Hanja Hangul Converter 0.0.5

1 License

Hanja Hangul Convert Project Homepage: http://kldp.net/projects/hanja/

Project Leader: Sung-il Kim (Masoris, masoris@gmail.com)

2 Describe Files

1. hanja.py: Simple Hanja Hangul Converter, coded by python.
2. hanconv.py: Hanja Hangul Converter Module
3. README.odt and README.pdf: This document what you read now.
4. dic0.txt: List to convert CJK Compatibility Forms to Han unification. This list do same thing with Unicode normalization algorithm.
5. dic1.txt: Hanja character list, which didn't use *duuembeobchik (두음법칙)*. In Korean Language in South Korea, when read sino-korean there are rule apply to first sound to make easy to pronounce. This rule is don't use in North Korea. You can download original hanja-hangul.ods file. (Han unification)
6. dic2.txt : Represent Phonetic Data in Unihan. (Han unification)
7. dic3.txt: Represent Phonetic Data in Unihan. (CJK Compatibility Forms)
8. dic4.txt: List of Sino-korean. List extracted form libhangul-0.0.4 and converted to use Unicode normalization algorithm. (Han unification)
9. dic5.txt: Exception list for covert Hanja to Hangul. Read 3.3 for detail. (Han unification)
10. dic6.txt: (experiment) Database to convert Hangul to Hanja.

3 How to use this database

3.1 You need to know first

These days there are two way to use Hanja by Unicode. The first way is to use both Han unification and CJK Compatibility Forms. This way used most time in Korean daily life. MS Word and Hangul word process support this way. In this case, The CJK Compatibility Forms are used for indicate pronunciation of Hanja. Most Hanja character in Korean has just one pronunciation, but some character are not (most times because of *duuembeobchik 두음법칙)*. So Hanja characters which have plural pronunciation are mapped plural code in CJK Compatibility Forms. So it makes easy to convert to Hangul.

But Unicode doesn't recommends to map a character to plural codes, and there are tool which names 'Unicode normalization algorithm' to make all each character map one code. So they don’t use CJK Compatibility Forms. Some web site such as Wikipedia support this way.

Therefore someone convert Hanja data in Wikipedia to Hangul by MS Word, it didn't converted correctly. To convert those Hanja data correctly, it needs Sino-korean dictionary, the file 'dic4.txt' will works for this.

In North Korean there are no *duuembeobchik* not likes in South Korean. So you only need 'dic1.txt' which didn't use the rule.

3.2 Convert Hanja text which use only Han unification

To convert Hanja text which only use Han unification (Hanja which converted by Unicode normalization algorithm likes **in Wikipedia**), Because of some Hanja character pronounced different way by words, so it needs Sino-korean dictionary 'dic4.txt'. And the problem is left characters which can't convert to use 'dic4.txt', those character doesn't have phonetic information in character, the recommend way is it convert to the sound which most use. So the recommended way is to convert those properties.

1. dic4.txt
2. dic5.txt
3. dic1.txt

When someone demands convert to this way, there are no confidence that, the all texts written by Han unification, so I recommend that convert Hanja first to use ‘dic0.txt’; it makes all Hanja to Han unification, then convert to Hangul.

'dic5.txt' makes result naturally, because of 'dic1.txt' doesn't contain *duuembeobchik* or other phonetic reason. 'dic5.txt' contains surname of Koreans and character which always use *duuembeobchik*.

3.3 Convert Hanja text which uses both Han unification and CJK Compatibility Forms.

To convert Hanja which use both Han unification and CJK Compatibility Forms like **MS Word** and Hangul word process do. It's simple. Just convert to use these databases. Those contain Unihan Represent Phonetic Data.

1. dic2.txt
2. dic3.txt

3.4 Convert Hanja text to Hangul (without duuembeobchik)

1. Covert Hanja to Han unification, to use 'dic0.txt'
2. And Covert to Hangul, to use 'dic1.txt'

4 hanja.py

'hanja.py' is python script to convert Hanja to Hangul.

4.1 Executive

To run 'hanja.py', do this in Linux terminal;

- $ chmod a+x hanja.py   // Give permission to hanja.py file.
- $ ./hanja.py   // And run it.

Or, You can run it by python interpreter;

- $ python hanja.py

4.2 Convert in Input Mode

To convert Hanja to Hangul in Input Mode is very simple.

```
$ ./hanja.py  #  First, Run python script.
Hanja Hangul Converter0.0.4    by Sung-il KIM (masoris@gmail.com)
Commands: exit(종료), mode(방식), reverse(역변환), list(목록)
Type Hanja to convert and press enter

Choose mode to convert Hanja to Hangul
1. Han unification only (Wikipedia, default)
2. Han unification and CJK Compatibility Forms (MS Word)
3. Without duuembeobchik (North Korean)
4. Apply Unicode normalization algorithm
5. (experiment) Compatible convert with both ways (1,2)
6. (experiment) Convert Hangul to Hanja
擇> 3  # Choose mode what you want. I choose 'With out duuembeobchik (North Korean)'in here.
Load dic1.txt file, 27496 of indexes
Total 27496 of indexes

入1> 韓國의 歷史  # Just input Hanja, what you convert, and press enter.
出1> 한국의 력사  # So computer convert the Hanja to Hangul directly. The result is right, the pronunciation of 歷史 is 력사 in North Korean.

入2> mode  # If you want change converting mode, Type 'mode' or '방식' and press enter.
Choose mode to convert Hanja to Hangul
1. Han unification only (Wikipedia, default)
```

```
2. Han unification and CJK Compatibility Forms (MS Word)
3. Without duuembeobchik (North Korean)
4. Apply Unicode normalization algorithm
5. (experiment) Compatible convert with both ways (1, 2)
6. (experiment) Convert Hangul to Hanja
擇> 1  # I choose '1. Han unification only (Wikipedia, default)', You can choose '1' by Just Enter
without type '1'. If you are type nothing when choose mode, it selected '1' default.
Load dic4.txt file, 32237 of indexes
Load dic5.txt file, 10 of indexes
Load dic1.txt file, 27496 of indexes
Total 59743 of indexes

Hanja Hangul Converter0.0.4    by Sung-il KIM (masoris@gmail.com)
Commands: exit(종료), mode(방식), reverse(역변환), list(목록)
Type Hanja to convert and press enter

入3> 韓國의 歷史  # And I typed same Hanja again.
出3> 한국의 역사  # The result is not same with before, because the pronunciation of '歷史' in not
same between South Korea and North Korea, it's '력사' in North Korean, and '역사' in South Korean.
Therefore the result '한국의 역사' is right, because I choose mode1.

入4> reverse  # The command 'reverse' or '역변환' make database reverse. It's experiment function
for test database. So do not use this function to convert Hangul to Hanja.
Convert Hangul to Hanja (Reverse)

入5> 한국의 역사
出5> 韓國  力士

入6> list  #  If you tyed 'list' or '목록', you can see all lists in database, what using now.
 ( ... )

入7> exit  # To exit programmer, just type 'exit' or '종료'.
```

4.3 *Convert in Terminal*

You can also convert text file, in type command in terminal.

1. $ ./hanja.py [File name to convert or String] # It will convert file or string by default mode, and print result on screen.
2. $ ./hanja.py [File name to convert or String] [Mode] # It will convert file or string by select

mode, and print result on screen.

3. $ ./hanja.py [File name to convert or String] [Mode] [File name to save result] # It will convert file or string by select mode, and save the result to file.

5 hanconv.py

This file is module for convert hanja to hangul.

5.1 *Available Functions*

5.1.1 convert([text], [mode], [reverse]) -> str

This function convert text hanja to hangul.

You can use those modes:

- 'unionly', (1): Han unification only (Wikipedia, default)
- 'uniandcomp', (2): Han unification and CJK Compatibility Forms (MS Word
- 'withoutduuem', (3): Without duuembeobchik (North Korean)
- 'uninormal', (4): Apply Unicode normalization algorithm
- 'compboth', (5): (experiment) Compatible convert with both ways (1, 2)
- 'hangul2hanja', (6): (experiment) Convert Hangul to Hanja

The numbers in parenthesis could be changed in version up.

5.1.2 getlenoflistfrommode(mode)

This function returns number of indexes in mode.

5.1.3 printlistfrommode(mode)

This function print all index list in mode.

5.2 *Example*

```
$ python  #Run python on hanja hangul converter directory
Python 2.4.4c1 (#2, Oct 11 2006, 21:51:02)
[GCC 4.1.2 20060928 (prerelease) (Ubuntu 4.1.1-13ubuntu5)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> import hanconv  #Import hanconv module
>>> print hanconv.convert('歷史')
```

```
역사
>>> print hanconv.convert('歷史', 3)
력사
>>> print hanconv.convert('역사', 1, True)
力士
>>> print hanconv.getlenoflistfrommode(1)
60019
```