Using the Japanese Analysis Engine and Access Method

Internationalization > Localization

2003-05-01

Apple Inc.
1 Infinite Loop
Cupertino, CA 95014
408-996-1010

Simultaneously published in the United States and Canada.

Contents

Figures and Tables

Introduction to Using the Japanese Analysis Engine and Access Method

The Language Analysis Manager and the Dictionary Manager use a plug-in architecture. Each allow you to provide plug-ins for analysis engines and dictionary access methods that are specialized for a language and a specific dictionary format. This document describes the analysis environments supported by the Apple Japanese language analysis engine and the dictionary access methods used by that engine.

Concepts

This chapter discusses the Japanese language analysis engine and its supported environments and dictionary access methods. It also provides information about the parts of speech returned in analysis results and lists those that are replaceable between dictionaries.

Japanese Language Analysis Engine

This section discusses the analysis environments supported by the Japanese language analysis engine, provides information about analysis results, and lists the content of the dictionaries used by the analysis engine.

Supported Analysis Environments

The Japanese language analysis engine provides the following analysis environments:

- “KanaKanjiConversion” - Kana-Kanji Conversion Environment

  Kotoeri, Apple's Japanese input method, uses this environment for kana-kanji conversion. It is capable of receiving hiragana character strings and converting them into text with kana and kanji mixed together. In the standard mode, "Basic Dictionary," and "Automatic Learning Dictionary" are used, as well as additional user dictionaries opened by Kotoeri.

  Correction and learning of analysis results are carried out by Kotoeri using the Language Analysis Manager. The types of learning supported are learning the priority of words which are the same part of speech but which have different notation, segment delimiter learning, and learning transliteration to katakana.

- JapaneseMorphemeAnalysis - Mixed Kana and Kanji Text Analysis Environment

  This is an environment provided to analyze normal Japanese which has kana and kanji mixed together. The "Redo conversion" of Kotoeri also uses this environment. It is impossible to convert numbers, symbols and alphabetical letters into readings. Additionally, readings of numerical string suffixes of numbers which come before do not change. In the standard mode, "Basic Dictionary”and “Single Kanji Dictionary” are used.

- JapaneseTextToSpeech - Japanese Text to Speech Environment

  This environment is provided to read out Japanese, and it is capable of converting Japanese with mixed kana and kanji into hiragana resembling the pronunciation. Symbols are not converted into readings, but numerals and letters of the alphabet are converted. In addition, changes in readings of numerical string suffixes, and replacement of the position of numerical prefixes are supported. In such cases, numerical suffixes and numerical prefixes are brought together as one, and treated as ordinary nouns. In the standard mode, “Basic Dictionary” “and “Single Kanji Dictionary” are used. With regard to learning, it works the same as the environment for analyzing kana and kanji mixed text.

Additional Information Contained in Analysis Results

In addition to the standard word information provided in the homograph node (`LAHomograph`) of the analysis results, the following information is returned:

- Data Character String (`keyAEText`, `typeUnicodeText`)

  For a kana-kanji conversion environment it contains a notation character string, for an environment for analyzing mixed kana and kanji text it contains a reading character string, and for the Text to Speech environment it contains a character string resembling the pronunciation.

- Dictionary Information (`keyAEHomographDicInfo`, `typeAEHomographDicInfo`)

  In all environments, structures defined in the following manner are returned if a word is obtained from the dictionary:

```
struct HomographDicInfoRec {
        DCMDictionaryID dictionaryID;
        DCMUniqueID uniqueID;
};
```

- Word Weight (`keyAEHomographWeight`, `typeShortInteger`)

  In all environments, the "weight" indicates the priority of that word is returned. This contains a value that indicates the priority within a group which has the same key character string and part of speech (larger numbers have priority), and users must be aware that it is not the absolute frequency of that word.

- Position of Word Accent (`keyAEHomographAccent`, `typeAEHomographAccent`)

  In the Text to Speech environment, an integer (1 byte) has been added which indicates the position of the accent of that word.

Figure 2-1 (page 10) shows the likely structure of an `LAHomograph` contained in the results of analyzing a character string in the "JapaneseMorphemeAnalysis" environment.

Figure 1-1 An LAHomograph

Analysis of 今日

LAHomograph	
Type	typeAERecord
	Key: keyAEText; Type: typeUnicodeTex; きょう
	Key: keyAEHomographDicInf; Type: typeAEHomographDicInf; DictionaryID = 1, UniqueID = 0
	Key: keyAEHomographWeigh; Type: typeShortInterge; 1

Dictionaries Used by Analysis Engine and Their Main Content

The main fields contained in the respective dictionaries used by an analysis engine and their specifications are given here. Actual dictionaries contain several other items of information apart from that shown here, but that content is private. With regard to the handling of data stored in the "'hins' - internal part of speech code" field of each dictionary, see "Parts of Speech" (page 14).

Basic Dictionary

Table 1-1 Dictionary properties for the basic dictionary

Property	Value	Data type	Comments
pDCMClass	kSCMBasicDictionaryClass	typeShortInteger	Dictionary class
pDCMPermission	kReadOnlyDictionary	typeShortInteger	Permission to read or write
pDCMListing	kDCMProhibitListing	typeShortInteger	Permission to list contents

Table 1-2 Dictionary fields for the basic dictionary

Field	Data type	Indexed	Maximum length	Comments
yomi	typeUnicodeText	Yes	40 bytes	Reading character string
hyok	typeUnicodeText	No	64 bytes	Notation character string
hins	typeShortInteger	No	2 bytes (fixed length)	Internal part of speech
hind	typeShortInteger	No	2 bytes (fixed length)	Word weight

Table 1-3 Dictionary fields for the basic dictionary, reverse lookup

Field	Data type	Indexed	Maximum length	Comments
yomi	typeUnicodeText	No	40 bytes	Reading character string
hyok	typeUnicodeText	Yes	64 bytes	Notation character string
hins	typeShortInteger	No	2 bytes (fixed length)	Internal part of speech
hind	typeShortInteger	No	2 bytes (fixed length)	Word weight
acnt	typeAEHomographAccent	No	1 byte (fixed length)	Position of accent
hton	typeUnicodeText	No	64 bytes	Pronunciation notation

Single Kanji Dictionary

Table 1-4 Dictionary properties for the single kanji dictionary

Property	Value	Data type	Comments
pDCMClass	kSCMBasicDictionaryClass	typeShortInteger	Dictionary class
pDCMPermission	kReadOnlyDictionary	typeShortInteger	Permission to read or write
pDCMListing	kDCMProhibitListing	typeShortInteger	Permission to list contents

Table 1-5 Dictionary fields for the single kanji dictionary

Field	Data type	Indexed	Maximum length	Comments
yomi	typeUnicodeText	Yes	40 bytes	Reading character string
hyok	typeUnicodeText	No	64 bytes	Notation character string
hins	typeShortInteger	No	2 bytes (fixed length)	Internal part of speech
hind	typeShortInteger	No	2 bytes (fixed length)	Word weight

Table 1-6 Dictionary fields for the single kanji dictionary, reverse lookup

Field	Data type	Indexed	Maximum length	Comments
yomi	typeUnicodeText	No	40 bytes	Reading character string
hyok	typeUnicodeText	Yes	64 bytes	Notation character string
hins	typeShortInteger	No	2 bytes (fixed length)	Internal part of speech
hind	typeShortInteger	No	2 bytes (fixed length)	Word weight
OnKn	typeUnicodeText	No	128 bytes	Chinese/Japanese reading character string

User Dictionary

Table 1-7 Dictionary properties for the user dictionary

Property	Value	Data type	Comments
pDCMClass	kSCMBasicDictionaryClass	typeShortInteger	Dictionary class
pDCMPermission	kReadOnlyDictionary	typeShortInteger	Permission to read or write
pDCMListing	kDCMProhibitListing	typeShortInteger	Permission to list contents

Table 1-8 Dictionary fields for the user dictionary

Field	Data type	Indexed	Maximum length	Comments
yomi	`typeUnicodeText`	Yes	64 bytes	Reading character string
hyok	`typeUnicodeText`	No	128 bytes	Notation character string
hins	`typeShortInteger`	No	2 bytes (fixed length)	Internal part of speech
hind	`typeShortInteger`	No	2 bytes (fixed length)	Word weight

Table 1-9 Dictionary fields for the user dictionary, reverse lookup

Field	Data type	Indexed	Maximum length	Comments
yomi	`typeUnicodeText`	Yes	64 bytes	Reading character string
hyok	`typeUnicodeText`	No	128 bytes	Notation character string
hins	`typeShortInteger`	No	2 bytes (fixed length)	Internal part of speech
hind	`typeShortInteger`	No	2 bytes (fixed length)	Word weight

Other Specifications

Morpheme analysis method: minimum cost method

Maximum input character string length: 200 byte

Type of characters to be used as reading and notation: no restrictions. However, Unicode is used as the dictionary's internal information and internal processing code, so characters which cannot be converted to Unicode cannot be used.

Dictionary Access Methods

The dictionary access methods for Japanese analysis engines provided by Apple have the name “DAM:Apple Backward Trie Access Method”. “DAM:” is a shared prefix which indicates that this module is a dictionary access method.

Search Methods

The following three search methods are supported by the access methods:

- `kDCMFindMethodExactMatch`, find an exact match.
- `kDCMFindMethodEndingMatch`, match to the end (bat matches combat, acrobat).
- `kDCMFindMethodBackwardTrie`, find backward Trie, match partial character string backward (flash matches ash, lash, flash).

In this access method it is also possible for each dictionary to have up to two indexed fields which can be used for searching (key fields). If setting up two key fields, when using `DCMCreateFieldInfoRecord` to create the field definition information specified in `DCMNewDictionary`, the key field created first is stored to the file named dictionary_1, and the key field created second to the file dictionary_2. These two files are stored in the file package which is externally treated as a dictionary.

Data Stored in a Dictionary

This access method is not concerned with the content of stored data, so it is possible to store all types of data in accordance with the definitions of the dictionary fields. However, there is the restriction that the maximum length for one data record is approximately 4,000 bytes, and the maximum length of data that can be stored in one field of a record is 255 bytes.

Dictionaries are managed internally in blocks of 4,096 byte units, and the maximum number of blocks that one dictionary can have is 65,534 blocks. The maximum number of records that can be registered varies according to the length of data stored (if one record is 50 bytes on average, it is approximately 5 million records).

Dictionary Properties

Fundamental dictionary properties such as `pDCMPermission`, `pDCMListing`, and `pDCMClass` are saved as a part of the management information of the main body of the dictionary. Other unknown properties are saved to the the Properties.plist file in the dictionary package.

With regard to both `pDCMPermission` and `pDCMListing` properties, changing from Read Only to Read/Write and data dump disabled to enabled direction respectively is not permitted, and `dcmPermissionErr` is returned.

Parts of Speech

Standard parts of speech set are defined as the parts of speech returned by the analysis, and parts of speech that are registered to dictionaries. Both are defined as 32-bit unsigned integers, but only the lower-place 16 bits are actually used, with the upper-place 16 bits reserved for future extensions. Parts of speech are classified hierarchically into three stages from rough classification to detailed classification, and it is possible to use a degree of detail in accordance with the objective.

Be aware that in the parts of speech contained in analysis results, in contrast to those used to indicate the form of utilization of parts of speech used by the lowest-place 4 bits, the form of utilization is not defined for parts of speech used when registering dictionaries, and it is only possible to register the word stems to dictionaries. Auxiliary verbs and post-positional words appear in analysis results, however they are not defined as parts of speech which can be registered in dictionaries.

The parts of speech field of actual dictionaries contain undisclosed internal parts of speech codes as `typeShortInteger`. The morpheme node of analysis results returned from analysis engines also contains parts of speech in the same format (`keyAEMorphemePartOfSpeechCode`), but by fetching this data in the form of a `typeAEMorphemePartOfSpeechCode`, a format conversion is carried out automatically by the coercion handler of the Apple Event manager, and the standard part of speech code given here is returned.

Similarly, when carrying out find/registration using the Dictionary Manager directly, the format conversion necessary to enable the data of the field specified in `kDCMJapaneseHinshiTag` to be handled in the format of `kDCMJapaneseHinshiType` is carried out.

Parts of Speech in Analysis Results

The parts of speech contained in analysis results are shown in Table 2-10. Bits 16 through 31 are unused. See Table 2-11 (page 17) for a list of all variations of parts of speech. Note that conjugations are classified into the following seven categories.

1. Stem
2. "Mizen" form
3. "Renyo" form
4. "Syusi" form
5. "Rentai" form
6. "Katei" form
7. "Meirei" form

The values in the Conjugations column of Table 2-10 refer to these seven classifications.

Table 1-10 Parts of speech information contained in analysis results

Rough classification (bits 12 - 15)	Medium classification (bits 8 - 11)	Strict classification (bits 4 - 7)	Conjugations (bits 0 - 3)
Nouns	Common noun		0
	Person name		0
		Surname	0
		First name	0
	Place name		0
		Place name with suffix	0
	Organization name		0
	Proper noun		0
	"Sa hen" nouns		0
	"Kei-do"nouns		0
"Rentai-shi"			0
Adverbs			0

Rough classification (bits 12 - 15)	Medium classification (bits 8 - 11)	Strict classification (bits 4 - 7)	Conjugations (bits 0 - 3)
Conjunctions			0
Interjections			0
Verbs	“5 dan” verbs	“KA gyo 5 dan” verbs	1-7
		“SA gyo 5 dan verbs”	1-7
		“TA gyo 5 dan verbs”	1-7
		“NA gyo 5 dan verbs”	1-7
		“MA gyo 5 dan verbs”	1-7
		“RA gyo 5 dan verbs”	1-7
		“WA gyo 5 dan verbs”	1-7
		“GA gyo 5 dan verbs”	1-7
		“BA gyo 5 dan verbs”	1-7
	“1 dan” verbs		1-7
	“Ka hen” verbs		1-7
	“Sa hen” verbs		1-7
	“Za hen” verbs		1-7
Adjectives			1-7
Adjective verbs			1-7
Prefixes			0
	Numerical prefixes		0
Suffixes			0
	Person name suffixes		0
	Place name suffixes		0
	Organization name suffixes		0
	Numerical suffixes		0
Non category	Single character		0
	Symbols		0
		Period	0

Rough classification (bits 12 - 15)	Medium classification (bits 8 - 11)	Strict classification (bits 4 - 7)	Conjugations (bits 0 - 3)
		Comma	0
	Numerals		0
	Independent words		0
	Idiomatic phrases		0
Auxilliary verb			1-7
Propositional particle			0

Table 1-11 All variations of parts of speech that can be contained in analysis results

All variations
Common noun
Person name
Surname
First name
Place name
Place name with suffix
Organization name
Proper noun
“Sa hen” nouns
“Kei-do”nouns
“Rentai-shi”
Adverbs
Conjunctions
Interjections
“KA gyo 5 dan” verbs
“SA gyo 5 dan verbs”
“TA gyo 5 dan verbs”
“NA gyo 5 dan verbs”

All variations
“MA gyo 5 dan verbs”
“RA gyo 5 dan verbs”
“WA gyo 5 dan verbs”
“GA gyo 5 dan verbs”
“BA gyo 5 dan verbs”
“1 dan” verbs
“Ka hen” verbs
“Sa hen” verbs
“Za hen” verbs
Adjectives
Adjective verbs
Prefixes
Numerical prefixes
Suffixes
Person name suffixes
Place name suffixes
Organization name suffixes
Numerical suffixes
Single character
Symbols
Period
Comma
Numerals
Independent words
Idiomatic phrases
Auxiliary verb
Propositional particle

Parts of Speech Replaceable Between Dictionaries

Table 2-12 lists the parts of speech that are replaceable between dictionaries. See Table 2-13 (page 20) for a list of all variations of parts of speech.

Table 1-12 Parts of speech information replaceable between dictionaries

Rough classification (bits 12 - 15)	Medium classification (bits 8 - 11)	Strict classification (bits 4 - 7)
Nouns	Common nouns	
	Person name	
		Surname
		First name
	Place name	
		Place name with suffix
	Organization name	
	Proper noun	
	“Sa hen” nouns	
	“Kei-do”nouns	
“Rentai-shi”		
Adverbs		
Conjunctions		
Interjections		
Verbs	“5 dan” verbs”	“KA gyo 5 dan” verbs
		“SA gyo 5 dan verbs”
		“TA gyo 5 dan verbs”
		“NA gyo 5 dan verbs”
		“MA gyo 5 dan verbs”
		“RA gyo 5 dan verbs”
		“WA gyo 5 dan verbs”
		“GA gyo 5 dan verbs”
		“BA gyo 5 dan verbs”
	“1 dan” verbs	

Rough classification (bits 12 - 15)	Medium classification (bits 8 - 11)	Strict classification (bits 4 - 7)
	“Ka hen” verbs	
	“Sa hen” verbs	
	“Za hen” verbs	
Adjectives		
Adjective verbs		
Prefixes		
	Numerical prefixes	
Suffixes		
	Person name suffixes	
	Place name suffixes	
	Organization name suffixes	
	Numerical suffixes	
Non category	Single character	
	Symbols	
		Period
		Comma
	Numerals	
	Independent words	
	Idiomatic phrases	

Table 1-13 All variations of parts of speech that are replaceable between dictionaries

All variations
Common noun
Person name
Surname
First name
Place name
Place name with suffix

All variations
Organization name
Proper noun
“Sa hen” nouns
“Kei-do”nouns
“Rentai-shi”
Adverbs
Conjunctions
Interjections
“KA gyo 5 dan” verbs
“SA gyo 5 dan verbs”
“TA gyo 5 dan verbs”
“NA gyo 5 dan verbs”
“MA gyo 5 dan verbs”
“RA gyo 5 dan verbs”
“WA gyo 5 dan verbs”
“GA gyo 5 dan verbs”
“BA gyo 5 dan verbs”
“1 dan” verbs
“Ka hen” verbs
“Sa hen” verbs
“Za hen” verbs
Adjectives
Adjective verbs
Prefixes
Numerical prefixes
Suffixes
Person name suffixes

All variations
Place name suffixes
Organization name suffixes
Numerical suffixes
Single character
Symbols
Period
Comma
Numerals
Independent words
Idiomatic phrases

Document Revision History

This table describes the changes to *Using the Japanese Analysis Engine and Access Method*.

Date	Notes
2003-05-01	Made a minor change to the title.
	Made minor edits, updated formatting.
	Removed enumerations that defined Japanese parts of speech. These are defined in *Inside Mac OS X: Language Analysis Manager Reference*.
1998-11-01	Preliminary document published under the title *"Using Apple Japanese Analysis Engine and Access Method."*

Document Revision History