JGloss User's Guide

Michael Koch

JGloss User's Guide

Michael Koch

Publication date 2013

Table of Contents

1. Introduction 1
    - About JGloss 1
    - Acknowledgements 1
    - License 1
2. Setting up JGloss 2
    - What you need 2
    - Running JGloss 2
3. Using JGloss 4
    - Quickstart 4
    - The Document View 5
    - The Annotation Editor 5
    - Importing text documents 5
    - Exporting annotated documents 5
        - HTML 6
        - Plain Text 6
        - LaTeX 6
        - Annotation List 6
    - The Dictionary Lookup Dialog 6
    - Text Parser selection 7
        - The Kanji parser 7
        - The ChaSen parser 7
    - The Preferences Dialog 8
        - General 8
        - Style 8
        - Dictionaries 8
        - Exclusions 9
    - JGloss menus 9
        - File 9
        - Edit 9
        - View 10
        - Annotation 10
        - Help 10

List of Figures

3.1. JGloss Document Window 4

Chapter 1. Introduction

About JGloss

JGloss is an application for adding reading and translation annotations to words in a Japanese text document. This can be done automatically and manually. When a text document is first opened, words will be looked up in dictionary files and the first reading and translation (if any) is used to annotate the word. The user can then edit the annotations: choose among the readings and translations found in the dictionaries, enter own readings and translations, remove annotations and add new annotations. The document can be exported as plain text with annotations, HTML (with support for the *Ruby Annotation* [http://www.w3.org/TR/ruby/] XHTML specification) or LaTeX.

The application is designed as a translation aid for people learning Japanese. With some new document, you can first skim the text and change the annotations to match the likeliest meaning of the word. Then you can print/export the text with annotations and start working on the details of understanding the text without having to resort to a paper dictionary all the time.

JGloss is written in Java and should work on any computer with support for Java 7 [http://www.java.com/].

Acknowledgements

JGloss is written by Michael Koch (`<tensberg@gmx.net>`). It owes many ideas to Jim Breen's work, particularly the WWWJDIC [http://www.dgs.monash.edu.au/~jwb/wwwjdic.html] and XJDIC [http://www.csse.monash.edu.au/~jwb/xjdic/]. The kanji parser is based on ideas from the WWWJDIC [ftp://ftp.monash.edu.au/pub/nihongo/www-jdict.ps.gz]. The character encoding detection uses code from Yasuhiro Tonooka's kcc (Kanji Code Converter). The French localization is contributed by Alexandre Beraud. Heinrich Künsting helped with the LaTeX export template design and wrote the LaTeX-CJK list template. Some of the file chooser icons are taken from the KDE project [http://artists.kde.org].

License

Chapter 2. Setting up JGloss

What you need

JGloss is a Java application. To run it, you will need a Java implementation that conforms to the Java 7 specification, e. g. Oracle's Java Runtime Environment [http://java.sun.com/] (JRE). The OpenJDK which is installed on many Linux distribution also works.

Your computer system should already be configured to work with Japanese text. You must have a Japanese font installed. Having a Japanese input method installed is not absolutely necessary, but very useful. As a test, if your web browser can display Japanese text, it should be possible to set up Java do do the same.

To use JGloss, you will need some dictionaries. Currently supported dictionary formats are:

EDICT2 [http://www.csse.monash.edu.au/~jwb/edict_doc.html]	EDICT dictionaries are Japanese to English word dictionaries. You can download them from the Monash Nihongo FTP Archive [http://ftp.cc.monash.edu.au/pub/nihongo/00INDEX.html#dic_fil] . Each dictionary also needs an index file. If no index file is found, it will be created automatically by JGloss and saved in the dictionary directory. Should the index file creation fail, for example because the directory is write-protected, the dictionary can't be used and an error message is shown.
Wadoku Jiten [http://www.bibiko.de/dlde.htm]	The Wadoku Jiten is an extensive Japanese-German dictionary. Current versions of the Wadoku Jiten are available in EDICT2 format from wadoku.de [http://www.wadoku.de/] To use it, download the file `wadokudict2_201....tar.bz2` from the download page [http://www.wadoku.de/downloads/] , unpack it and add the unpacked file `wadokudict2` to the list of dictionaries used by JGloss in the dictionary dialog.
KANJIDIC [http://ftp.cc.monash.edu.au/pub/nihongo/kanjidic_doc.html]	KANJIDIC dictionaries contain information about individual kanji, among other things readings and translations. You can find KANJIDIC dictionaries at the same location as the EDICT dictionaries.

Running JGloss

JGloss requires no installation. On a Windows system with the Java Runtime Environment installed, double-clicking the `jgloss-...jar` file should start the application. To start JGloss from a shell, change to the directory which contains the JAR file and enter **`java -jar jgloss-...jar`**. JGloss has some command line options: **`java -jar jgloss-...jar [option] file ...`**.

`-h` , `--help` , `/?`	Shows a short help message with the list of options.
`-i` , `--createindex`	Creates index files for the dictionary files given on the command line. The index files will be saved in the current directory. JGloss tries to create an index file for a dictionary file automatically when none is found. If this fails, for example because a normal user has no write permissions for the dictionary directory, you can log in as a privileged user (e. g. administrator or root) and use this option to create the index files.

`-f, --format`	Prints the format of the dictionary files given on the command line.

Chapter 3. Using JGloss

Quickstart

When you start JGloss for the first time, the welcome wizard is shown. This dialog will guide you to the steps required to import your first Japanese text. After you have completed the wizard, the annotated text will be shown in the document window.

Figure 3.1. JGloss Document Window

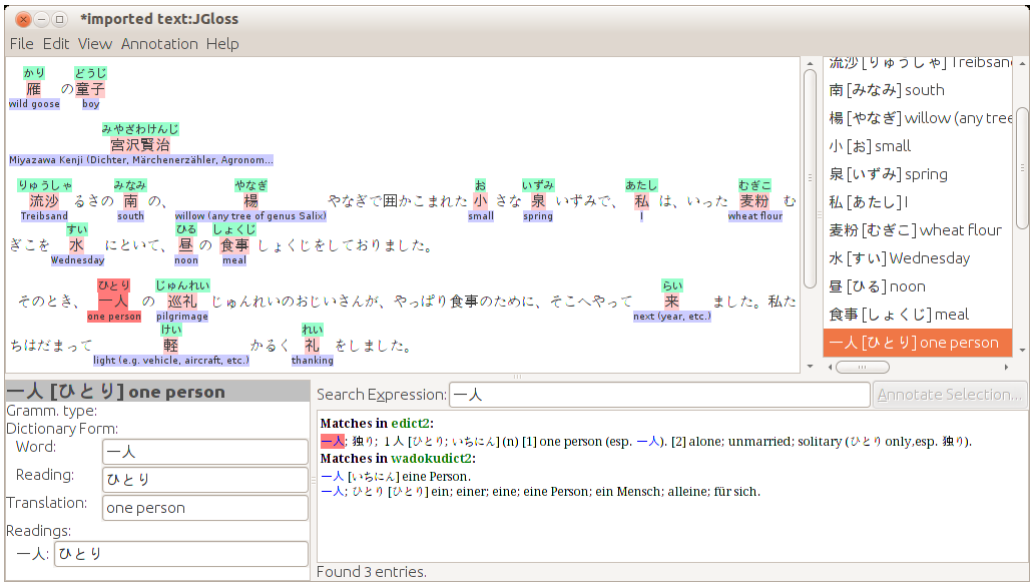

Figure 3.1, “JGloss Document Window” shows the JGloss window after some text is imported. The top-left part of the window shows the annotated document. An annotated word is shown in light red, with the reading annotation above and the translation annotation below the word. The top-right part of the window shows the list of annotations.

The bottom-left part of the window contains the annotation editor. For the selected annotation the annotation editor shows the current reading and translation, and dictionary form of the entry. You can use the annotation editor to change the reading and translation of a word by editing the text in the text fields.

The bottom-right part shows dictionary search results for the currently selected text. If you select an annotation, the annotated word is automatically searched. Click on a result to use it as the new reading or translation. If you select some unannotated text in the top-left document view, it will also be automatically searched. Click on the Annotate selection... button to add an annotation for the selected text.

The heuristics used for generating the annotations are not perfect. For example, if you import a text using the EDICT2 dictionary file, 来る is assigned the reading きたる instead of くる and 一人 is assigned the reading いちにん instead of the more common ひとり. This happens because the application picks the first reading and translation found in the dictionary. Also, the algorithm used for verb/adjective deinflection can produce false results. What follows is that you should not always trust the automatic annotations, the document will require some editing to be correct.

When you have finished editing your document, you can export it to different formats. Select HTML from the Export entry in the File menu. In the file chooser you can select which type of annotations will be written. Select a file name and save the HTML file. If you are using Microsoft Internet Explorer which already (more or less) supports the *Ruby Annotation* XHTML specification,

the ruby will be rendered above the annotated word by the browser. Translations will be shown in a popup window if you move the mouse over a word.

The Document View

The top-left part of the window shows the annotated document. Annotated words will have a colored background. You can change the colors or switch them off in the preferences dialog and toggle the display of reading and translation annotations in the View menu. If you left-click on an annotated word, the annotation will be selected in the annotation editor. A right-click will select the word and pop up a context menu with options for this annotation (see the section called “Annotation”). To look up the currently selected text in the dictionaries, select Dictionary Lookup from the Edit menu.

The Annotation Editor

The annotation editor is used to edit the reading and translation annotations of the currently selected annotated word in the document. It is displayed in the bottom-left quarter of the JGloss window (see Figure 3.1, “JGloss Document Window”).

The dictionary form (word, reading and tranlation) is used for the vocabulary list created by the various export formats. Below that, the list of readings for the word as it appears in the document is shown. If a word has no kanji characters, the reading will span the whole word. Otherwise, a reading annotation is added to every kanji substring of the word.

Importing text documents

To create a new annotated document, select Import or Import Clipboard from the File menu. Import Clipboard will create a new document from the content of the system clipboard, the import options can be configured in the General Preferences dialog.

Selecting Import will show the Import dialog. In the text field you can enter an HTTP URL or the file name of a local document. The document should be plain text. There is some support for importing HTML documents, but it will only work for documents with simple layout. If you click the Choose File button, a file chooser dialog will pop up and let you select a local file. The character encoding of the file can be selected in the Character encoding ... popup menu. You can usually leave it on the <auto> setting, which will make JGloss auto-detect the encoding of the file. If the auto-detection fails for a file and the document is not displayed correctly, you can select the encoding manually. You can choose the text parser for the automatic annotation in the Text Parser part of the dialog (see the section called “Text Parser selection”). Clicking the Import button will import the selected file.

If the imported document contains a reading for a annotated word, it will be used for the reading annotation, and the first translation found with this reading will be used for the translation annotation. Otherwise, the first reading and translation found in a dictionary will be used for the annotation. You can change the annotations later manually.

Note

The JGloss application is quite resource intensive. It can take a rather long time for the newly imported document to be displayed. You should consider splitting a long text in several shorter files before you import it.

Exporting annotated documents

JGloss supports exporting of annotated documents in several formats, described below. You can select one of the formats from the Export submenu in the File menu.

A common option for all formats is the character encoding of the generated file. What encoding you use depends mainly on what application you want to use the exported file with. Modern web browsers should support all of the encodings. For other applications, if you are working on Windows, you should try Shift-JIS and on Linux EUC-JP. If the document or your annotations contain characters not in ASCII or the Japanese character set (e. g. German umlauts), you should use UTF-8, which can represent all characters.

HTML

If you export the document in HTML format, the document can viewed in any web browser that supports display of Japanese characters. The document title set in JGloss will be used as the title of the HTML document. The markup defined in the *Ruby Annotation* specification is used to embed the annotations, browsers which support it will render the annotations above/below the annotated words. Translations are shown using JavaScript in a floating window and the status bar of the browser when the user moves the mouse over an annotated word. In Firefox and other browsers which don't support ruby annotations but support the neccessary JavaScript functions, reading annotations will be shown in a floating window above the annotated word and in the status bar. In other browsers, the reading annotations will be shown in the document after the annotated word, the translations are not available.

Plain Text

The plain text export function will generate a text document similar to the originally imported document. Annotations will be written after the annotated word, enclosed in brackets.

LaTeX

The LaTeX export function will generate a text document in LaTeX format. There are several document style variants you can choose from by selecting the corresponding template from the template chooser. You can choose the document font size from the Font size menu.

Standard LaTeX can't handle Japanese documents. JGloss therefore uses the CJK macro package, which adds support for far eastern scripts. In order to use it, you will have to download and install the macro package and the corresponding font package. On Ubuntu Linux, simply install the package latex-cjk-japanese (type **`sudo apt-get install latex-cjk-japanese`** in a command prompt). Once the package is installed, you can process the files generated from JGloss with the commands latex or pdflatex.

Annotation List

The annotation list export function will write a text file listing all annotations in the document. The dictionary form and the selected translation of the annotated word is used. Annotations will only be written once, duplicate entries will be skipped. You can use the generated text as a basis for a vocabulary list.

The Dictionary Lookup Dialog

You can look up words in the dictionaries from the word lookup dialog. Enter a Japanese or English word in the Enter Expression text field. The search is performed as soon as you stop typing. The results will be displayed in the lower area of the dialog window. The part of the entry which matched the expression will be highlighted blue for each result line.

You can select the search mode from the Search Options part of the dialog. Exact matches will only return entries where the entered expression is identical to the word, reading or one of the translations of the entry. Starts with Expression and Ends with Expression will find entries where the word, reading or one of the translations have the search expression at the beginning/end. Any

Match will return entries where the search expression appears anywhere within the result. Note that the dictionary formats may not support all of the search options. For example, the EDICT implementation does not do "ends with" searches for readings so you might not get all possible matches if you use this search mode.

You can limit the search to a single dictionary by selecting the Search Dictionary radio button and selecting the dictionary from the popup menu to the right. If the Search all Dictionaries radio button is selected, all dictionaries made available to JGloss in the preferences will be searched.

The options below Filter Results will limit the search results to dictionary entries which are marked with the selected attributes. Not all dictionary format support the attributes. The checkboxes will be disabled depending on the selected dictionary. Search Fields lets you limit the search to the word, readings or translation parts of the dictionary entries.

Field Match Mode controls if the the search option apply to a word in the field (e. g. translation) or the whole field. E. g. if you select Exact match and search for **`character`**, the translation *Chinese character* will only be found if the field match mode is Match words.

Text Parser selection

JGloss can use two different parsers for the automatic annotation of Japanese text, the Kanji parser and the ChaSen parser. Select the parser by clicking the respective radio button. The ChaSen parser will only be available if the chasen program is installed (see below).

If the annotate first occurrence option is selected, each word in a document is only annotated the first time it appears. This decreases the RAM usage and the time it takes to display the document. The Guess paragraph breaks option controls how line breaks in the imported document are converted to paragraph breaks. If the option is selected, JGloss tries to determine if a line break in the imported document signifies the end of a paragraph or if it is used for formatting reasons only. In the second case, it will be ignored.

Some documents you can find on the internet already have reading annotations added to kanji words in the form of some hiragana enclosed in brackets after the kanji word. The parsers can generate reading annotation entries for these words. You can select the brackets used in the document for reading annotations with the Brackets used... box. If the document contains no reading annotations, you can select none or simply ignore this setting.

The Kanji parser

The Kanji parser is built into JGloss. A simple heuristic is used for choosing words to annotate: for a sequence of katakana characters, the whole sequence is treated as one word and looked up. For a sequence of kanji characters followed by hiragana characters, the algorithm first looks for possible inflected forms in the hiragana string and will try to find words that consist of the kanji word and the dictionary form of the inflected forms that appear in the hiragana string. If no match is found, only the kanji part is looked up. If still no match is found in any of the dictionaries, prefixes of the kanji word will be tried and if this leads to a match the process will be repeated with the remainder. A consequence of this method is that hiragana words will never be annotated automatically even if they are in the dictionaries.

The ChaSen parser

The ChaSen parser uses the ChaSen morphological analysis program to decompose Japanese text in words and to derive the base form of inflected words. It is slower than the Kanji parser, but will annotate hiragana words as well as kanji and katakana words. It also does a better job of deinflecting verbs and adjectives. You can download ChaSen from the ChaSen homepage [http://chasen-legacy.sourceforge.jp/] . On Ubuntu Linux, you can simply install the package chasen (**`sudo apt-get install chasen`**). After installation, you have to set the path to the

chasen executable in the preferences dialog . It usually is `/usr/bin/chasen` under Unix or `c:\Program Files\chasen21\chasen.exe` under Windows.

The ChaSen program is used to generate a list of words with their reading and base form from the parsed text. The words will be looked up in the dictionaries, and if an entry is found, an annotation will be generated. If no dictionary entry is found and the word is not inflected, kanji substrings will be tried. A reading annotation with the reading output by ChaSen is also added if the reading returned by ChaSen is different from the first reading found in the dictionaries. Since the ChaSen program uses its own set of dictionaries to detect words, it might not recognize words which are found in the dictionaries used by JGloss but not in the ChaSen dictionaries.

The Preferences Dialog

The preferences dialog contains four panels, one with general preferences, one for setting the visual appearance of the annotated document, one for managing the dictionaries and one for editing the list of words excluded from annotation. You can access the dialog by selecting Preferences from the Edit menu.

General

You can select the window opened on startup with the Open empty JGloss document and Open Word Lookup dialog radio buttons. The function of the left mouse button is changed in the Left-clicking an annotated word section of the dialog. The text parser used when importing the clipboard content can be selected in the Import Clipboard Parser section of the dialog. See the section called “Text Parser selection” for details. You can set the location of the ChaSen parser program in the ChaSen executable text field. It usually is `/usr/bin/chasen` under Unix or `c:\Program Files\chasen21\chasen.exe` under Windows. If the program cannot be found, the ChaSen parser will not be available.

Style

The Japanese User Interface Font lets you choose the font used in the display elements of JGloss. If the default Java fonts don't contain Japanese characters or you want to use a different font, select the Use this font radio button and choose a font from the list. Note that not all fonts in the list can display Japanese characters.

The other font selection options determine the fonts used in the word lookup result list and in the document view. Select a font by using the popup menu. The font size can be selected from the popup menu to the right, or you can enter the font size manually if it is not in the list. You can select a different background color by clicking on the button with the color label, or disable the use of a background color by unchecking the set background color checkbox. The Highlight color ... button lets you select the color which is used for highlighting the currently selected annotation in the document view.

Dictionaries

In this dialog you can set the dictionaries which are used when importing a text or adding annotations to a document manually. Download dictionaries... opens a dialog which lets you download and install common dictionaries. Click on Add dictionary file to add one or more dictionaries to the list which you have downloaded yourself. JGloss currently supports dictionaries in EDICT2 and KANJIDIC format. To remove a dictionary from the list, select it and click Remove entry.

Since the automatic annotation process will search the dictionaries in the order in which they are displayed in the list and will use the first entry found as default annotation, you should put your preferred dictionary at the top. Select one of the dictionaries and click Move entry up or Move entry down do move it to the desired position.

Exclusions

This dialog lets you manage the list of words excluded from automatic annotation. When you import a document, no annotations will be added for the words in this list. You can export and import the list by using the corresponding buttons. The format of the list is simply one word per line.

JGloss menus

File

Import — Creates a new document by importing a text file, annotating it on the fly. See the section called “Importing text documents” for details.

Import Clipboard — Creates a new document by importing the content of the clipboard. See the section called “Importing text documents” for details. This item will only be enabled if the clipboard currently contains some text.

Open — Open a JGloss annotated document.

> **Note**
>
> The application can take rather long to display a document for the first time after it is loaded, especially for larger documents.

Open Recent — Open a JGloss annotated document by selecting it from the list of recently opened files.

Save — Saves the current annotated document in the JGloss file format. If the file name has not yet been determined, a file chooser dialog will be shown.

> **Note**
>
> The JGloss file format is a simple XML-based format with JGloss-specific elements.

Save As — Saves the current document in the JGloss file format under a new name.

Export — The entries in this submenu let you export the current annotated document in several formats. See the section called “Exporting annotated documents” for details.

Print — Print the annotated document. The current settings for font sizes and colors will be used in the printed document.

Close — Closes the document window. If the document has unsaved changes, a warning dialog will be shown. If all document windows and the word lookup dialog are closed, the JGloss application will quit.

Edit

Cut/Copy/Paste — These have the standard functionality. Note that because the document view is not editable, the Cut and Paste items are always disabled.

Dictionary Lookup — Selecting this item will show the Dictionary Lookup dialog (see the section called “The Dictionary Lookup Dialog”). If some text is selected in the current document, this text will be automatically searched using the current dialog settings.

Annotate Selection	This will let you annotate to the currently selected text. **Note** Annotations cannot overlap. If the selected text already contains annotations, these annotations will be deleted. Also, an annotation cannot span paragraphs. If text from more than one paragraph is selected, the annotation will end at the end of the first paragraph.
Document Title	Brings up a dialog which lets you set the title of the document. The title is used when the document is exported in HTML or LaTeX format.
Preferences	Selecting this item will show the preferences dialog (see the section called “The Preferences Dialog”).

View

Show Readings	This item toggles the display of reading annotations in the document view.
Show Translations	This item toggles the display of translation annotations in the document view.
Show Annotation Tooltips	If this item is selected, a window will pop up if you move the mouse over an annotation in the document view which show the dictionary form, reading and translation.

Annotation

The items in this entry manipulate the annotations of the entry currently selected in the annotation editor. The items also appear in the context menu for annotation editor items and annotated words in the document view.

Remove Annotation	Removes the annotation from the currently selected word. It will be removed from the annotation editor, and the word will be changed to normal text in the document view.
Add to Exclusions	Adds the word of the currently selected annotation to the list of words excluded from automatic annotation (see the section called “Exclusions”).

Help

Welcome	Shows the welcome wizard which also shown the first time you start JGloss.
About JGloss	Shows a dialog with information about the JGloss authors and license.