LDC Word Aligner
Version 1.20

Stephen Grimes
sgrimes@ldc.upenn.edu

October 2010

1. INSTALLING THE PROGRAM

A. Windows

B. Linux

C. From source

2. LOADING THE PROGRAM

A. In Windows:

Double-click on program shortcut.
Or double-click on .wa file (if filename extension assocation functional)

B. From the command line - several examples

I. Without opening a file

The user changes to the appropriate directory of the executable, or the user types the entire path to the exectuable, followed by the filename of the alignment file. By convention the program uses the .wa file extension.

/path/to/tool/alignerCH.py -f /path/to/file/filename.wa

II. Specifying an annotation file

/path/to/tool/alignerCH.py -f /path/to/file/filename.wa

III. Speficying input files and an annotation file to create from them

/path/to/tool/alignerCH.py -s /path/to/source/chinese.tkn //
-t /path/to/english/english.tkn -f /path/to/file/parallel.wa

IV. Opening the file to a specific sentence. Let's go to the third sentence.

/path/to/tool/alignerCH.py -i 3 -f /path/to/file/filename.wa

If the user does not use -i to specify a sentence to start at, the tool will load the first sentence for which annotation has not been completed.

3. USE OF THE PROGRAM

The user begins by assigning tags or links to words. When a link is created, the existing tags are used to determine the link type. If a linked word has its tag updated, the new tag is used to determine the new link type. Tags are selected by right clicking in the box of the word to be annotated and choosing from the context menu.

To create a link, the user first selects all words by left-clicking on them. When all words have been selected, the space bar is pressed or the "Correct" button is selected. For "Incorrect" links, the "i" button is pressed or the "Incorrect" button is selected.

In order to annotate a word as "Not translated", it is assigned one of three tags: "Not Translated: Context Obligatory", "Not translated: Context optional", or "Not translated: Incorrect". It is the immediately colored orange, and the token appears in the link table as being aligned to <not translated>.

For a complete list of possible tags and possible links, see the end of this document. For help in determining the appropriate use of the tags and link types, please refer to the annotation guidelines.

Links may be deleted by selecting the link to delete from the lower right link table. However, individual words may be removed from a larger link by double clicking on the box of the word. If the word is the last remaining word from one language in a link, the entire link is removed.

Tokens can be added to existing links by selecting the existing link (either click on a word that is part of the link or select the link in the table). Once the existing link is selected, select the word to be added and press space bar (or "i" for an incorrect link).

The user may save the file at any time by selecting File->Save from the menu. The program automatically saves when advancing to the next sentence.

The Undo button removes the last link in the link table. This is usually the last link added, although the order of links in the link table may change if the sentence is reloaded.

The Next and Previous buttons may be used for sentence navigation. Alternatively, the user may change to any sentence by double clicking on that sentence in the source or translation boxes on the upper right side. The Refresh button reloads the current sentence.

Sorting the table. If the Edit->Toggle Sorting feature is selected from the menu, this enables sorting of the table by left-clicking a column heading. This is particularly useful for sorting by link type. Sorting is automatically turned off when a new sentence is loaded.

4. COLORS

4.1 LDC Chinese configuration

Yellow: Unannotated tokens.
Blue: Selected tokens.
Lavender: Terminal links.
Green: Composite links.
Orange: Not translated.
Red: Incorrect.

4.2 LDC Arabic configuration

Yellow: Unannotated tokens
Purple: Selected tokens
Green: Correct Link
Orange: Incorrect Link
Light Blue: Not translated, Correct Link
Light Red: Not Translated, Incorrect Link

4.3 Default configuration

Uses same colors as LDC Arabic configuration.

5. KNOWN ISSUES / BUGS

In order to tag a word, the user may have difficulty if right-clicking on the word itself. Instead click in an area of the box where there are no words or tags.

Currently one word is not permitted to be a member of multiple links. This behaivor is by design.

6. POSSIBLE TAGS AND LINKS

6.1 LDC Chinese configuration

TAGS

FUN Function
DEC DE-clause
DEM DE-modifier
DEP DE-possessive
TEN Tense/Passive
OMN Omni-function-preposition
POS Possessive
TOI To-infinitive
SEN Sentence Marker
MEA Measure-word
DET Determiner/demonstrative
CLA Clause marker
ANA Anaphoric-reference
LOC Local context marker
RHE Rhetorical
COO Not Translated: Context obligatory
CON Not Translated: Context optional
INC Not Translated: Incorrect
TYP Typo
MTA Meta

LINKS*

SEM Semantic
FUN Function
DEP DE-possessive
DEC DE-clause
DEM DE-modifier
GIF Grammatically Inferred Function
GIS Grammatically Inferred Semantic
LOC Contextual
INC Incorrect
NTR Not Translated
CON Not Translated: Context optional
COO Not Translated: Context obligatory

*Link types are not annotated by the user but instead determined based upon the combination of tags assigned to component tokens of the link.

6.2 LDC Arabic configuration

TAGS

MET Metadata
TYP Typo
GLU Unmatched and glued
TOK Tokenization error
MRK Markup attached

LINKS

COR Correct
INC Incorrect

6.3 Default configuration

By default there are no tags available. To specify tags, edit the tags.txt file.

LINKS

COR Correct
INC Incorrect