Formatting Data for PanLex

August 6, 2012

Abstract

All data (selected from sources) for PanLex must be formatted and normalized. This document explains the process of locating a source in the PanLex database that has not yet been entered, normalizing and formatting the source for PanLex, and finally adding it to the PanLex database.

1 Accessing the source

All the digitized sources are stored at a remote station, which can be accessed by going to panlex.net/panlex. Sources follow a naming convention as follows: the three letter ISO code for the source language, the three letter ISO code for the target language, the author's last name. If there is no known author, then the initials of the title are used (for example, EFD for English-French Dictionary). For example, for the English-French dictionary, the source would be named fra-eng:EFD. Since there is a length limit in PanLex, but not in panlex.net/panlex → Sources, the names are not always identical.

1. Go to panlex.net/panlex
2. Click on Sources
3. Click on queued (these are the sources that have not yet been entered)
4. For each source, there's typically one file, with the data, and a secondary folder, which will have extras (usually meta-data).
5. Click on the file to get access to the source data. This is typically what you will be normalizing.

If the file is in .pdf format, it can be converted to .html as follows:

1. Open the PDF file in Acrobat Professional.
2. Extract the dictionary pages into a new window.
3. Save that part as a new PDF file.
4. Save the PDF file as HTML (from Acrobat).
5. Additionally, use pdftotext, using Acrobat's HTML export. This will save as a text file
6. Compare the text file to the .html file, and use whichever one has a better output.

Having the file in .html format is desirable, as it provides extra formatting tags which can be used to parse the lines into appropriate sub-parts (for example, the source word, the target translation, part of speech, etc.).

2 Editing the source

Sometimes it will be necessary to edit the PanLex catalog record for the source, either before or after formatting the data for entry to PanLex. This is done in PanLem.

1. Go to panlex.org
2. Click on Try It
3. Click on PanLem (in the body of the text)
4. Click on Source → Edit
5. Sign in with your username and password
6. You will see a list of all the sources. Click on the one you want.
7. You will then see a table containing meta-data associated with the source.
8. Click on the button under 'change' for the one you want to edit, and edit appropriately.
9. When done editing, click on Submit at the bottom of the page.

What the values in the righthand column of the first table stand for is described below:

name This is the name assigned to the source, following the naming convention described in Section 1.

World Wide Web The url where the source was obtained, if there is one.

ISBN ISBN if there is one

author The author or authors of the source

title The title of the source (not the name given for identification in PanLex)

publisher The publisher of the source

year The year of publication

good-number This number refers to the quality of the source, with rating 0-9. The higher the number, the better the source is judged to be. This number helps PanLex rate how likely a given translation is, based on its source.

source-number Any number the editor wants to assign for the source.

fact-other Any other facts that are of interest about the source. Do not put editorial comments here; those go under fact-other in the next table down (the editors table, which only the editor sees).

permission-kind The kind of permission for use of the source. For example, creative commons license

right-text Quoted from the source itself, if there is a line about how the source may be used.

right-person The person to contact with questions about rights of usage.

right-email-address The email address of the person to contact about rights of usage.

The second table is viewable only by editors, and contains more editorial-type meta-data. What the values in the righthand column of the second table stand for is described below:

good-number-edit-necessary-1/0 refers to whether or not the quality of the resource has been determined. (0=determined, 1=undetermined). This will be changed once the source has been edited, and an appropriate value for good-number has been entered).

file-difficult How difficult the file is likely to be to edit, from 0-9 with 9 being most difficult.

file-submit-necessary-1/0 whether the [normalized/processed/converted] source file needs to be submitted (0=no, 1=yes).

expression-edit-necessary-1/0 whether a submitted file needs further editing

expression-edit-done-language-(aaa-bbb-ccc) which of the languages in the source are finished (require no further editing)

file-address name of the folder, or directory, containing the files of the source. Typically, this is the same as the name, except that the colon is replaced with a hyphen.

fact-other any further editorial comments

The third table, language, contains the languages in the source file. The fourth table, file-kind, shows the kind(s) of file of the source (pdf, xml, etc). You can click on edit-person to update the identity of the editor. The first time, you would be adding the identity of the editor (most likely you); you can later add or delete other users who are entitled to edit the data from the source.

3 Formatting the source

Whatever kind of formatting the source is in, it must be converted to a form that PanLex can read. PanLex files have a standard format, with one entry taking up several lines. However, it is also possible to put the data into tabular format, and then use a number of existing scripts to transform this into the PanLex format. Often this is easier, and is what is described here.

The tabular formatting is relatively simple:

- each line contains one entry
- each tab-separated column contains one piece of information about the entry, which will vary depending on what information the source contains. For example, the first column might be one language, the second a translation in another language, the third a translation in a third language, the fourth the part of speech, etc.
- order of the columns does not matter, with one constraint: any

columns that provide additional data (word class, metadata, etc) about expressions in a particular column must immediately follow that column.

- do not put a heading; you can define which column is which using the shell script

Most formatting scripts are written in perl. The most recent and general past scripts can be found in panlex.net/panlex → tools → tabularize. Other scripts can be found in panlex.net/panlex → sources → used → aaa-bbb-Name → secondary. It is a good idea to make your scripts as general (and readable) as possible, so that they are re-usable for any updates of the sources.

Once you have the data in tabular format, you can use the shell script main.sh to convert to PanLex format. main.sh (and accompanying perl scripts) can be found at panlex.net/panlex → tools → serialize.

4 Serializing

To use the serialization scripts found on panlex.net/panlex, make sure you have downloaded the latest versions of all scripts. Open main.sh, and look through the available scripts included. Not all will be needed for a given source, and which scripts are needed will depend on the source and the kinds of information it has.

The use of each script is documented in the comments of main.sh.

General options that will need to be changed for each script are to replace aaa-bbb-Author with the particular name of the file being processed. For example, if the file were named pak-por-Ferreira-1.txt, then aaa-bbb-Author should be replaced with pak-por-Ferreira. The number at the end of a file name is what is referred to in the comments of main.sh as *version of the input file.* After deleting or commenting out unneeded scripts, go through each script and alter this number such that for the first, the number matches the number at the end of the input file name (in the example case, this number would be 1). Each number after should increase by 1, as each script will produce an output file with this number augmented by 1. The output file of one script is the input file for the next, so it is important to make sure that these numbers are entered correctly.

The purpose of each perl serializing script is described briefly below:

apostrophe.pl This is to ensure that apostrophes are entered in a standard way into PanLex. There are actually many kinds of apostrophes in Unicode. If a language is being entered into PanLex for the first time, then a decision needs to be made as to which apostrophe will be used for that language. Descriptions of three kinds of apostrophes and how they are most commonly used are described in the comments for this script.

extag.pl This script tags expressions as expressions recognizable by PanLex. This needs to be done for any file, as an entry without any expression is not a legitimate entry in PanLex. Additionally, expressions are split into either synonymous translations or separate expressions (meaning change), as appropriate.

exdftag.pl This is useful if the tabularized file sometimes contains definitions in parentheses, or longer expressions which are more appropriately categorized as definitions.

dftag.pl This is useful if the tabularized file contains a particular column of definitions

mitag.pl This is useful if the tabularized file contains a particular column of meaning identifiers (typically a number; useful for finding the entry in the original source, as the meaning identifier can be matched).

wctag.pl This is useful if the tabularized file contains a particular column of word class. Word class from the original source will be transformed, using this script, to the appropriate PanLex name (more specific information, such as the particular noun class of a word, will be tagged as metadata). If you suspect that the word classes identified in the source are not at all standard, it may be necessary to alter wctag.pl to include them. Standard word classes are defined in the hash *my %wc*, with the source name of the word class as the key and the PanLex name for it as the value.

mdtag.pl This is useful if the tabularized file contains a particular column of metadata, that cannot be described as a definition, a meaning identifier, a word class, or a domain. Alter the metadata tag as appropriate, so that it describes what kind of information it is. The default is *gram*.

dmtag.pl This is useful if the tabularized file contains a particular column of the domain (for example, *plant*, *tool*, etc).

normalist.pl This is one of two normalizing scripts. Normalization will be discussed more in section 5.

mnsplit.pl This is useful if the tabularized file contains multiple meanings in a single entry (ie, not synonymous translations). This script will split these multiple meanings into separate entries. An example of multiple meanings for a word in French, *pâte*, in English are *dough*, *pasta*. While the spelling in French is the same for these two words, they do not have the same meaning.

wcshift.pl This is useful if word classes are prepended to something, but otherwise conform to PanLex specifications.

normalize.pl This is the second of two available normalizing scripts. Which to use will depend on the format of the input file.

out-simple-0.pl This is the first of several final scripts, which will transform the tabular file into the style of PanLex. Use out-simple-0.pl if the source has no other data besides expressions (making it a simple file) and there are more than two languages.

out-simple-2.pl Use out-simple-2.pl if the source is bilingual, has no other data besides expressions, and the first column contains only a single expression (with no synonyms).

out-full-0.pl Use out-full-0.pl if the source has data in addition to expressions (word class, definitions, etc), and contains more than two languages

out-full-2.pl Use out-full-0.pl if the source is bilingual, has data in addition to expressions, and the first column contains only a single expression (with no synonyms).

5 Normalization

Normalization of expressions is very useful, since if a single expression (that is, an expression with the same meaning and form) is entered in two ways into PanLex, PanLex will treat the single expressions as two different expressions. In turn, this limits PanLex's ability to link expressions between languages, based on a common translation into a third language for each.

Both normalist.pl and normalize.pl match expressions to what is already in PanLex, suggesting alternate expressions, or not, depending on the score parameters. Thus, it is advisable to normalize only if the language of the column to be normalized is already very well-represented in PanLex.

Even so, there are likely to be normalization errors, where an expression is altered to another expression with a higher score. For example, in Portuguese, there are two different words, *avô* and *avó*, which are very close in form. Parameters in normalize.pl should be adjusted to prevent PanLex from normalizing *avó* to *avô* or vice versa, since one or the other will have a higher score (scores of words are based on the number of different sources that they appear in, and each source's quality estimate, in PanLex). This may require some experimentation, as each source will differ as to which are the best parameters.

Even so, perfect normalization is unlikely from this algorithm alone. To get the best quality data, it may be helpful to correct the resulting file by hand (expressions to be altered/deleted are tagged *exp*, making it easier to search for them).

In all cases, it is helpful to have some knowledge of the language(s) whose expressions are to be normalized, as this will help with spotting errors in normalization.

6 Uploading the normalized source

Once a source has been formatted for PanLex, the last step is to upload it. This can be done at panlex.org → Try It → PanLem → file - submit. Enter username and password, then choose the appropriate file name. This may or may not exactly correspond to the one from panlex.net/panlex. Choose whether a simple file (out-simple-n.pl was used), whole file (out-full-n.pl was used), or xml (not discussed here, but useful if the original source file was in xml) is being submitted.

Before submitting, find the file (aaa-bbb-Author-final.txt) and click *Check*. This will show any impossible lines (lines that PanLex can't read or doesn't allow). It will not check for other errors, however, so look over the file before submitting it and see if there are normalization errors, mislabeled expressions or definitions, etc.).

When it is reasonably certain that there are no errors, click on *approve* - *more* or *delete* - *replace* - *whole* if a previous submission is to be replaced.