A Trouble with Computer Arrangement of Natural Words

HIROHIKO NISIMURA*

A large list of sorted words and phrases is a basic material for researches in linguistic information processing. A computer arrangement of phrases differs partially from conventional arrangement and is found insufficient for human use when the phrases are spelled with hyphen, period, and other special characters; as a computer locates a string in the list according to the collating sequence of characters.

On the contrary, a conventional arrangement of words in dictionaries is based on key strings composed of 26 letters excluding special characters. A large list for human examination will be sorted by a computer utilizing such key strings derived from the original spellings.

A sorting key consists of three parts, a primary key, an intermediate key and a minor key. The primary key string is composed of letters, justified left and space-filled, which are extracted from the original spelling. The intermediate key string denotes the variation of letters, such as capitallization and hyphenation. The variations may be denoted by figures as follows:

Alphabet		*Japanese kana*	
space	0	space	0
special character	3	special character	1
upper case letter	6	small hirakana	2
lower case letter	8	small katakana	3
		voiceless hirakana	4
		voiceless katakana	5
		voiced hirakana	6
		voiced katakana	7
		semivoiced hirakana	8
		semivoiced katakana	9

The minor key string is the original spelling as itself.

An example of sorting key is as follows:

primary key	*intermediate key*	*minor key*
allfoolsday	68806888830688	All Fools' Day

The discussion is applicable to Alphabetic words and Japanese words.

This paper first appeared in Japanese in Joho-Shori (the Journal of the Information Processing Society of Japan), Vol. 10, No. 1 (1969), pp. 21-25.

* Electrotechnical Laboratory, MITI, Tokyo.