What is a multiword expression?

Lexicographic criteria for including MWEs in bilingual dictionaries

by **Jack Halpern** <jack@cjki.org>
The CJK Dictionary Institute (cjk.org)

This paper discusses some of the fundamental principles for the selection of headwords in bilingual dictionaries. A basic assumption in lexicography is that the linguistic units of one language map to those of another language, but even in close language pairs like Spanish and English there are numerous exceptions. In some language pairs, such as English and Japanese, *cross-linguistic lexical anisomorphism* (CLA) is so great that it becomes literally impossible to map certain words and phrases across these languages.

A string of words can be segmented into components in multiple ways. Linguists may disagree on how to combine these components to form meaningful linguistic units. The primary task of the lexicographer is to decide which combination of components qualifies as a dictionary headword, which may depend on the application (language learning, MT, morphological analysis). In this paper we define these units based on decades of experience in bilingual lexicography, especially the compilation of an extremely comprehensive full-form Spanish-English dictionary for an innovative machine translation project. We will define the following linguistic units, noting that the definitions may be rigorous or mutually exclusive.

Lexical unit is the smallest distinctive linguistic unit that associates meaning with form. The near-synonym *lexeme* emphasizes the members of an inflectional paradigm, rather than a specific wordform. *Free word combination* is a meaningful free sequence of words that follows the rules of syntax but has no lexical status. *Phraset* is a free combination of words that is recurrently used to express concepts but has no monolingual lexical status but maps cross-lingually. *Collocation* is a recurrent combination of words co-occurring more often than by chance whose meaning is (mostly) compositional and transparent. *Multiword expression* consists of multiple words that together function as a single lexical unit. *Word group* is a group of multiple contiguous orthographic words that have no special significance other than that the words occur contiguously.

It is hoped that a detailed discussion of these concepts will contribute to headword selection based on (mostly) objective criteria.

About Jack Halpern

Jack Halpern (春遍雀來), CEO of The CJK Dictionary Institute, is a lexicographer by profession. For sixteen years was engaged in the compilation of the New Japanese-English Character Dictionary, and as a research fellow at Showa Women's University (Tokyo), he was editor-in-chief of several kanji dictionaries for learners, which have become standard reference works.

Jack Halpern, who has lived in Japan over 40 years, was born in Germany and has lived in six countries including France, Brazil, Japan and the United States. An avid polyglot who specializes in Japanese and Chinese lexicography, he has studied 15 languages (speaks ten fluently) and has devoted several decades to the study of linguistics and lexicography.

On a lighter note, Jack Halpern loves the sport of unicycling. Founder and long-time president of the International Unicycling Federation, he has promoted the sport worldwide and is a director of the Japan Unicycling Association. Currently, his passions are playing the quena and improving his Chinese, Esperanto and Arabic.

The CJK Dictionary Institute

The CJK Dictionary Institute, Inc. (CJKI) specializes in CJK and Arabic computational lexicography. The institute creates and maintains CJK (Chinese, Japanese and Korean) and Arabic lexical databases currently covering approximately 50 million entries. Located in Saitama, Japan, CJKI is headed by Jack Halpern, editor-in-chief of the world-renowned New Japanese-English Character Dictionary and of various other CJK dictionaries.

CJKI plays a leading role in helping the IT industry penetrate the lucrative East Asian market by providing software developers with high quality dictionary data. This includes comprehensive databases of general vocabulary, proper nouns and technical terms for CJK languages, including Chinese dialects such as Cantonese and Hakka. CJKI also maintains databases and romanization systems of Arabic proper nouns, a large-scale Spanish-English dictionary, and various multilingual databases of proper nouns and geographic data.

CJKI has become one of the world's prime sources for CJK lexical resources. It is contributing to CJK and Arabic information processing technology by providing high-quality lexical resources and professional consulting services to some of the world's leading software developers and IT companies, including Fujitsu, Sharp, Sony, IBM, Google, Microsoft, Yahoo, Amazon and Baidu.

The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 JAPAN
Phone: 048-473-3508 Fax: 048-486-5032
Email: jack@cjki.org   Web: www.cjk.org