# A tool for Cross-Language Pair Annotations: CLPA

August 28, 2006

This document describes our tool called Cross-Language Pair Annotator (CLPA) that is capable to automatically annotate cognates and false friends in a French text. The tool uses the Unstructured Information Management Architecture (UIMA)[1] Software Development Kit (SDK) from IBM and Baseline Information Extraction (BaLIE)[2], an open source java project capable to extract information from raw texts.

## 1 Tool Description

CLPA is a tool that has a Graphical User Interface (GUI) capability that makes it easy for the user to distinguish between different annotations of the text. We designed the tool as a java open source downloadable kit that contains all the additional projects (Balie and UIMA) that are needed. It can be downloaded from the following address: CLPA[3].

The tool is a practical follow up for the research that we did on cognates and false friends between French and English. Since one of our main goals is to be able to use the research that we did in a CALL tool that is capable to help second language learners of French, CLPA is intended to be the first version of such a tool. At this point, the tool uses as knowledge a list of 1766 cognates and a list of 428 false friends. The list of false friends contains a French definition for the French word and an English definition for the English word of the pair. The tool offers and easy management of the resources. If the user would like to adjust/use other lists of cognates and/or false friends he/she needs to add the new source files to the resource directory of the project. The directory can be found in the home project directory.

[1] http://www.research.ibm.com/UIMA/
[2] http://balie.sourceforge.net/
[3] www.site.uottawa.ca/ ofrunza/CLPA.html

UIMA is an open platform for creating, integrating, and deploying unstructured information management solutions from a combination of semantic analysis and search components. It also has different GUI document analyzers that make it easy for the user to visualize the text annotations.

UIMA offers CLPA the GUI interface and an efficient management of the annotations that are done for a certain text. The user can select/deselect the cognate or false friend annotations. By default, both type of cross language pairs are annotated.

BaLIE, is a trainable java open source project that is capable to perform the following tasks: Language Identification, Sentence Boundary Detection, Tokenization, Part of Speech Tagging and Name Entity Recognition for English, French, German, Spanish and Romanian.

BaLIE is the project that provided the tokenization and part-of-speech tagging tasks for the French texts. The tokenization is done using a rule based method and the part-of-speech by using a probabilistic part-of-speech tagger, QTag[4].

In a single run, the tool is capable to annotate not only one document, but a directory that contains more than a single text document. UTF-8 is the format chosen to represent a document. The reason why we chose this format is due to the French characters and also for a consistency with the other projects that are used by CLPA the BaLIE project.

The following figures, provide snapshots of the interface that the user will see after the annotation process is completed.

The user, has to click on one of the text annotations to obtain additional information (e.g. what position in the text does the chosen word starts, what position does it end, the French definition of the French false friend word, the English definition of the English false friend word, etc.) about the chosen annotation.

## 2 Tool Capabilities

In its early stage of existence, the first version of CLPA is capable to annotate cognates and false friends between French and English in a French text. The cognate and false friend knowledge that the tool has is provided by lists of pairs of cognates and false friends. For now we intended high accurate lists instead of automatically produced lists.

In addition to the colored annotations (cognates are annotated with one color and false friends with another color), the tool provides other useful

[4]http://www.english.bham.ac.uk/staff/omason/software/qtag.html

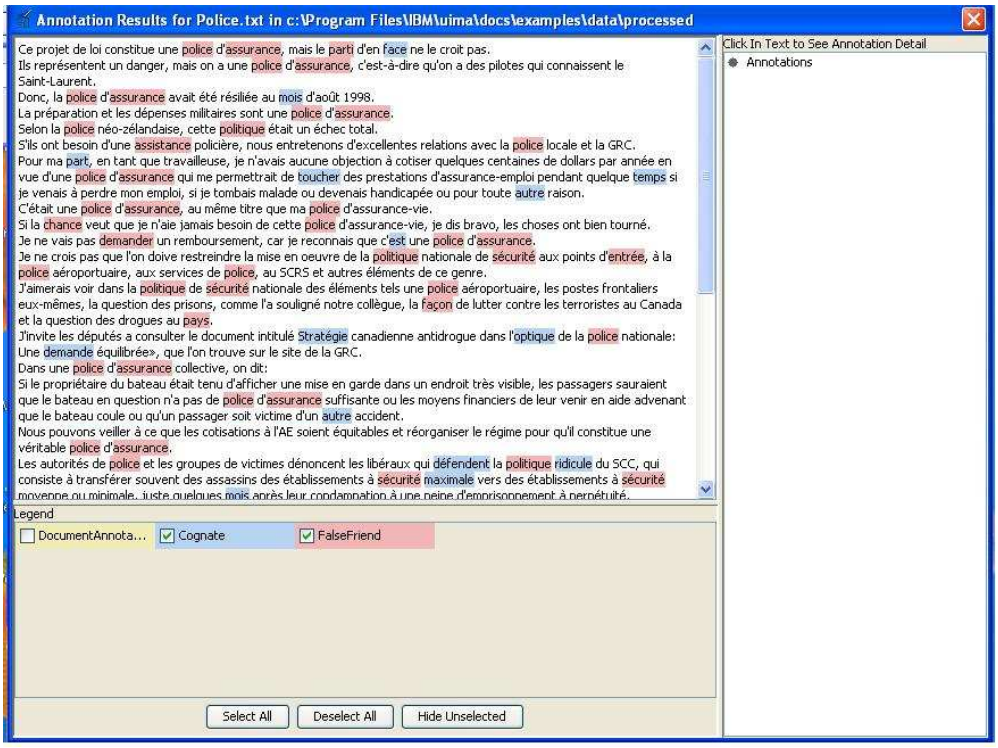


Figure 1: Cognate and False Friend annotations.

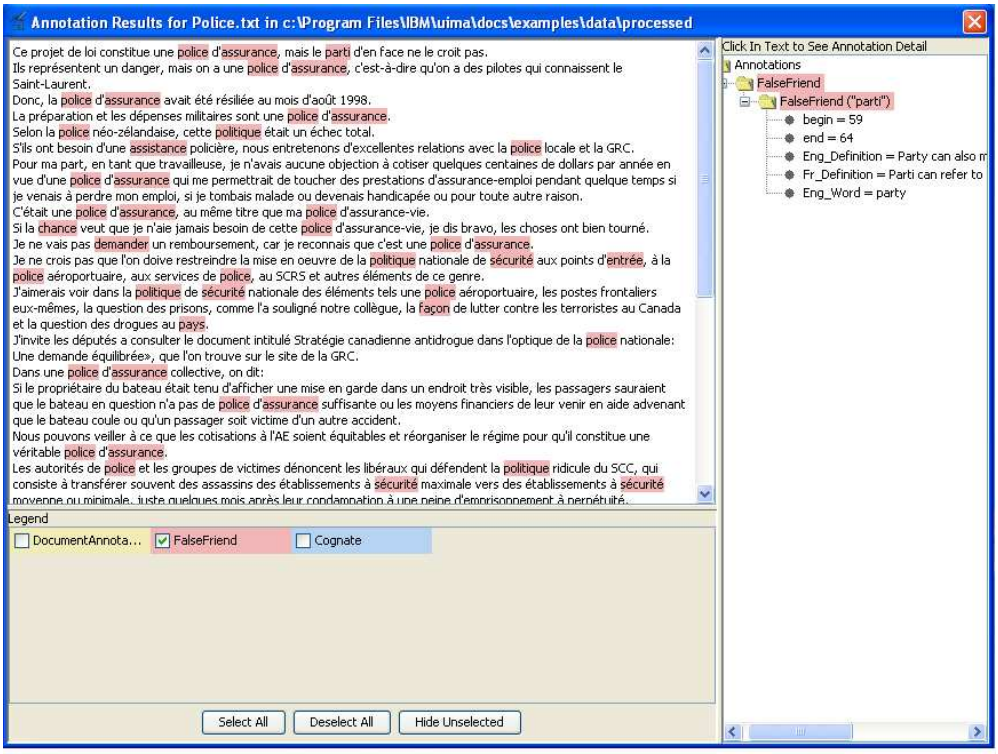


Figure 2: False Friend annotations.

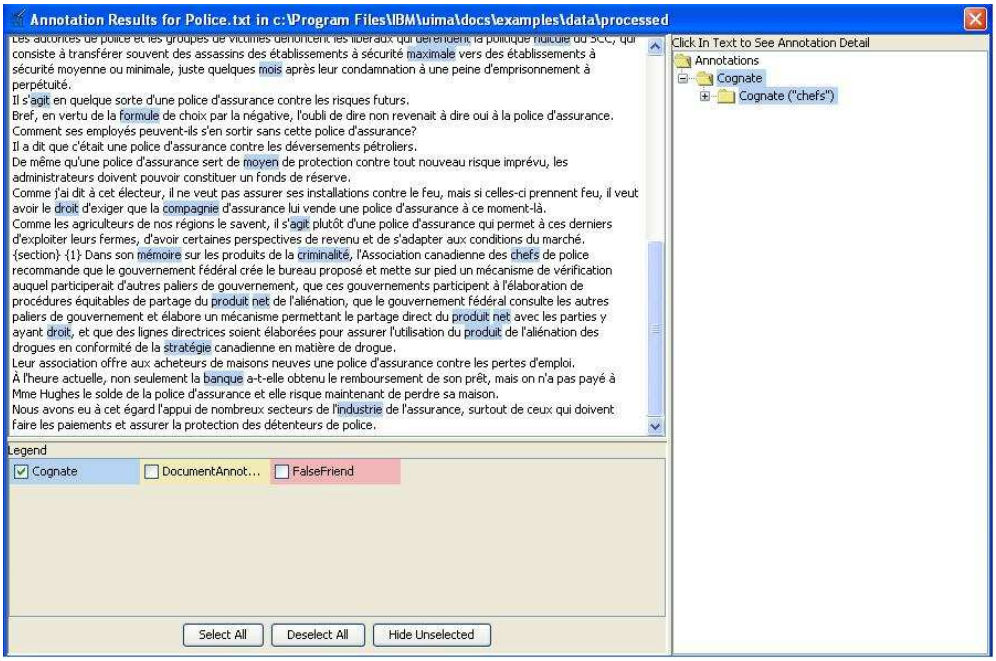


Figure 3: Cognate annotations.

information about the annotations. For the cognate words, it provides the position in the text and the English cognate word. For the false friend words it provides: the position in the text, the English false friend word/words, the definition of the French word, and the definition of the English words. The definitions were collected from the same resource[5] as the false friend word pairs.

The lists that the CLPA uses to annotate the French texts are free to download and can be used for future research. They are contained in the same package with the tool.

The annotations that the tool does are only for French content words: nouns, adjectives, adverbs and verbs. We have chose to annotate only the content words not to introduce some false alarms (e.g. the French word *pour* can be either adverb (*pro*), or preposition (*for; to*), and it is a false friend with the English word *pour* that is a verb), and also because they are of more interest for second language learners.

Since BaLIE can provide information regarding the part-of-speech tag for each token in the text, it was easy for us to make the distinction between the content and close class French words.

The tool does not lemmatize the text, and for this reasons some words might not be annotated or some errors might be introduced. Some annotations might be missed because the words are not in the base form and some errors might be introduced because the inflected form corresponds to the base form for another word (e.g. the verb *être* has the singular third person *est* form that corresponds to the base form of the cardinal point *est* that is cognate with the English word *east*).

The annotation will be done only for the tokens in the text that have the same form as the pair of words in the lists, the base form of the French and English words. For the next version of the tool, we will have the lemmatization step performed on the text before we do the annotation step.

Both UIMA and BaLIE are java projects that can be easily downloaded and used with Eclipse[6] SDK. In fact, UIMA has some of the features to be easily used with Eclipse. For both projects, documentation on how to install the projects is available from the corresponding web pages. For CLPA, the web page will provide instructions on how to install and put all the resources together so they can be ready to run for French text annotations.

[5] http://french.about.com/library/fauxamis/blfauxam_a.htm
[6] http://www.eclipse.org/