

Pedagogical Lexicography Applied to Chinese and Japanese Learner's Dictionaries

Jack Halpern
The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi
Saitama 352-0001 JAPAN
jack@cjki.org

Abstract

Chinese dictionaries, both monolingual and bilingual, suffer from several drawbacks that render them mostly inadequate for the serious learner of Chinese. This paper introduces *The CJKI Chinese Learner's Dictionary (CCLD)*, a work designed to satisfy the needs of learners by presenting abundant lexicographic information on the most frequently used characters and compounds. The paper also analyzes some of the shortcomings of existing dictionaries, discusses some differences between Chinese and Japanese compilation strategies, and describes an effective character lookup method adopted in Chinese dictionaries for the first time.

1. Background and Aims

Though Chinese is the most widely spoken language in the world, the lack of pedagogically effective dictionaries puts the learner at a disadvantage compared with those of other major world languages. Traditional Chinese dictionaries, often rooted in classical Chinese, have various shortcomings and thus do not adequately meet the needs of learners of Chinese as a foreign language. These shortcomings include inaccurate or archaic equivalents, historical sense ordering, absence of lexical categories, inefficient lookup methods, poor design, and a failure to distinguish bound morphemes from free ones.

CCLD aims to address the shortcomings of existing dictionaries in a systematic manner. The approach is based on the same principles used to compile two kanji dictionaries that have gained wide acceptance in Japanese language education. The first is Kenkyusha's *New Japanese-English Character Dictionary* (1990, 1993), and the second is *The Kodansha Kanji Learner's Dictionary* (1999). These are part of a series of kanji dictionaries and applications, referred to as Kanji Integrated Tools (KIT), designed to help learners master kanji (six have been published so far). CCLD is the first release of a new series of dictionaries and tools for the study of hanzi, referred to as Hanzi Integrated Tools (HIT).

The primary aim of CCLD is to provide an in-depth understanding of how Chinese characters are used. This is achieved by presenting abundant information on the forms, meanings, readings and functions of the most frequently used characters and compounds in a user-friendly design that promotes understanding and stimulates a desire to learn.

2. Compilation Techniques

The techniques used for compiling CCLD are based on a systematic approach that emerged from the author's several decades of experience in compiling learner's dictionaries and in studying 14 languages. Computational lexicography was combined with the latest advances in DTP technology to produce a work that meets the specific needs of beginning and intermediate learners. In addition, various tools were used to perform sanity checks and validation at every stage of the project to ensure data integrity, accuracy and consistent implementation of editorial policy.

This dictionary is firmly committed to the descriptive approach. Unlike most existing dictionaries, the aim was to record usage as it actually occurs in the living language, not to cover obscure or archaic meanings occurring only in Classical Chinese. Although dozens of dictionaries were consulted, no meaning was included merely on the authority of other dictionaries. Word meanings were extracted from actual occurrences, while character meanings were determined by such methods as componential analysis to extract the semantic, syntactic and morphological features for each sense.

Several criteria were used to select the entry characters and compound words. First, all words that appear in the latest edition, as well as in older editions, of the well-known Chinese proficiency test *hànyǔ shuǐpíng kǎoshì* (汉语水平考试) (HSK) were included. The rest of the entries were selected on the basis of frequency statistics extracted from a corpus called Chinese Gigaword (LDC), possibly the largest Chinese corpus ever compiled. However, the selection of entries for learner's dictionaries is not a mechanical process that can be based on raw frequency alone. Native Chinese editors manually reviewed the items selected on the basis of frequency, rejecting unsuitable ones and adding some that were beyond the frequency threshold but were deemed useful to the learner.

3. Lexical Categories

Chinese lexical units often belong to multiple lexical categories. A verb like 合作 *hézuò* 'cooperate', for example, can be used as a noun meaning 'cooperation', while 矛盾 *máodùn* 'contradiction; contradictory' can function both as a noun and as a stative verb. The common occurrence of such cross-categorical lexical units makes it seem as if Chinese lexical categories are not well defined. There is a grain of truth to this statement, and sometimes it almost seems as if speakers are free to switch lexical categories at will (not unlike the English *but me no buts*).

Classical Chinese scholars neglected the study of lexical categories, and, as pointed out by Dong, the issue has not attracted the interest of lexicographers and most Chinese dictionaries do not provide POS codes (Dong: 24-32). As Western linguistic concepts penetrated Chinese linguistic studies in the first half of the 20th century, Chinese lexicographers began to realize that lexical categories are a legitimate part of Chinese grammar, and POS codes slowly began to appear in reference works. Various pedagogical dictionaries published outside China gradually started to provide POS codes, but dictionaries published in China still lag behind. Probably the most complete analysis of Chinese lexical categories was undertaken by Yu Shiwen (1998).

There are three issues related to POS codes:

1. *Lack of POS codes.* A recent survey by CJKI (Survey, 2011) showed that of the 19 dictionaries that provide no POS codes, 12 (or 63%) are published in China, including such well-known monolingual dictionaries as *Xinhua Zidian* (2004) and bilingual dictionaries like *The Contemporary Chinese Dictionary* (2002). The failure of many dictionaries to provide POS codes is linguistically untenable, and most inconvenient to learners. As in other languages, Chinese lexical units are governed by syntactic constraints that determine their behavior as members of specific lexical categories. For example, verbs cannot (normally) be modified by intensifiers like 很 *hěn* ‘very’, while stative verbs like 漂亮 *piàoliang* ‘(be) beautiful’ can be modified by intensifiers but cannot take direct objects.
2. *Limited POS codes.* The CJKI survey also showed that one third (10 out of 31) of the dictionaries that do give POS codes limit them to single-character (monomorphemic) entries, as is the case with *A Comprehensive Chinese-English Dictionary* (2004), the most comprehensive bilingual Chinese dictionary today. It goes without saying that Chinese polymorphemic entries (compound words) belong to specific classical categories, just like any other lexical unit, so there is absolutely no justification for this limitation.
3. *Incomplete POS codes.* The survey further showed that POS codes are often incomplete or inaccurate. 合作 *hézuò* ‘cooperate; cooperation’, for example, is both a verb and noun but 17 out of 21 dictionaries label it only as a verb. 通常 *tōngcháng* ‘usually; usual’ is both an adverb and adjective but only one dictionary labels it as such, while another dictionary mislabels it as a noun. From a pedagogical point of view, such errors are of paramount importance. To the learner, it is not at all obvious that the verb 合作 can also be used as a noun. Learners cannot just assume that any verb can be used as a noun, and vice versa, because this is often not the case. For example, the noun 电话 *diànhuà* ‘telephone’ cannot be used as a verb, while the verb 喝 *hē* ‘to drink’ cannot be used as a noun.

An important feature of CCLD is that it gives POS codes for all entries, including compounds, as shown in Figure 1. Since the POS codes in CCLD were determined on the basis of semantic analysis of the syntactic role of each sense, rather than on the authority of other dictionaries, they are accurate, complete and reliable.

4. Transparency and Productivity

An important characteristic of Chinese characters is their ability to convey meaning; that is, their logographic and morphographic nature (Wang, Inhoff, Chen: 259). They are also highly morphologically productive — by combining a few thousand characters, countless compound words are formed. This is similar to Latin and Greek roots in English, e.g. *hydr-*, *aqua-* in *hydrophobia* and *aquarium*. Unlike English, the corresponding Chinese and Japanese words 恐水病 and 水族館 share the grapheme 水, which transcends their pronunciation and makes it clear at a glance what they mean; that is, they are semantically transparent.

A salient feature of both CCLD and KIT dictionaries is the central role that semantic

transparency and morphological productivity have played in determining editorial policy. As is explained below and as shown in Figure 1, character meanings are presented in a manner (1) that shows how compounds are composed from their constituents (semantic transparency) and (2) that enables users to infer the meanings of compounds not listed in the dictionary (morphological productivity).

5. Semantic Levels

The meanings associated with a single character may be quite complex. In both Chinese and Japanese, characters can have *morphemic meanings* (meanings of bound forms) and *lexemic meanings* (meanings of free forms). In Japanese, character meanings operate on four distinct but interrelated semantic levels (levels of meaning):

- L1: as an *on* (Chinese-derived) word element (morphemic)
- L2: as a *kun* (native Japanese) word element (morphemic)
- L3: as an *on* (Chinese-derived) free form (lexemic)
- L4: as a *kun* (native Japanese) free form (lexemic)

These levels may interact in a complex way, from partial or absolute equivalence to total nonequivalence, and on each level may have numerous meanings and multiple functions (bound morpheme, affix, counter etc.) For example, on L1 著 *cho* means ‘write, publish; conspicuous; literary work’, on L3 ‘literary work’, and on L2 and L4 ‘write, publish’ (著す *arawasu*) and ‘conspicuous’ (著しい *ichijirushii*).

In Chinese, character meanings can be classified into several levels based on degree of boundness, such as free, semibound (measure words), bound and free in restricted contexts (such as idioms). Zhang combines the above levels with monosemy and polysemy (2001: 33-41). For example, he assigns the level "polysemic morphemic/lexemic" to 兵 *bīng*, since it has multiple senses, some of which are lexemic or morphemic only, while some are both morphemic and lexemic.

A major drawback of the absolute majority of Chinese dictionaries is that they have no labels to distinguish morphemic meanings from lexemic ones, which is especially problematic for monomorphemic entries. For example, the first sense for 展 *zhǎn* in *A Comprehensive Chinese-English Dictionary* (2004) is ‘open up...unfold’. Since there is no label to show that it is used as a bound morpheme, the user is misled to think that 展 can be used as a free word meaning ‘unfold’.

A major feature of CCLD is the explicit indication of semantic levels. As can be seen in Figure 1, the semantic level for each sense of the entry character is distinguished by various symbols that indicate the degree of boundness, i.e. free form (△), bound form (▲), and free or bound form (△). In addition, a system of functional labels such as [suffix] and [prefix] and various others indicate different types of affixes. This not only ensures that the user does not confuse bound morphemes with free words, but also illustrates the character's wide range of morphological productivity.

6. Character Meanings

6.1 The Equivalent

Every effort has been made to present precisely worded character meanings in a manner that helps the learner understand them in-depth. Equivalents are grouped in a manner that shows how the single senses are related to each other through the core meaning, and may include various explanatory glosses and other devices to supplement the meaning (see Figure 1). Equivalents also include sense division numbers and various symbols and labels such as functional labels, status labels, POS labels and temporal labels, followed by numerous compounds that illustrate each sense.

6.2 Core Meaning

A salient feature of both CCLD and all KIT dictionaries is the presentation of a *core meaning*. This is a concise keyword that provides a clear grasp of the central or most fundamental concept that links the principal senses of a character into a single conceptual unit. Figure 1 shows a snippet of the CCLD entry for 留 that illustrates the pedagogical efficacy of core meanings.

By grasping that the central concept represented by 留 is KEEP, the learner can see how such seemingly unrelated senses as ‘reserve’, ‘detain’, and ‘concentrate on’ are variants of the same basic concept. Color coding (red capitals) is used to identify the core meanings, making it clear how the senses are interrelated. The core meaning thus integrates widely differing senses into a single conceptual unit. It is useful to learners in several ways:

1. It concisely conveys the character's most fundamental meaning.
2. It acts as the central pivot that interrelates the principal senses to each other.
3. It serves as a mnemonic that encapsulates the character's multiple senses.

Though linguistically it is often not possible to isolate a single sense from which all other senses can be logically derived, the core meaning comes close to playing that role. It often represents the direct, psychologically most dominant, meaning — in Edward Sapir’s words, its “conceptual kernel” (1921: 24-41) — the meaning that would occur to a native speaker presented with the character in isolation. Since the core meaning functions as a concentrated thought package and appeals to the learner’s powers of association, it serves as an effective learning aid.

6.3 Interrelatedness of Senses

Chinese characters can be highly polysemous, with many senses apparently unrelated to each other. To stimulate a deeper understanding of character meaning, whenever possible senses are presented in a manner that shows their interrelatedness, as shown in Figure 1. In CCLD, this is achieved by *logical* ordering of senses, sense disambiguation glosses, and various typographic devices such as capitalized core words and indented sense division numbers that establish a logical hierarchy between the senses.

In contrast, the vast majority of Chinese dictionaries list senses in chronological order, beginning with the original meaning of the character. For example, the first meaning given for

本 *běn* is almost invariably ‘root (of a tree)’, misleading the user to believe that this rare sense is important. Although historical ordering is of value to the scholar, it is not useful to the learner since it does not reflect contemporary usage, and, more importantly, because archaic senses often appear first without any indication of their morphemic and temporal status.

Figure 1: Snippet of CCLD entry for 留
(Note: some senses and compounds are omitted)

留

1646

■ 2-5-5

► KEEP ► STAY

liú

㊶ 留 ㊷ 留

田	HSK-4	S10-5-5	GB3384
102	B0741	㊶2580	U7559

㇀ ㇁ ㇂ ㇃ ㇄ ㇅ ㇆ ㇇ ㇈ ㇉

1 2 3 4 5 6 7 8 9

留

10

① ㊶ [original meaning] **KEEP in place, ask (someone) to STAY** *VB*[△]

- b** KEEP for future use, reserve, leave behind[△]
- c** KEEP in custody, detain[△]
- d** KEEP one's mind on, concentrate on[△]

a 我们留她吃晚饭。 *Wǒmen liú tā chī wǎnfàn* *EH* We asked her to stay for dinner.
留步 *liúbù* *EA* don't bother to see me out (said by departing guests to host)
挽留 *wǎnliú* *VB* urge [persuade] someone to stay

b 留言 *liúyán* *VO* leave a message [comments]; *NC* short message
留念 *liúniàn* *VB* keep[accept] as a souvenir; *NC* souvenir
保留 *bǎoliú* *VB* reserve, retain; *NC* reservation

c 扣留 *kòuliú* *VB* detain, arrest
拘留 *jūliú* *VB* detain, intern

d 留心 *liúxīn* *VO* be careful, take care
留意 *liúyì* *VO* be careful, look out, keep one's eyes open

② (KEEP one's hair growing) **grow (one's hair), wear** *VB*[△]

留长发 *liú chángfà* *EH* wear long hair
留胡子 *liú húzi* *VO* grow a moustache [beard]

③ ㊷ (remain in a given condition) **STAY, remain** *VB*[△]

- b** STAY abroad to study[△]

a 大战后, 她一直留在中国。 *Dàzhàn hòu tā yīzhí liú zài Zhōngguó* *EH* She has remained in China since the war.
留级 *liújí* *VO* fail to advance in school grade
留学 *liúxué* *VO* study abroad
停留 *tíngliú* *VB* stay temporarily, stop over

b 留美 *liú Měi* *VO* study in the U.S.
留日 *liú Rì* *VO* study in Japan

Logical ordering presents the senses in a manner that makes the relation or similarity between them and the core meanings self-evident. Whenever possible this is done by letting the core meaning function as a central pivot, with the various senses clustered around it in a manner that allows them to be perceived as a logically structured whole. Though sense frequency and importance also serve as criteria in establishing sense order, logical interrelatedness often takes precedence for the sake of pedagogical efficacy.

As can be seen in Figure 1, 留 has several distinct senses, all of which are clustered around the core meanings KEEP and STAY in a manner that shows their differences and similarities. If the senses:

- KEEP for future use, reserve, leave behind
- KEEP in custody, detain
- KEEP one's mind on, concentrate on

were presented as shown below:

- reserve, leave behind
- detain
- concentrate on

they would appear to be an arbitrary list of unrelated senses, rather than as a structured and semantically integrated unit. Unfortunately, it is often not possible to cluster the single senses around the core meaning in the neat manner shown for 留.

7. Compounds and Examples





Another distinctive feature of CCLD is that compound words are grouped together *under the senses which they illustrate*. Though this makes them somewhat harder to locate, it allows learners to gain a deeper understanding of character meaning. The primary aim of the compounds and examples is to provide high-frequency, maximally useful examples for each sense, especially on the morphemic level. Unlike traditional dictionaries, compounds in which the entry character occurs non-initially are also listed, which illustrates the character's word-building function in a variety of contexts (Figure 1).





This format enables users to decompose even semantically opaque compounds into their constituents. Not only does this help the learner pinpoint the specific sense in which each entry character is used, but it also enables her to infer the meanings of compounds of a similar pattern but not listed in the dictionary; that is, it illustrates the character's morphological productivity. For example, although 留英 *liúyīng* may be opaque at first sight, the user can easily infer that it means 'study in England' from sense 3a 'STAY abroad to study'.

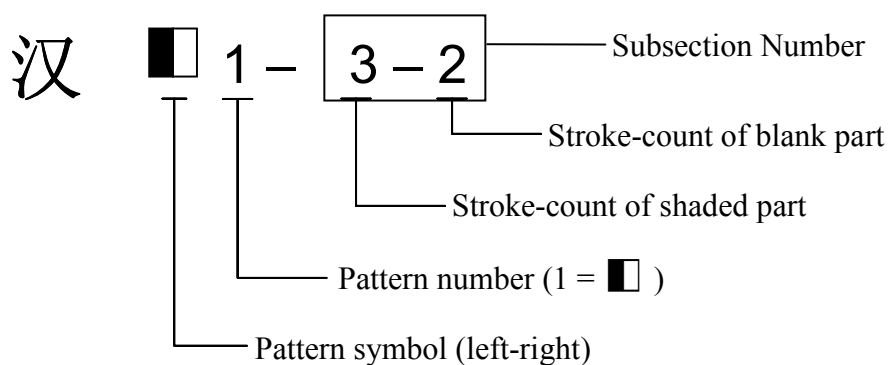
8. Innovative Lookup System



The lack of an efficient system for ordering Chinese characters has long been a source of frustration to dictionary users. Looking up characters by traditional radicals is a time-consuming, unreliable process (Paton, 2008: x). Although alternative systems have been devised, none has achieved the speed and simplicity required to meet the practical needs of the learner.

Figure 2: The SKIP Indexing Scheme

No.	Pattern	Examples
1	 LEFT-RIGHT	相 ₄₋₅ 代 ₂₋₃ 情 ₃₋₈ 街 ₃₋₉ 町 ₅₋₂ 翻 ₁₂₋₆ 髓 ₁₀₋₉ 伺 ₂₋₅
2	 UP-DOWN	示 ₁₋₄ 二 ₁₋₁ 三 ₁₋₂ 言 ₁₋₆ 系 ₁₋₆ 雀 ₄₋₇ 券 ₆₋₂ 春 ₅₋₄ 寺 ₃₋₃ 空 ₃₋₅ 文 ₂₋₂ 亭 ₂₋₇ 堯 ₂₋₆ 当 ₃₋₃ 南 ₂₋₇ 支 ₂₋₂
3	 ENCLOSURE	進 ₃₋₈ 辻 ₄₋₂ 刀 ₁₋₁ 司 ₁₋₄ 石 ₂₋₃ 考 ₄₋₂ 医 ₂₋₅ 臣 ₃₋₄ 旬 ₂₋₄ 載 ₆₋₇ 尾 ₃₋₄ 病 ₅₋₅ 肉 ₄₋₂ 凶 ₂₋₂ 回 ₃₋₃ 国 ₃₋₅
4	 SOLID	下 ₃₋₁ 耳 ₆₋₁ 雨 ₈₋₁ 子 ₃₋₁ 由 ₅₋₂ 自 ₆₋₂ 坐 ₇₋₂ 重 ₉₋₂ 中 ₄₋₃ 十 ₂₋₃ 手 ₄₋₃ 本 ₅₋₃ 由 ₅₋₂ 自 ₆₋₂ 坐 ₇₋₂ 重 ₉₋₂

A major feature of CCLD is the speed and facility with which entries can be looked up by beginners. In 1990 KIT dictionaries introduced a new scheme, called SKIP, which enables users to quickly look up characters as accurately as in alphabetical dictionaries (1990). Each character is classified under one of four visually distinct geometrical patterns:  1 left-right,  2 up-down,  3 enclosure, and  4 solid. Within each group the characters are further subdivided by stroke-count, as shown below:



For example, 汉 can be divided into left and right parts and is classified under pattern  1. It contains three strokes in the shaded part (彳) and two strokes in the blank part (又), so it appears under SKIP number  1-3-2. The main body of the dictionary is ordered according to SKIP numbers, and within each SKIP number the characters are further classified under various criteria, such as shared elements, to enable the user to quickly zoom in on the desired entry.

Since SKIP is a reliable and logically consistent system that can be learned by beginners in a short time, it has gained popularity in many kanji dictionaries, including online and electronic dictionaries. SKIP was originally invented specifically for KIT dictionaries. Its adoption in CCLD marks the first time that this system has been used in any Chinese dictionary.

9. Future Work

This paper introduced a new Chinese-English dictionary based on a pedagogically oriented editorial policy designed to satisfy the special needs of non-native learners by addressing the major shortcomings of previous works. Every effort has been made to meet those needs with a rich set of features, many not described here. These include accurate pinyin readings, traditional Chinese and Japanese character forms, study level codes (HSK), stroke order diagrams, numerous cross-references, frequency statistics, character codes, and multiple indexes to look up characters by pinyin, by SKIP pattern, or by traditional radical.

The number of learners of Chinese worldwide is said to exceed 30 million, which has led to a constantly growing demand for pedagogically effective learning aids. The CJK Dictionary Institute is dedicated to meeting this need through the ongoing development of numerous electronic dictionaries and pedagogical applications, dozens of which have already been released for both Chinese and Japanese. These include mobile versions of CCLD and KIT dictionaries, KIT spinoffs such as the German and Spanish editions, the world's most comprehensive Chinese-English electronic dictionary, applications for mastering Japanese and Chinese vocabulary, and more.

It is hoped that lexicographers and educators around the world will continue to contribute to this effort through advice, project proposals, constructive criticism and, above all, through direct collaboration.

References

New Japanese-English Character Dictionary. Tokyo: Kenkyusha, 1990.

The Kodansha Kanji Learner's Dictionary. Tokyo: Kodansha International, 1999.

HSK Test English Home Page [on line]. [Beijing: Beijing Language and Culture University]. http://www.hsk.org.cn/Index_E.aspx [Access date: June 24, 2011].

Chinese Gigaword Product Page [on line]. [Philadelphia: Linguistic Data Consortium]. <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T09> [Access date: June 24, 2011].

Dong, X. (1999). "Cong weici dao tici de zhuanhua tan hanyu cidian biao zhu cixing de biyaoxing" [Nominalization and the need for lexical categories in Chinese dictionaries]. *Lexographical Studies* 1. 24-32.

Yu, S., et al. (1998). *The Grammatical Knowledge-base of Contemporary Chinese -- A Complete Specification*. Beijing: Tsinghua University Press.

The CJK Dictionary Institute, Inc. (CJKI). "Chinese Dictionaries". Internal survey. June 19, 2011.

Xinhua Zidian dishiban [Xinhua Dictionary 10th Edition]. Beijing: The Commercial Press, 2004.

Xiandai Hanyu Cidian Hanying Shuangyu [The Contemporary Chinese Dictionary: Chinese-English Edition]. Beijing, Foreign Language Teaching and Research Press, 2002.

Hanying Zonghe Dacidian [A Comprehensive Chinese-English Dictionary]. Dalian: Dalian University of Technology Press, 2004.

Wang, J., Inhoff, A., Chen, H. (1999). *Reading Chinese Script: A Cognitive Analysis*. Mahway, New Jersey: Lawrence Erlbaum Associates.

Zhang, L. (2001). "Cidian shiyi zhong de ciyi he yusuyi" [Lexemic and morphemic meanings in dictionary definitions]. *Lexographical Studies* 2. 33-41.

Sapir, E. (1921). "The Elements of Speech". *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace. 24-41.

Paton, S. (2008). *A Dictionary of Chinese Characters: Accessed by Phonetics*. Abingdon: Routledge.