

CHDICT Data Format

Table of contents

- CHDICT Data Format1
- Table of contents1
- Versions1
- Source format2
- References.....2
- Considerations2
- XML DTD.....2
- Explanation.....2
- Parts of speech3
- References.....3
- Considerations3
- Annotation.....4
- Definition of parts of speech.....4
- Regions.....4
- Styles.....4
- Fields4
- References.....4
- Definition of parts of fields.....5
- Open issues.....5
- Regions and styles.....5
- More annotation for expressions.....5
- Examples stored in a separate resource5

Versions

Version	Date	Description
1.0	December 28, 2006	Initial version

Source format

References

1. CEDICT database: <http://www.mdbg.net/cedictwiki/start>
No hierarchical structure
2. JMDict DTD: http://www.csse.monash.edu.au/~jwb/jmdict_dtd_h.html
Complex DTD

Considerations

CHDICT aims for a compromise between structured representation and ease of processing. The two fundamental decisions are: a) to store dictionary data in XML format; b) to keep the dictionary bilingual, with Chinese distinguished as the source language, and a small bias towards simplified writing.

XML DTD

```
<?xml version="1.0"?>
<!-- Version 1.0 -->
<!DOCTYPE CHDICT [
<!ELEMENT dict (entry*)>
<!ELEMENT entry (id, status, hanzi+, pinyin, cnf, sense+)>
<!ELEMENT id (#PCDATA)>
<!ELEMENT status (#PCDATA)>
<!ELEMENT hanzi (#PCDATA)>
<!ATTLIST hanzi var (trad|simp)>
<!ELEMENT pinyin (#PCDATA)>
<!ELEMENT cnf (#PCDATA)>
<!ELEMENT sense (pos, region?, field?, style?, meas?,
(gloss|expl)+, ant*, syn*, xmp*, xpr*)>
<!ELEMENT pos (#PCDATA)>
<!ELEMENT region (#PCDATA)>
<!ELEMENT field (#PCDATA)>
<!ELEMENT style (#PCDATA)>
<!ELEMENT meas (#PCDATA)>
<!ELEMENT gloss (#PCDATA)>
<!ELEMENT expl (#PCDATA)>
<!ELEMENT ant (#PCDATA)>
<!ELEMENT syn (#PCDATA)>
<!ELEMENT xmp (hanzi, trans)>
<!ELEMENT trans (#PCDATA)>
<!ELEMENT xpr (hanzi+, pinyin?, (gloss|expl)+)>
]>
```

Explanation

id	A unique integer linking entries in the engine sources to database entries under revision.
status	Status of the entry in the online system. Four entities are defined: &st_none; is not used in CHDICT &st_approved; for entries approved by an editor &st_edited; for entries modified by a user and not yet approved by an editor &st_unrevised; for automatically generated and not yet edited/approved entries
hanzi	Every entry must contain two hanzi variants, indicated by the var attribute as traditional or simplified. Both variants are always listed even if they are identical. Neither is necessarily a unique key in the dictionary, see different pinyin readings.
pinyin	Tones indicated by numbers 1-5; neutral always indicated; syllables separated with spaces. "ü" is represented as "u:". E.g., "nu:3 er2". No distinction is made between

	upper and lower case. Different readings of the same character/word are listed as different entries.
cnf	Combined normal frequency calculated from character frequencies, used for ranking lookup results.
sense	Different meanings as well as different Chinese PoS readings are listed as separate senses.
pos	Part of speech; must take a value from the PoS entities.
region	Optional: region-specific reading. Must take a value from the region entities.
field	Optional: field-specific reading. Must take a value from the field entities.
meas	Optional: space-separated list of applicable measure words (simplified).
gloss	One translation of the sense.
expl	Explanation section (rendered in different type when displaying the entry)
ant	Optional: space-separated list of antonyms (simplified).
syn	Optional: space-separated list of synonyms (simplified).
xmp	An example sentence (simplified) with the headword and the sentence's translation.
trans	Translation as plain text.
xpr	Multi-word expression with the headword. Has a very similar structure to the main sense, but PoS, metadata (region & field) and semantics (synonyms and antonyms) are not indicated. Only a single hanzi variant (simplified or traditional) is required, and pinyin is optional.

Parts of speech

References

1. CEDICT database: <http://www.mdbg.net/cedictwiki/start>
PoS information is randomly indicated with inconsequent notation
2. HanDeDict Chinese-German dictionary: http://www.chinaboard.de/chinesisch_deutsch.php
PoS categories available in the new entry editing interface
3. The CJK Dictionary Institute: <http://www.cjk.org/cjk/samples/chinpos.htm>
Sophisticated annotation policy
4. Praktisches Chinesisch, *Kommerzieller Verlag, 2004 Beijing*
Word lists contain annotations
5. Magyar-kínai, kínai-magyar kisszótár, Óri Sándor, *Kossuth Kiadó, 2004 Budapest*

Considerations

Sources vary greatly in their depth of annotation, from virtually non-existent (1 and 5) to very specific (3). CHDICT aims for a compromise that human editors can navigate easily.

The naïve assumption that Chinese makes no distinctions between parts of speech, or conflates categories such as adjectives and verbs, is discarded. At the same time, the discussion on whether prepositions exist as a separate category or are in fact a special class of verbs is ignored.

The following compromises are made regarding parts of speech:

1. **Detail vs. simplicity.** The fundamental decision is including PoS information as such. However, various types of proper nouns are not distinguished on the PoS level; verbs are not sub-classified (transitive/intransitive etc).
2. **Categories suited to Chinese vs. Latin conventions.** Chinese is distinguished as the source language, therefore PoS information always applies to the Chinese form. Measure words and particles are distinguished. Although the existence of prepositions can be argued against, they are retained for (Latin) convention.

Key decisions regarding entry structure:

1. **Senses, glosses and explanations.** Not only are senses indicated structurally but also glosses and explanation sections within a single sense. The aim is to enhance the precision of

retrieval and to separate “translations” of a word from “metatext” when something is stated about the word or its translation.

2. **Expressions.** The CEDICT data contains many headwords that have a clear internal grammatic structure, e.g. V+Obj (打电话), V+Compl (记住) etc. CHDICT lists these under senses, and part of the editorial policy has to do with “where to list what.”

Annotation

PoS categorization applies to senses individually. If the same word can function in different roles (e.g., 帮助 as “help” or “to help”), separate senses are listed with different PoS.

Editing policy in the Hungarian glosses, with respect to the PoS of the Hungarian translation:

- Nouns: Except for pluraliatanta, singular nominal is used. Unless idiomatically required or needed to make a semantic distinction, no definite/indefinite article is used.
- Verbs: Third person, singular, present tense, indefinite object is used except in multiword lexemes requiring otherwise.
- Affixes: when the Chinese meaning is expressed by an affix, all variants go to the gloss (“-val/-vel”, “leg-”). Such cases always require an explanation.

Definition of parts of speech

Entity	Part of speech	中文	O/C	Note
&pos-adj;	adjective	形	open	
&pos-adv;	adverb	副	open	
&pos-conj;	conjunction	连	closed	
&pos-int;	interjection	叹	closed	
&pos-meas;	measure word	量	semi	
&pos-n;	noun	名	open	
&pos-part;	particle	助	closed	
&pos-prop;	proper noun	专	open	
&pos-prep;	preposition	介	closed	Includes “postpositions” such as 上
&pos-pro;	pronoun	代	closed	
&pos-v;	verb	动	open	
&pos-x;	other	别	semi	

Regions

In progress.

Styles

In progress.

Fields

References

1. HanDeDict Chinese-German dictionary: http://www.chinaboard.de/chinesisch_deutsch.php
“Area” categories available in the new entry editing interface

Definition of parts of fields

Entity	Field
&fld-arch;	architecture
&fld-art;	fine arts
&fld-astr;	astronomy
&fld-bio;	biology
&fld-buddh;	buddhism
&fld-chem;	chemistry
&fld-comp;	computers
&fld-eco;	economics
&fld-surn;	surname
&fld-food;	food and drink
&fld-geog;	geography
&fld-geol;	geology
&fld-givn;	given name
&fld-hist;	history
&fld-law;	law
&fld-ling;	linguistics
&fld-lit;	literature
&fld-math;	mathematics
&fld-med;	medicine
&fld-met;	metereology
&fld-mus;	music
&fld-org;	organization
&fld-pers;	person
&fld-phil;	philosophy
&fld-phys;	physics
&fld-pol;	politics
&fld-prov;	proverb
&fld-psy;	psychology
&fld-relig;	religion
&fld-sport;	sports
&fld-tech;	technology

Open issues

Regions and styles

More research is needed to establish the list of regions and fields indicated in the dictionary.

More annotation for expressions

Adding region, style and field to expressions is being considered.

Examples stored in a separate resource

Currently, examples are listed under senses directly. JMDict has moved examples into an independent resource, only linking then to the dictionary through ID's. A common resource of

Chinese example sentences could be created with translations in English/German/Hungarian/other languages.