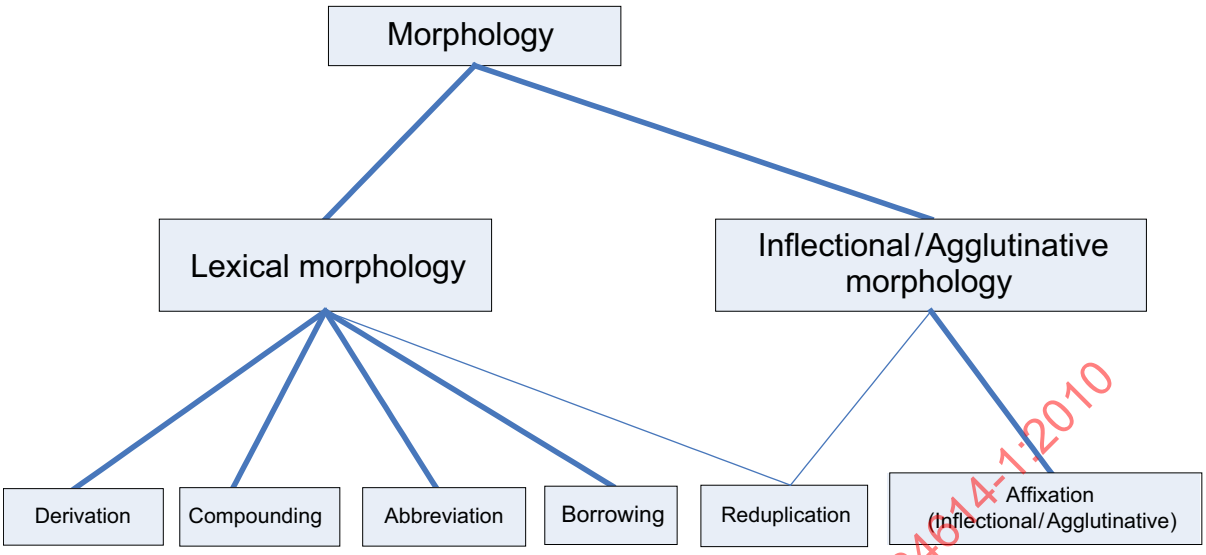

Figure 2 — The system of morphology in languages

Multiword expressions (MWE) include compounds, idioms, proverbs and familiar quotations. See Figure 3. Compounds include word compounds and phrasal compounds. The meaning of a word compound cannot be derived from the meaning of its individual parts. For example, “White House,” as in the US Presidential residence, refers to a unique concept, not just a house that is white. The meaning of a phrasal compound, however, can be derived from the meaning of its individual parts. For example, “apple pie” is a pie made of apples. Although “blueberry pie” is similarly a pie made of blueberries, the former is considered a phrasal compound (see Introduction and the example of 2.20), and therefore comprises one WSU, because the combination of the words “apple pie” is frequent and is even used in the idiomatic expression, “American as apple pie”, whereas “blueberry pie” does not have those properties.

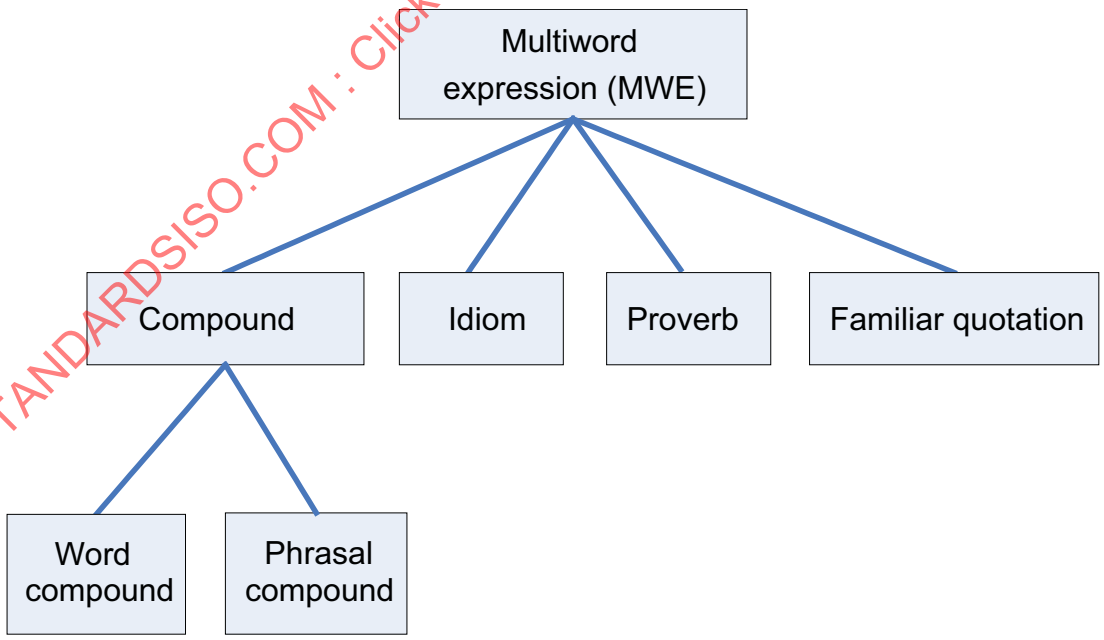

Figure 3 — Types of MWE

WSUs consist of word forms and other character strings. The character strings include or are mixed with numeric or foreign characters, punctuation marks or some other character strings such as radicals in Chinese text or consonant or vowel characters in Korean text. For example, “Bravo!” contains the exclamation mark in the word.

NOTE 4 WSUs contain bound morphemes in some cases such as, in Korean, nominal suffixes “-e” in “hakkyo-e” (school-at) and “-ga” in “hakkyo-ga” (school-nominative) that are treated as belonging to a unique part of speech called “josa” (auxiliary part of speech).

See Figure 4.

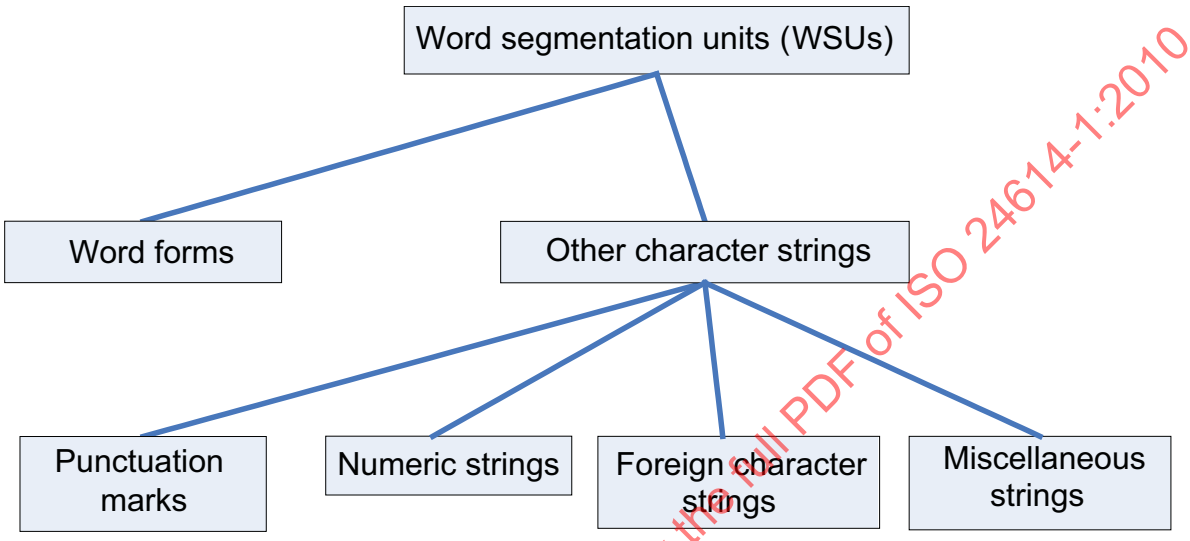

Figure 4 — Types of WSU

3.2 Resources that can facilitate word segmentation

The process of word segmentation within a particular language domain can benefit from the following components and resources:

a) a relevant lexicon;

b) an affix list, including prefixes, suffixes, and infixes, if any;

c) a bound morpheme list, other than affixes;

d) specification for the morphology of the language — to specify the output of word segmentation on the basis of language-dependent phenomena, under the principles described in Clause 4;

e) a representative corpus of a language.

In order to ensure compatibility in word segmentation of different texts (or of one text with different tools) and to ensure that segmentation provides comparable numbers when applied to counting the tokens (see 3.3) of a text document, the resources mentioned in a) to e) above must be described in detail with respect to their contents.

3.3 Process of word segmentation

Figure 5 outlines the process of word segmentation.

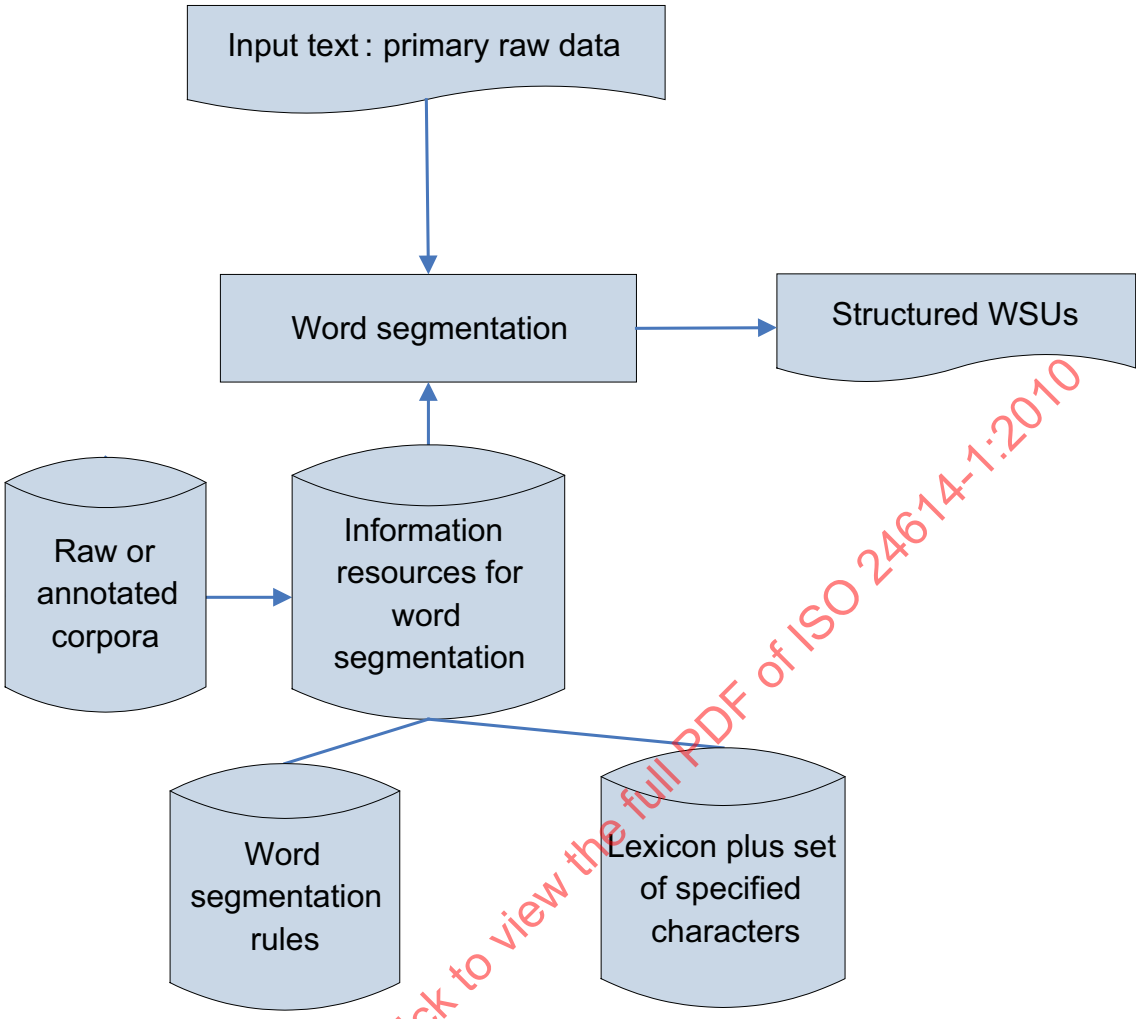

Figure 5 — Word segmentation process

Given primary raw data, the text is segmented into characters and marked with location indexes, and is then segmented into appropriate basic units according to ISO 24612. Raw and annotated corpora provide a basis for constructing a lexicon, which contains word forms and possibly a list of bound morphemes and characters. A set of word segmentation rules is also provided. The corpora, word segmentation rules and lexicon together constitute the resources necessary for transforming the basic segmentation into a segmentation comprising WSUs.

An example in Chinese in graph format is shown in Figure 6.

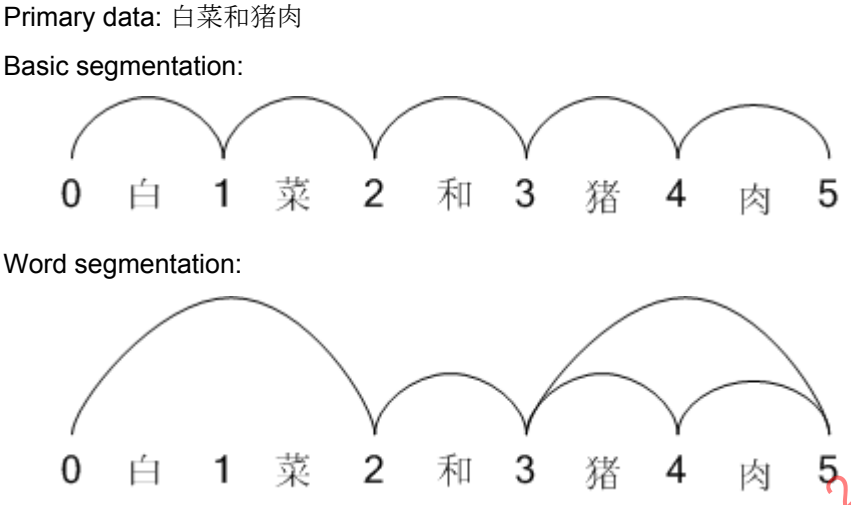

Figure 6 — Illustration of basic segmentation and word segmentation

At the basic segmentation level, each character is marked as a span between two location indexes (e.g. the first character "白" in Figure 6, marked with a span <0,1>). At the linguistic annotation level of word segmentation, the first output "白菜" ("white vegetable") is identified as being a word, marked with a span <0,2>, because the two characters cannot be considered independently. The second unit is a single-character word "和" ("and"). The third unit "猪肉" ("pig meat", "pork") is a phrasal compound, marked with a span <3,5>, with an internal structure which consists of two WSUs "猪" ("pig") and "肉" ("meat"), marked with spans <3,4> and <4,5>, respectively. In the latter case there are two WSUs because the two characters can exist independently and can each contribute to the meaning.

Word segmentation applies to raw text. Word segmentation results in the division of a given text into a sequence of WSUs; a WSU can have an inner segmentation structure when alternative segmentations are allowed. Given a sentence "John left the United States of America" in a fragment of text, it can first be split into segments called "tokens" on the basis of some segmentation rules, here simply on the basis of spacing (for languages that do not use spaces, such as Chinese, different rules need to be used to perform tokenization). Then, by referring to a lexicon, a string of some of these segments, such as "the United States of America", can be treated as a single linguistic unit called a "word" or as an MWE, which is considered to be a type of word. The results of the second stage depend on the contents of the lexicon; some lexicons may not include the whole string "the United States of America" as a lemma, but just "United States of America," or even just "United States".

4 General principles of word segmentation

4.1 The universal principle of morphology

A universal principle and basic foundation of ISO 24614 is that every language has words and smaller units called "morphemes".

4.2 Principles for validating a WSU

4.2.1 General

Two sets of language-independent principles for validating word segmentation units are given: one from a linguistic perspective and the other from a practical perspective. Language-specific exceptions are described in the other parts of ISO 24614, which deal with specific languages. Different principles may apply in different situations, even for identical strings of text.

4.2.2 Principles from a linguistic perspective

a) **Principle of bound morpheme**

If a bound morpheme is attached to a word, then the result is one WSU (e.g. “un” as a bound morpheme in “unhappy”).

b) **Principle of lexical integrity**

The application of syntactic rules does not consider the internal structure of a word. If a word candidate satisfies this principle, then it is likely to be a single WSU. For example, nothing can be inserted between the second and third individual tokens in “the White House,” when referring to the US Presidential residence, such as “the White clean House”, whereas one can say “the white clean house” when referring to any white house.

c) **Principle of unpredictability of a word's meaning from its subparts**

If a word candidate has a property of semantic unpredictability, then it is a single WSU. For example, a “blackboard” is not necessarily black; many are green. Therefore, this word comprises one WSU.

d) **Principle of idiomatic use**

If a sequence of word forms is used idiomatically, then it is treated as a single WSU (e.g. “kick the bucket” used as an idiomatic expression).

e) **Principle of non-productivity**

If a word candidate is unproductive in formation, then it is likely to be a single WSU. For example, “白菜”, literally “white vegetable”, is an unproductive Chinese word, since the character meaning “white” cannot be replaced by a character meaning any other colour; the resulting combination does not exist in Chinese.

4.2.3 Principles from a practical perspective

a) **Principle of frequency**

Frequency is a basic criterion for quantifying the degree of lexicalization of a word candidate. A highly frequent word or sequence of words is likely to be a single WSU.

b) **Gestalt principle** (from cognitive science)

Things are likely to be perceived as a whole. This principle gives evidence for including some phrasal compounds as lemmas in the lexicon even though they appear on the surface to be separate items.

c) **Principle of prototype members in categories** (from cognitive linguistics)

According to the prototype theory regarding the mental lexicon, prototype members in categories are more salient than non-prototype members. They are more accurately remembered in short-term memory and more easily retained and accessed in long-term memory by human beings. This principle provides a rationale for including some phrasal compounds which can serve as prototypes in a productive word