日中韓辭典研究所
The CJK Dictionary Institute

Hanzi-to-Pinyin/Zhuyin Converter (H2X)

Overview and Goals

H2X is a Hanzi-to-Pinyin (H2P) conversion system for Simplified Chinese and a Hanzi-to-Zhuyin (H2Z) conversion system for Traditional Chinese. It can also be expanded to other romanization systems, such as Yale and Wade-Giles. Collectively, we will refer to such a hanzi transcription system as "H2X," where *X* stands for any phonemic transcription such as pinyin, zhuyin or romanized Cantonese. Below, *X* will often be referred to by the term *reading*.

H2X can be used, among others uses, to:

- aid native speakers in reading difficult names or characters
- aid learners to read Chinese texts
- enable ambiguous search based on homophones (explained below)
- sort hanzi by pinyin or zhuyin (useful for name lists and the like)

Conversion Ambiguity

An obvious and major issue with H2X is the one-to-many ambiguity of thousands of characters, the so called *polyphonic* hanzi (多音字 *duōyīnzì*), such as *lè* and *yuè* for 乐, resulting in numerous homophones. The disambiguation strategy for accurate H2X conversion is to tokenize the text so as to isolate individual words, then to look up in word-level hanzi mapping tables, which almost completely eliminates ambiguity. This requires the following components

1. Simple word tokenizer
2. Word-level H2X mapping tables
3. Character level H2X mapping tables

There are two kinds of homophonic ambiguity (the implications of which are described below):

1. ***Homotonic***: reading and tone are identical, such a 网陆 (resulting from input errors) and 网路, both **wǎnglù.**
2. ***Heterotonic***: reading is identical but tone different, such as 网炉 **wǎnglú** (input error) and 网路 **wǎnglù.**

Note that for the purposes of converting to the correct reading, the tokenizer need not be as robust as for other applications since the goal is not to extract tokens per se, but to segment just accurately enough so that the correct reading is determined. Thus the H2X tokenizer can be based on a simplified tokenization algorithm, which CJKI can provide, independently of the main tokenizer.

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

Ambiguous Search

H2X conversion on both the homotonic and heterotonic levels can have a major benefit: enabling ambiguous search as well as retrieval of documents even if the keywords are input erroneously, such as 网陆 for 网路. The system should thus support four conversion modes:

1. *Toneless* pinyin for Simplified Chinese
2. *Toned* pinyin for Simplified Chinese
3. *Toneless* zhuyin for Traditional Chinese
4. *Toned* zhuyin for Traditional Chinese

This means that if the search engine is properly tuned it will retrieve not only homotonic homophone pairs like 网路/网陆 **wǎnglù,** but also heterotonic pairs like 网路/网炉 (**wǎnglù** vs. **wǎnglú**), in which the tones are different but the readings are identical.

Features of H2X Converter

The system should eventually have the following capabilities/features:

1. The source string is first extracted by the (simple) tokenizer.
2. The string is looked up in a comprehensive word-level H2X mapping table covering general vocabulary, proper nouns and technical terms.
3. Support both query errors and document errors.
4. Character level H2X mapping tables. The readings (pinyin or zhuyin) have been proofread and fine tuned over the years and include the following features:
    a. The first reading has been carefully selected to ensure it is the most common.
    b. The order of the other readings in the case of one-to-many mappings is based on frequency of use.
    c. Rarity flags enable selecting a mode in which rare and historical readings are ignored so as to reduce ambiguity (at the slight risk of error).
    d. Possibly, provide flags to indicate order of priority when a reading is used in names as opposed to general vocabulary.
    e. A major feature is that SC readings are clearly distinguished from TC readings when they are heterotonic, e.g. SC **qī** as opposed to TC ㄑㄧ (zhuyin for **qí**) for 期. See details at: http://www.cjk.org/cjk/samples/chinpin.htm.

5. The H2X conversion algorithm that (eventually) supports the following features:
    a. word level conversion
    b. character level conversion
    c. picklist of candidates in case of one-to-many ambiguities
    d. option to ignore rare/historical readings
    e. possible option to fine tune output to proper nouns
    f. select SC or TC reading
    g. output in zhuyin

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

h. output in any romanization system, such as Wade-Giles and Yale.
i. output in IPA broad transcription.

Resources for H2X Conversion

CJKI can provide the following comprehensive mapping tables and robust algorithm for H2X conversion:

1. SC-to-pinyin word mapping table
2. TC-to-zhuyin word mapping table
3. SC/TC to/from pinyin/zhuyin character mapping table
4. H2X conversion algorithm with advanced options

For your reference, if in the future you wish to support Hanzi-to-Cantonese conversion, we can also provide the following:

1. Hanzi-to-Cantonese mapping table
2. Mapping table for the eight Cantonese romanization systems
3. Hanzi-to-Cantonese conversion algorithm

# # #

About The CJK Dictionary Institute, Inc.

The CJK Dictionary Institute, Inc. (CJKI) specializes in CJK (Chinese, Japanese and Korean) and Arabic lexicography. CJKI is headed by Jack Halpern, editor-in-chief of the *The Kodansha Kanji Learner's Dictionary* and various other dictionaries that have become standard reference works for studying Japanese. CJKI is one of the world's prime sources for CJK and Arabic lexical resources, and is contributing to CJKA information processing technology with its high-quality lexical resources.

Description	Japanese	Chinese
General vocabulary	350,000	500,000
Companies and organizations	600,000	55,000
Personal names	2,620,000	243,000
Place names	444,000	170,000
Personal name variants	3,500,000	6,000,000
Technical terminology	1,800,000	5,500,000
Orthographic variants	80,000	-
Bilingual English	320,000	800,000
Others	72,000	160,000
Total	**9,786,000**	**13,428,000**

Japanese and Chinese Data Coverage at a Glance

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