SLOVENSKI STANDARD
SIST ISO 24614-1:2013

01-julij-2013

Upravljanje z jezikovnimi viri - Segmentacija v besede v pisnih besedilih - 1. del: Osnovni pojmi in splošna načela

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

Ta slovenski standard je istoveten z: **ISO 24614-1:2010**

ICS:

01.020	Terminologija (načela in koordinacija)	Terminology (principles and coordination)
01.140.10	Pisanje in prečrkovanje	Writing and transliteration
35.240.30	Uporabniške rešitve IT v informatiki, dokumentiranju in založništvu	IT applications in information, documentation and publishing

SIST ISO 24614-1:2013 en,fr,de

iTeh STANDARD PREVIEW

(standards.iteh.ai)

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 2010

ISO 24614-1:2010(E)

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.

---

iTeh STANDARD PREVIEW

(standards.iteh.ai)

COPYRIGHT PROTECTED DOCUMENT

Contents

Page

iTeh STANDARD PREVIEW
(standards.iteh.ai)

ISO 24614-1:2010(E)

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.

iTeh STANDARD PREVIEW

(standards.iteh.ai)

INTERNATIONAL STANDARD   **ISO 24614-1:2010(E)**

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,

1

— 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.