Chinese Utilities Help

July 2016

Introduction

This package contains utilities for working with Chinese Hanzi and Hanyu Pinyin in Fieldworks. Converters are provided for generating:

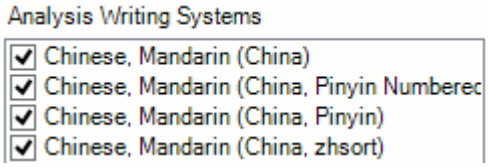
- numbered Pinyin from Chinese Hanzi (characters)
- diacritic Pinyin from numbered Pinyin
- a sort field to correctly sort a reversal index

The sort index and Pinyin formatting follow the conventions used in 现代汉语词典[1].

Fieldworks Configuration

Writing Systems

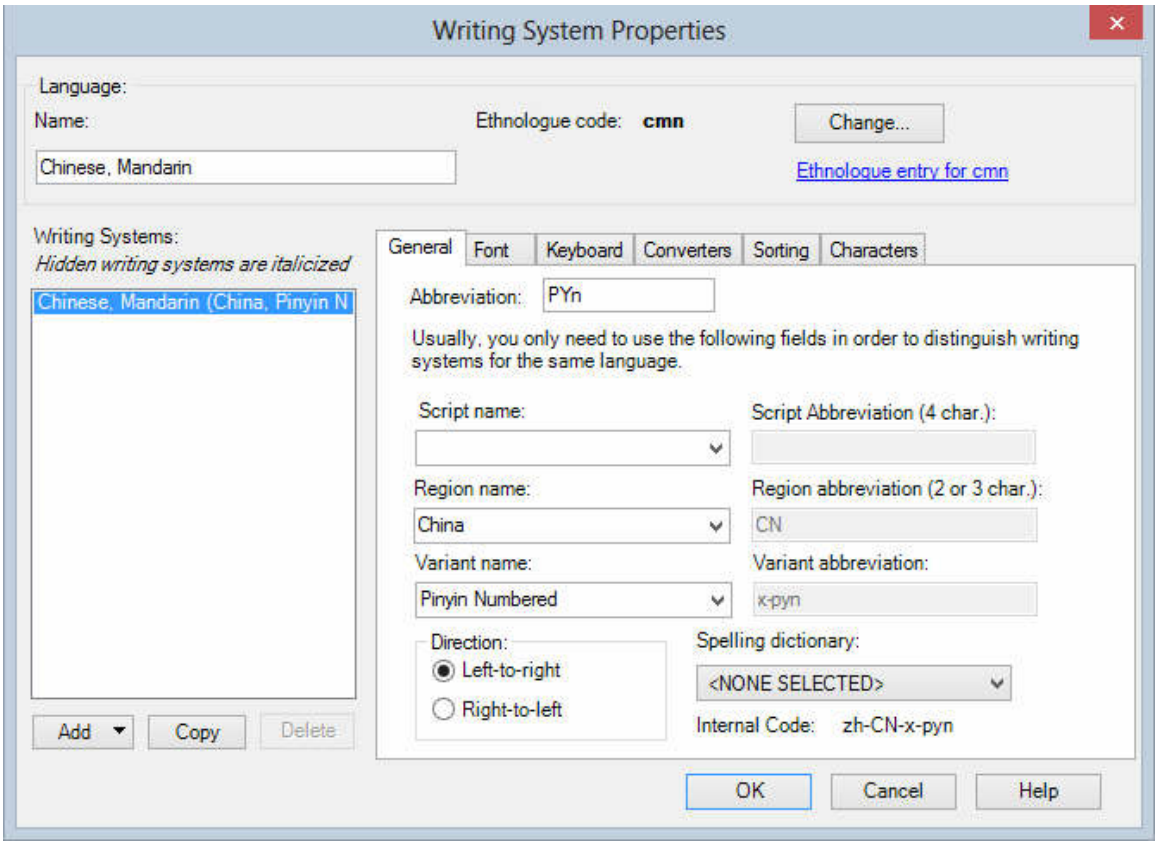
In Fieldworks configure the following Chinese analysis writing systems as seen in the Writing Systems tab of the Project Properties dialog box:



1. Click Add and select "Chinese, Mandarin (China)"
2. Click Add again and select each of the Chinese, Mandarin writing systems exactly as shown above.

[1] (Xian Dai Han Yu Ci Dian) The Contemporary Chinese Dictionary, Foreign Language Teaching and Research Press, 2002.

Click the Modify button to see the Writing Systems Properties dialog:



Notice the *Internal Code* in the bottom right corner. It should match the following table[2]:

Abbreviation	*Variant name*	*Internal Code*
Chn		zh-CN
Pyn	Pinyin Numbered	zh-CN-x-pyn
PY	Pinyin	zh-CN-x-py
HZs	zhsort	zh-CN-x-zhsort

It is the Internal Codes that are critical since it is those exact names (except for capitalisation and '-'/'_' differences) that are used by the FLExTools modules.

Reversal Index

To create a Chinese reversal index go to the Lexicon | Reversal Indexes area and use the menu Insert | Reversal Index. In the dialog select, "Chinese (China)." (Note: from Fieldworks 8.3 reversal indexes are automatically created.)

[2] For compatibility with pre-FW7 projects the following codes are also supported: cmn, cmn__X_PYN, cmn__X_PY, cmn__ZHSORT

Starting with Fieldworks 7.1 any sorting configured in the Bulk Edit Reversal Indexes area will be applied to the Reversal Index view. So, when outputting the reversal index to XHTML (for Pathway) you need to configure the Bulk Edit Reversal Indexes to be sorted by the *'Chinese, Mandarin (China, zhsort)'* field before exporting.

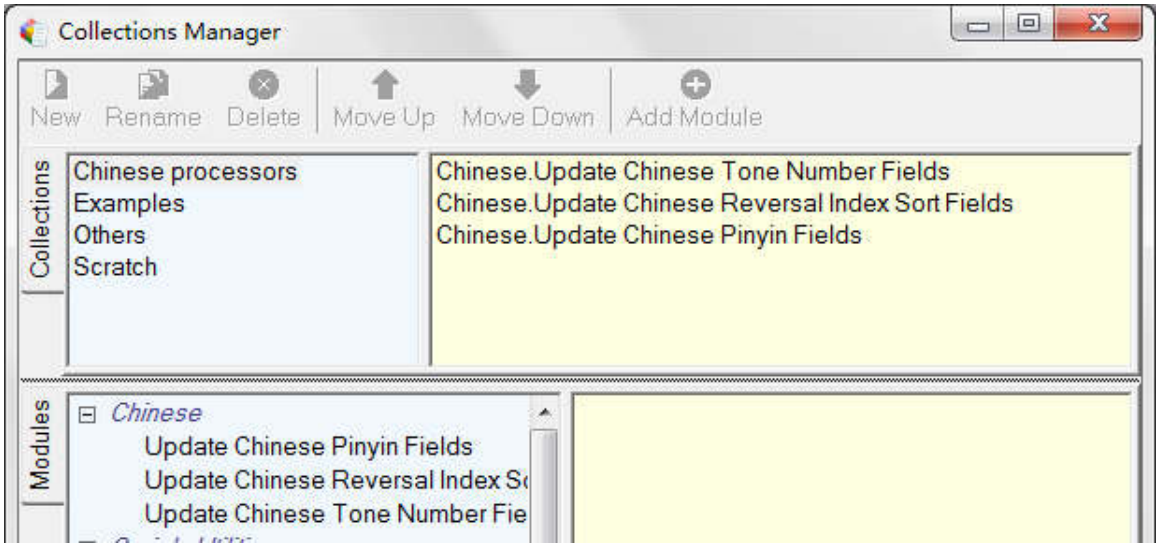
The sort field can be omitted from the dictionary output using the configuration dialog (Tools | Configure | Dictionary.)

FLExTools Configuration

Setup

Install FLExTools if necessary and install this package in the Modules folder.

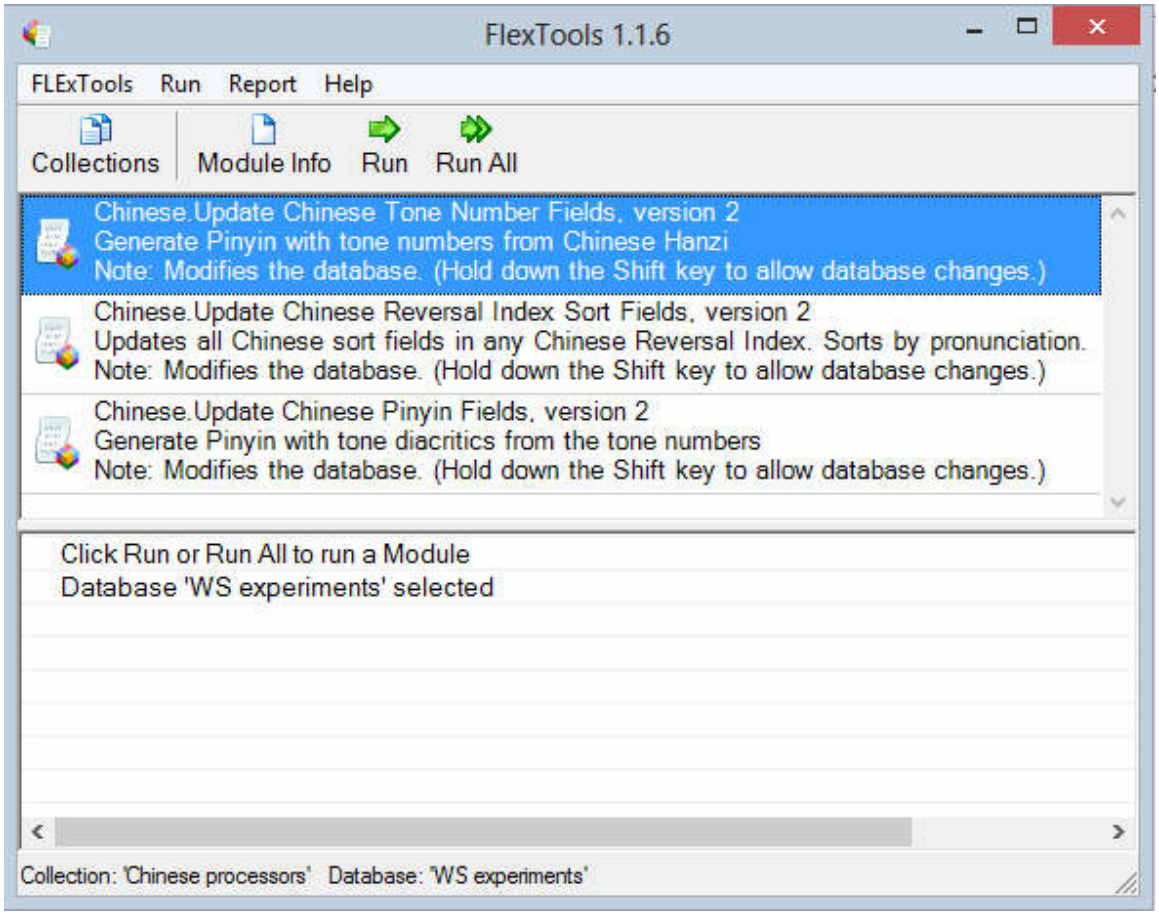
1. In FLExTools create a new module collection by clicking the Collections toolbar button, then clicking New in the Collections Manager dialog.
2. Type a name for the collection, e.g. "Chinese processors"
3. Add the three Chinese modules and order them as shown here:



4. Click the dialog's close button.

Usage

The main FLExTools window should now look like this:



Select your Fieldworks database with the menu FLExTools | Choose Database. Use the Run or Run All buttons to run one module at a time, or to run them all as a batch process.

FLExTools allows you to do a 'dry run' to see what changes will be made before committing to the changes (like the Preview and Apply options in Fieldworks' Bulk Edit tools.) By default, clicking Run will not make changes to the database. When you are happy with the operation, hold down the Shift key while clicking Run and the database will be updated as necessary.

NOTE:

- Please backup your database before making any changes with FLExTools.
- The Fieldworks database has to be closed when running FLExTools.

The next section explains what these modules do.

FLExTools Modules

Update Chinese Tonenumber Fields

This module populates the *Chinese, Mandarin (China, Pinyin Numbered)* field from the *Chinese, Mandarin (China)* field for:

- all glosses in the lexicon;
- all forms in the Chinese Reversal Index.

Punctuation is included in the tone number field to match the formatting in 现代汉语词典. E.g. 'lu4//yin1', 'jiao3.zi5'.

Ambiguities in the Pinyin are included as a list of possibilities separated by a vertical bar '|', e.g. 'zhong1|zhong4'.

If the tone number is already present then it is checked against the Chinese and any inconsistencies reported (e.g. if the Chinese has been changed without updating the tone number.) However the tone number field will not be overwritten to avoid losing any manual edits.

After running this module use a filter in Fieldworks to find all the ambiguities (search for '|') and manually edit the tone number field.

Update Chinese Reversal Index Sort Fields

This module populates the sort field *Chinese, Mandarin (China, zhsort)* in the Chinese Reversal Index. The sort field is generated from the 汉字 *Chinese, Mandarin (China)* field **and** the Pinyin with tone numbers (*Chinese, Mandarin (China, Pinyin Numbered)*) field. Any entries with an ambiguous tone number field will not be updated.

The sort field produced by this Module orders Chinese by pronunciation, then by stroke count, and finally by stroke order. This follows the

现代汉语词典. Thus:

- san < sen < shan < sheng < si < song (三 < 森 < 山 < 生 < 四 < 送)
- lu < lü < luan < lüe (路 < 绿 < 乱 < 掠)
- (stroke count) 录 < 录音 < 路 < 路口
- (stroke order) zhi4 with 8 strokes: 郅 < 制 < 质 < 治

Generate Reversal Sort Field Only

For cases where Pinyin isn't needed in the Lexion (glosses), this module updates the tone number field and the sort field in the Chinese Reversal Index only. The tone numbers and sort strings are the same as the modules above.

Update Chinese Pinyin Fields

This module populates the *Chinese, Mandarin (China, Pinyin)* field from the *Chinese, Mandarin (China, Pinyin Numbered)* field for:

- all glosses in the lexicon;
- all forms in the Chinese Reversal Index.

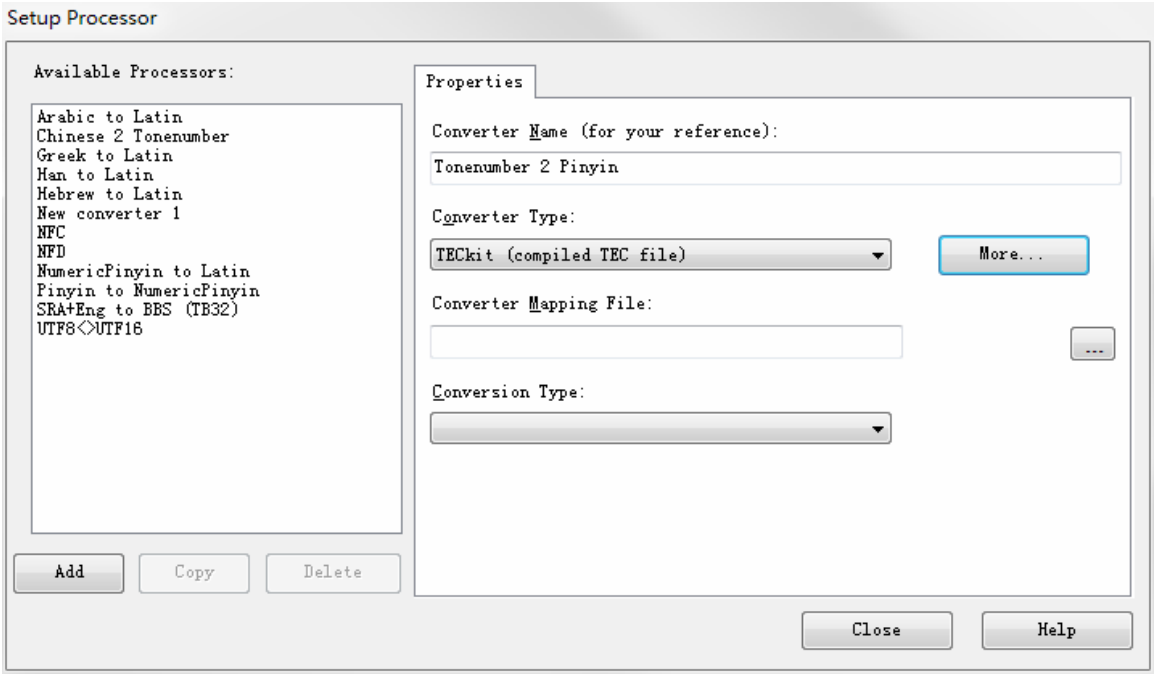
If the tone number field has any unresolved ambiguities then the Pinyin field is cleared, otherwise the Pinyin field is always overwritten (when holding Shift to enable database changes.)

Bulk Edit Process

In some situations it may be more convenient to use the Bulk Edit feature of Fieldworks Language Explorer to update the Pinyin fields. Two Python scripts are included for this purpose, both found in the Chinese/Lib directory:

- BulkEdit_HZ_2_Tonenumber.py
- BulkEdit_Tonenumber_2_Pinyin.py

These can be installed in FLEx via the Bulk Edit Process tab. Click Setup to get this dialog:



1. Type in a name for the converter as shown above, and then click the *More* button.
2. Choose ‘Python Script’ and click Add.

3. In the Python Script dialog choose the Setup tab then browse for one of the files listed above using the '…' button in the top right. The Function field will be automatically filled in
4. Click on 'Save in System Repository', and accept the suggested name.
5. Finally click OK, then Close (you may get a warning message about an 'Invalid mapping file', which can be ignored.)

Now, in the Process drop-down list, choose the new converter.

NOTES:

- The BulkEdit_HZ_2_Tonenumber converter must load the Chinese dictionary for every row of data so it is very slow.
- The ICU Pinyin converters (see 'NumericPinyin to Latin' and 'Pinyin to NumericPinyin' in the screen-snap above) are not compatible in their behaviour with these tools, so they are not recommended.
- The sort string calculation requires access to two fields at once (the Chinese *and* Tonenumber fields), which isn't possible with Bulk Edit.

Trouble-shooting

The processing tools described in this document assume certain data formatting standards and so they produce errors or warnings when these are not met. Such rigorous checking helps to ensure consistency and high quality Chinese and Pinyin formatting. If you are getting errors that you don't understand, review the following sections. Double-click the warning messages to jump to the relevant entry and edit it.

Unknown Chinese or Invalid Punctuation

The Chinese field must contain only Chinese characters and full-width punctuation. Half-width punctuation and extraneous spaces will produce a warning like this:

刚(完了): [Use Chinese (wide) punctuation: u'(']

Traditional Chinese characters will also produce an error.

If valid characters produce this error, then please use the contact details to request they be added to the database.

Chinese and Tone-number mismatch

If the Chinese and tone number fields don't match, then a warning is produced, but the tone number field is not changed. E.g.:

⚠ 三 san1 shi2 [Expected "san1"]

⚠ 枣红色 zao3 hong2 se4 [Expected "zao3hong2 se4|shai3" or "zao3 hong2se4"]

Missing Chinese Data

If it is discovered that a Chinese word used in the dictionary or reversal index is missing from the database then the tone number field can be manually edited. To avoid future warnings about this entry, please use the contact details to request that the word be added to the database.

Appendix A: Sort String

The Chinese sort string is designed to generate *phonetically* sorted dictionary order based on a standard ANSI sort. It has the following form:

<pinyin spelling> <pinyin tone number> <stroke count> <stroke types>:

- Pinyin spelling: U-diaresis is represented with the digit '9' before the tone number so that it sorts before 'a', but after '5', thus generating the order: 'lu', 'lü', 'luan', 'lüe', 'lun', 'luo.'
- Tone number: numerals 1-4 for the tones, and 5 for neutral tone so the nuetral tone sorts after the other tones.
- Stoke count: represented as two digits with '@' for 0 and A-I for 1-9
- Stroke types: a series of digits (1-5) representing the stroke types in order, i.e. horizontal line (一), vertical line (丨), left slash (丿), dot (丶), and straight stroke with bending tip, or extended bending stroke (乛, 𠃌, 𠃍, 亅, 𠄎, 乙, 𠂉, etc.)