CC-CEDICT WIKI

Syntax

*These are guidelines on what CC-CEDICT entries **should** look like. CC-CEDICT still has many old entries which do not comply to these rules yet.*

Basic format

The basic format of a CC-CEDICT entry is:

```
Traditional Simplified [pin1 yin1] /English equivalent 1/equivalent 2/
```

For example:

中國 中国 `[Zhong1 guo2] /China/Middle Kingdom/`

Additionally:

- The Chinese word should consist of one or more Chinese characters, without any spaces in it
- The Mandarin pinyin should follow in the format below:
  - It should have a space between each pinyin syllable
  - Each pinyin syllable should have a tone number. Use 5 for the light tone (e.g. ni3 hao3 ma5)
  - Raw tones should be used:
    - Tone sandhi is **not** indicated (e.g., ni3 hao3 is not changed to ni2 hao3)
    - Although "yi" and "bu" have various modifications in tone, depending on what follows them, these are **not** indicated in writing (e.g., "one horse" is pronounced "yi4 pi3 ma3" but written "yi1 pi3 ma3", and "not enough" is pronounced "bu2 gou4" but written "bu4 gou4")
    - Word-related changes to neutral tone, however, **are** indicated. These are especially common with reduplicated forms (e.g., use ma1 ma5, not ma1 ma1; ba4 ba5, not ba4 ba4; kan4 kan5, not kan4 kan4; xiang3 xiang5 ("take under consideration"), not xiang3 xiang3). This isn't limited to reduplicated forms, e.g., ming2 bai5, not ming2 bai2; cong1 ming5, not cong1 ming2.
      It's best to keep in mind that Pinyin is about Mandarin words, not Chinese characters.
  - For pinyin that uses the ü, represent it with a u followed by a colon (e.g. nu:3 ren2)
  - Capitalize pinyin for proper nouns (e.g. **B**ei3 jing1)
- The English definitions should be separated with the '/' character (e.g. /English equivalent 1/equivalent 2/).
- American English should be used for the English definitions
- Do not add definite or indefinite articles (e.g. "a", "an", "the", etc) to English nouns unless they are necessary to distinguish the word from another usage type or homonym

Punctuation

Middle dot

Middle dots are often used for separating western names:
珍・奥斯汀 珍・奥斯汀 [Zhen1・Ao4 si1 ting1] /Jane Austen (1775-1817), English novelist/

A double width middle dot is used in the Chinese, a single width middle dot padded with spaces on both sides is used in the pinyin.

Comma

Commas are sometimes used in Chinese proverbs:
人為財死，鳥為食亡 人为财死，鸟为食亡 [ren2 wei4 cai2 si3 , niao3 wei4 shi2 wang2] /Human beings die in pursuit of

wealth, and birds die in pursuit of food/…/

A double width comma is used in the Chinese, a single width comma padded with spaces on both sides is used in the pinyin.

Retroflex finals

There are 3 kinds of R-ised words that use the 兒/儿 character:

1. 兒/儿 is not-optional because it's its own syllable (usually meaning "son," so daughter is actually "girl son") - 女兒/女儿 nǚ'ér
2. 兒/儿 is not-optional because it changes the definition of the word and is tacked on to the preceding syllable - 头兒/头儿 tóur (leader) as opposed to 头 tóu (head)
3. 兒/儿 is an optional northern pronunciation (er2hua4) and is tacked on to the preceding syllable - 花兒/花儿 huār (flower) as opposed to 花 huā (flower)

These 3 cases should be formatted as follows:

1. 女兒 女儿 [nu:3 er2] /daughter/
2. 頭兒 头儿 [tou2 r5] /leader/
3. 花兒 花儿 [hua1 r5] /erhua variant of 花/flower/

Please note: words ending with 'r5' (such as 'hua1 r5') are presented as a -r joined with the previous syllable (eg. 'huar1') in some dictionaries using CC-CEDICT, such as the MDBG Chinese-English dictionary [http://www.mdbg.net/chindict/chindict.php].

Taiwanese pronunciation

CC-CEDICT follows "standard Mandarin" as used in P.R.China. Mandarin as used in Taiwan sometimes has slight variations in the pronunciation, these can be listed as follows:
叔叔 叔叔 [shu1 shu5] /(informal) father's younger brother/uncle/Taiwan pr. shu2 shu5/

Taiwan doesn't use the light tone so, we do not list Taiwan pronunciations when they consist only of saying "don't use the light tone". When a character has a "Taiwan pr." notice, then all of its compound need not mention it.

General principles

Various trivial style things:

- Don't use parts of speech. Instead try to give an indication of grammatical usage within the English definition. CC-CEDICT is a human readable descriptive dictionary, not a resource intended for machine processing.
- Abbreviations etc cf e.g. i.e. do not need any further punctuation.
- Extended meanings indicated by lit. .. fig. combination when appropriate or when a common expression refers back to a classical incident or chengyu, one can refer to it with cf (incident in Records of the Historian).

Choice of entries and translations

The current CC-CEDICT database contains a considerable number of infelicities, inaccuracies, omissions, and actual errors. As an ideal, new entries should be checked against 2 or 3 different sources (e.g. the online and paper dictionaries). Care is needed, since the dictionaries copy from one another – an entirely bogus entry in CC-CEDICT is copied uncritically onto thousands of websites within a few months.

A Chinese word for which a Google query with the following syntax results in many thousand of hits should probably be added to CC-CEDICT, with translations corresponding to the main usages.

```
+"combination of characters"
```

(the +"" combination forces Google to match both a whole word and to ignore variants)

General principles of translation

The English should be meaningful, not horribly ugly, and bear a close relation to the Chinese meaning. It should correspond to something that could be used naturally by an English speaker (I think Arthur Waley has some advice saying that just because a text is about magnetohydrodynamics, it doesn't follow that it has to be horribly ugly).

On the other hand, a translation always loses something, and the translator can compensate by substituting an English equivalent (e.g. a biblical or Shakespearian allusion in place of a Confucian idiom).

Names of persons should say dates if possible (birth, death, years in which the person was active in a certain role, etc), what interest the person has (writer, general, pop star, etc), brief indications of CV (e.g. took part in a revolution, was murdered, wrote famous book, etc). For example:
胡錦濤 胡锦涛 [Hu2 Jin3 tao1] /Hu Jintao (1942-), president of PRC from 2003/

Names of plants, animals, musical instruments should give common name and scientific name when appropriate; there is a particular problem of how specific the word is – a plant may mean a minor variety within a species, or may refer to an entire taxonomic family. Different writers will use it to mean the common family, or the particular item of salad on their plate at present.

Most words have more than one meaning, and more than one grammatical function. Care is needed not to concentrate only on a specific occurrence to the exclusion of others. e.g . the actual occurrence may be a verb in the past participle (say “overthrown”) whereas the word may also mean “destruction”, “to topple” etc.

There are 20,000 Chinese characters in the more advanced dictionaries, of which many are obscure, never used, and will not have correct definitions in online or paper dictionaries. This is the boundary of knowledge. (Exactly the same applies to big English dictionaries.) These obscure characters appear on modern websites, and one sometimes needs to give a definition. It is reasonable to admit (precise meaning unknown), and give an indication of what one can deduce.

Ambiguity due to homonyms

Sometimes words used in the English definitions can have multiple meanings. If the Chinese word does not have these additional meanings, additional information should be provided to prevent ambiguity:
首都 首都 [shou3 du1] /capital (city)/

The text between the parentheses is “meta-information”; it is not a direct part of the translation, merely to prevent ambiguity.

References

The English definitions can contain references to other Chinese words. These should be noted as follows:
漢字|汉字[Han4 zi4]

For example:
股指 股指 [gu3 zhi3] /stock market index/share price index/abbr. for 股票指數|股票指数[gu3 piao4 zhi3 shu4]/

Classifiers

Classifiers (also called “Measure words”) can be listed using the following syntax:
避風港 避风港 [bi4 feng1 gang3] /haven/refuge/harbor/CL:座[zuo4],個|个[ge4]/

Classifiers follow the 'reference' syntax, are prefixed by 'CL:' and separated by a comma (no additional spacing).

The classifier words itself can be described using:
/classifier for small round things (peas, bullets, peanuts, pills, grains etc)/

Variants

Many characters have variants, sometimes more than one, sometimes with identical meaning or quite different meanings. Some choice of variants found in texts on websites will arise because of the different input methods, and the user may have had no intention of using the variant.

You can get rough usage frequency information by searching the alternative word forms in Google. Please use this syntax to make sure that Google doesn't perform any automatic variant translations:

```
+"word"
```

Additionally you can use Google's advanced search to specify the language to either 'Chinese (Traditional)' or 'Chinese (Simplified)' to prevent Japanese web pages from influencing the results. For example:
789 Chinese (Traditional) pages for +“撐竿跳高”
17,700 Chinese (Simplified) pages for +“撑竿跳高”
1,750 Chinese (Traditional) pages for +“撐杆跳高”
66,900 Chinese (Simplified) pages for +“撑杆跳高”

It often happens that Google tells you that +“Xx” occurs 200 times more frequently than +“XX”, in which case Xx should be in CC-CEDICT as a regular entry, and XX only as “XX XX [pin1 yin1] /variant of Xx/definition/”.

When there are alternative forms of the same expression, and the less common form is at most 5 times less common, the less common entry should have /also written ../ referring to the more common form, e.g. 撐竿跳高 撑竿跳高 [cheng1 gan1 tiao4 gao1] /pole-vaulting/also written 撐杆跳高|撑杆跳高/.

Romanization of foreign languages

When transcribing foreign words in definitions, please use the following romanization methods:

- Japanese: Modified Hepburn [http://en.wikipedia.org/wiki/Hepburn_romanization]
- Korean: Revised Romanization of Korean [http://en.wikipedia.org/wiki/Revised_Romanization_of_Korean]

If an alternative romanization method is more popular for a certain word, that version can be added as an additional translation.

---

format/syntax.txt · Last modified: 2012/01/28 23:00 (external edit)