UIMA Apertium Translator

1. BASIC INFORMATION

Tool name

UIMA Apertium Translator

Overview and purpose of the tool

This tool translates text from a source language into a target language. It operates on text that has previously been tokenised and morphologically analysed, and POS-tagged. Target language tokens are assigned POS tags and morphological analyses. The Apertium Translator is a module of Apertium machine translation system[1] (Armentano-Ollet et al., 2006). The provided tool can currently operate on a subset of the languages that are supported by the Apertium system, namely: English, Spanish, Calatan, Galician, Portuguese and Basque.

NOTE: The morphological analysis required and POS tagging prior to running the transfer component MUST be carried out by running the Apertium morphological analyser (which also performs tokenisation), followed by the Apertium POS tagger.

The tool is provided as a UIMA[2] (Ferrucci et al., 2006) component, specifically as Java archive (jar) file, which can be incorporated within any UIMA workflow. It can be run within the U-Compare text mining platform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record), since the types of annotations it produces are compliant with the U-Compare. However, U-Compare does not currently support visualization of the output of this tool, since multiple subjects of analysis (sofas) are used. The output of the tool may, however, be visualized using the UIMA annotion viewer.

The Apertium morphological analyser and POS tagger are also available as a UIMA components (called ApertiumMorpho and ApertiumPOS, respectively).

A short description of the algorithm

The transfer module of Apertium uses bilingual dictionaries and structural transfer rules to perform the translation. Morphological dictionaries are used to generate appropriate surface forms in the target language. See the Apertium documentation (http://wiki.apertium.org/wiki/Documentation) for more information.

[1] http://www.apertium.org/

[2] http://uima.apache.org/

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The tool can most easily be run using the U-Compare platform (see separate META-SHARE record). The only requirement to run U-Compare is for Java 6 to be installed.

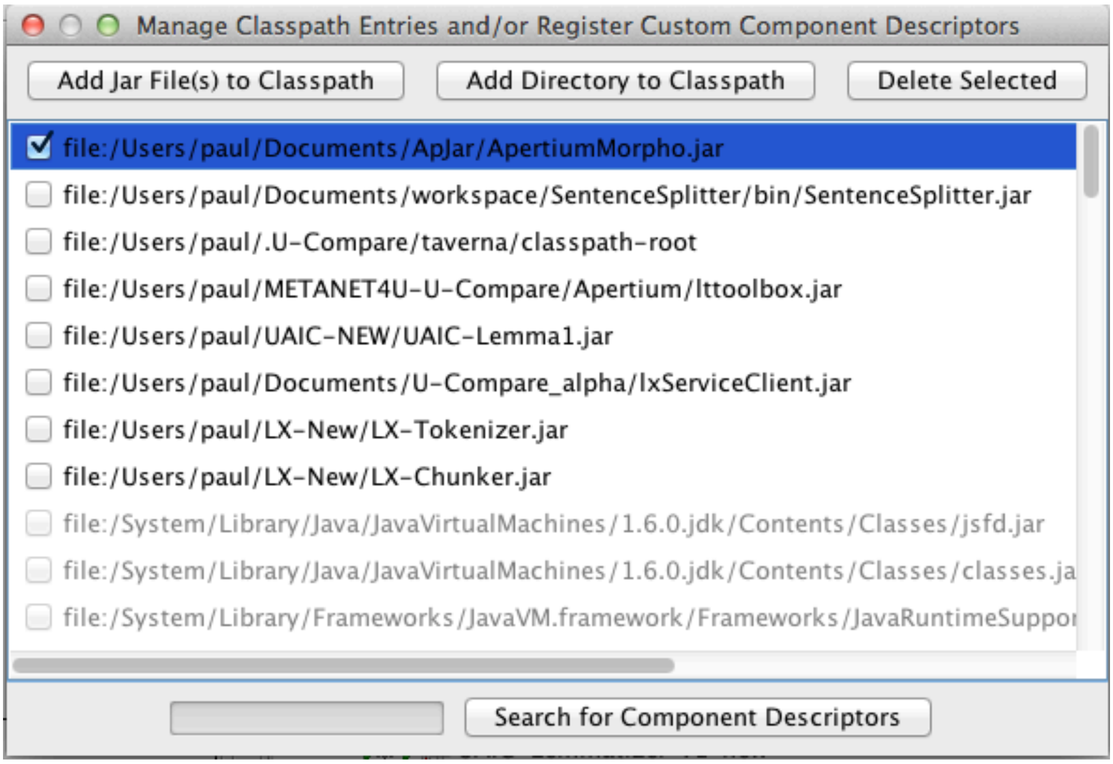
To run the tool as a UIMA component independently of U-Compare, Apache UIMA must be installed (see http://uima.apache.org/). UIMA must currently also be installed in order to use the UIMA annotation viewer, to view the output of the module (although support for MT components in U-Compare is planned.

Installation

In order to run the Apertium Translator in U-Compare, it must be imported into U-Compare. Information about downloading and running U-Compare can be found in the documentation associated with the META-SHARE record for the U-Compare Workbench or the U-Compare website: http://nactem.ac.uk/ucompare/.

Importing the UIMA component (provided as the file ApertiumMorpho.jar) is carried out in U-Compare as follows:

1) From the “Library” menu in the U-Compare Workbench, choose the item “Register External Components (Edit Classpath)”. This will cause a window to appear that allows external components to be managed. It is shown in Figure 1.

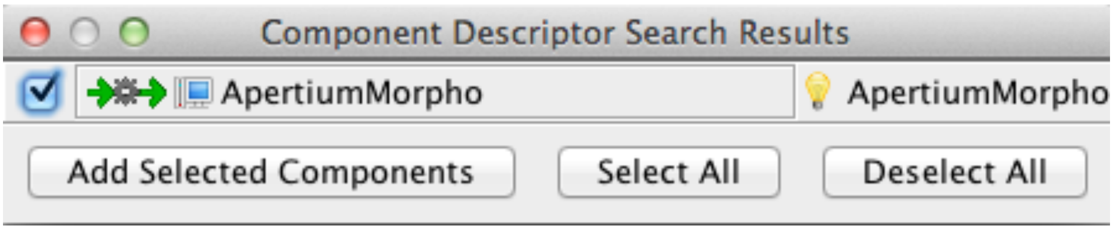

Manage Classpath Entries and/or Register Custom Component Descriptors
Add Jar File(s) to Classpath
Add Directory to Classpath
Delete Selected
Search for Component Descriptors

☑ file:/Users/paul/Documents/ApJar/ApertiumMorpho.jar
□ file:/Users/paul/Documents/workspace/SentenceSplitter/bin/SentenceSplitter.jar
□ file:/Users/paul/.U-Compare/taverna/classpath-root
□ file:/Users/paul/METANET4U-U-Compare/Apertium/lttoolbox.jar
□ file:/Users/paul/UAIC-NEW/UAIC-Lemma1.jar
□ file:/Users/paul/Documents/U-Compare_alpha/lxServiceClient.jar
□ file:/Users/paul/LX-New/LX-Tokenizer.jar
□ file:/Users/paul/LX-New/LX-Chunker.jar
□ file:/System/Library/Java/JavaVirtualMachines/1.6.0.jdk/Contents/Classes/jsfd.jar
□ file:/System/Library/Java/JavaVirtualMachines/1.6.0.jdk/Contents/Classes/classes.ja
□ file:/System/Library/Frameworks/JavaVM.framework/Frameworks/JavaRuntimeSuppor

Figure 1: External component management window

2) Click on the button "Add Jar File(s) to Classpath", and browse to the location where the file "ApertiumTranslator.jar" has been saved, and click on "Open". The will cause the file to be displayed in the list of files in the external component manager window.

3) Ensure that box next to the file "ApertiumTranslator.jar" is checked in the external component manager window. Then, click the "Search for Component Descriptors" button. A "Component Descriptors Search Results" window will appear (Figure 2).



☑ ApertiumMorpho ApertiumMorpho

Figure 2: Component Descriptor Search Results window

4) Check the box next to "ApertiumTranslator", and click on the "Add Selected Components" button. The name of the component will then appear in the library on components on the right hand side of the main U-Compare workbench window, under "Custom components", and it can then be used in workflows.

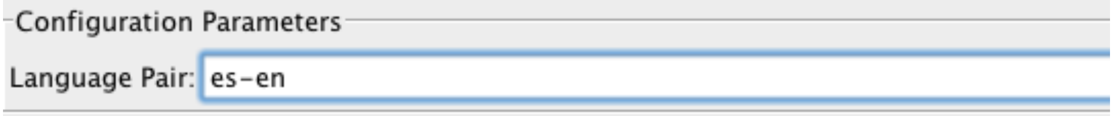
Execution instructions

Within U-Compare, the tool can be executed through inclusion in workflow. This can be done simply by dragging and dropping it onto the workflow canvas using the graphical user interface of the U-Compare workbench. See the META-SHARE record "U-Compare Workbench" for more details. Alternatively, it can be incorporated into other UIMA-based workflows, by following the documentation on the Apache UIMA site: http://uima.apache.org/

The component must be configured before use, to tell the component the source and target languages to be used during the translation. This is done by specifying a value for the "languagePair" parameter. The value of the "languagePair" parameter consists of two-letter codes for the languages, joined with a hyphen, where the first language is the source language, and the second is the target language, e.g. "pt-es" is used when Portuguese is the source language and Spanish is the target language.

Possible values of the languagePair attribute that can currently be used are as follows: "en-es", "es-en", "gl-es", "es-gl", "es-pt", "pt-es", "es-ca", "ca-es" and "eu-es". If a non-valid value is entered, then the language pair will default to "en-es", i.e., POS tagging for English will be carried out

When being run within U-Compare, the value of the "languagePair" parameter can be set by clicking on the 🔧 icon. This will cause a parameter configuration window to appear, allowing the user to enter the appropriate language pair string. A part of this window is shown in Figure 3. The value entered means that the component will be configured to carry out morphological analysis for Spanish.



Configuration Parameters

Language Pair: es-en

Figure 3: Configuration of LanguagePair parameter

NOTE: When running the comoment in U-Compare, it is recommended to place a component of type "Xmi Writer CAS consmer" as the end of the workflow. This will cause the contents of the CAS to be written to an xmi format file, which can subsequently be displayed using the external UIMA Annotation viwer.

Input/Output data formats

Input data formats

The tool operates on text that has previously been tokenised, morphologically analysed and POS tagged. The type of morphological analysis and POS tagging expected are those produced by the Apertium morphological analyser and POS tagging modules. Hence, the UIMA component corresponding to the Apertium morphological analyser (ApertiumMorpho), followed by the UIMAcomponent corresponding to the Apertium POS Tagger (ApertiumPOS) MUST be run prior to this POS tagger. This will ensure that annotations of type "ApertiumToken", which store both morphological and POS information, will be added to the UIMA Common Analysis Structure (CAS), which are required as input to the translator component.

Output data format

The result of running the Apertium POS tagger is that a new "view" of the text (called targetSofa) is created. This contains the translated text, in which each token has an associated "ApertiumToken" annotation, storing morphological and POS information about the generated token.

Integration with external tools

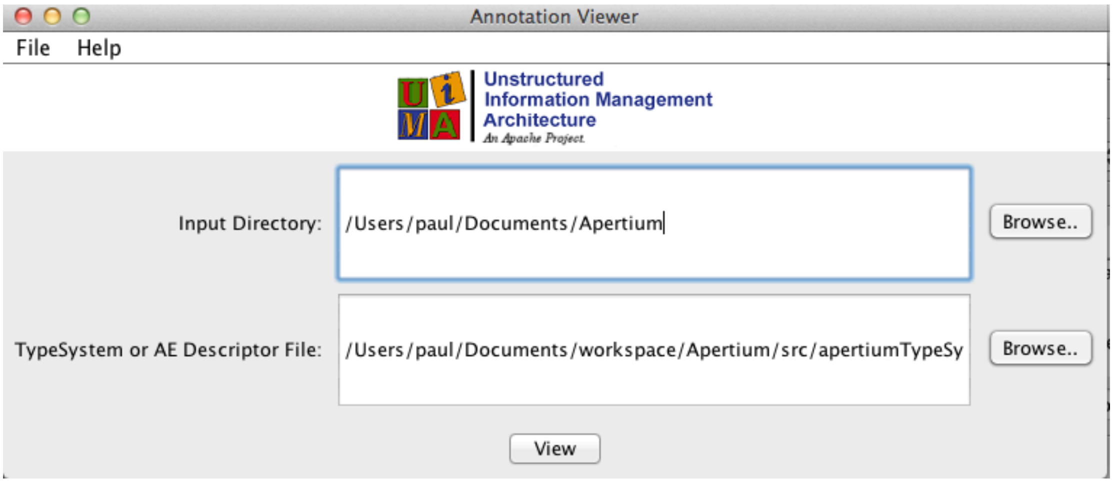
As mentioned above, the tool can only be run within U-Compare framework, or more generally, within the UIMA framework.

3. CONTENT INFORMATION

After the workflow ApertiumMorpho->ApertiumPOS->ApertiumTranslator-Xmi Writer CAS consumer has been run, the results can be inspected using the UIMA Annotation viewer. When the UIMA Annotation viewer is first started, a window appears (shown in Figure 4) in which a couple of configuration parameters must be set. These are:

1) The directory where the output XMI files are stored,

2)The type system descriptor file. This should be: apertiumTypeSystemDescriptor.xml, which can be found within the src directory of the ApertiumTranslator.jar file.

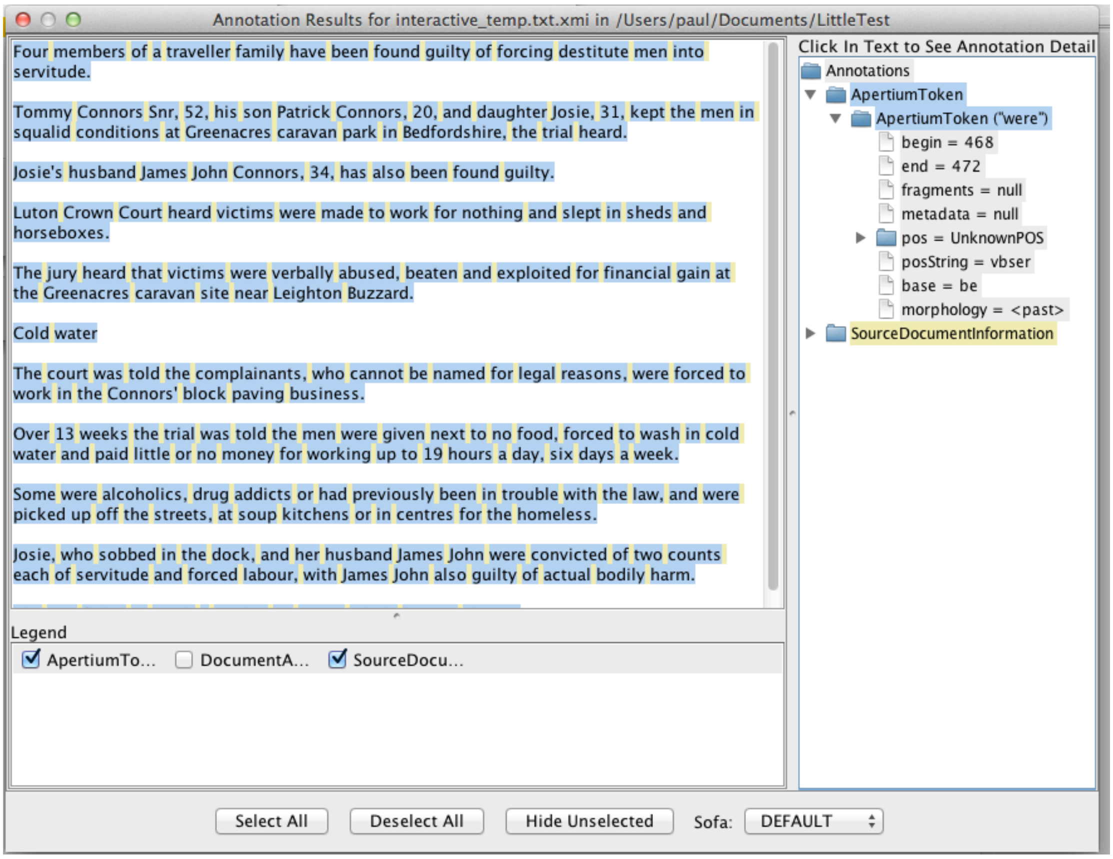


Input Directory: /Users/paul/Documents/Apertium

TypeSystem or AE Descriptor File: /Users/paul/Documents/workspace/Apertium/src/apertiumTypeSy

Figure 4: UIMA Annotation Viewer configuration

Clicking on the "View" button in the configuration window will cause a list of xmi files in the specified directory to be viewed. If an Input Text Reader has been used in U-compare, then the name of the file will be "interactive_temp.txt.xmi". Clicking on the name of an xmi file will cause an annotation viewer window containing that file to be displayed. The viewer is illustrated in Figure 5.



Annotation Results for interactive_temp.txt.xmi in /Users/paul/Documents/LittleTest

Four members of a traveller family have been found guilty of forcing destitute men into servitude.

Tommy Connors Snr, 52, his son Patrick Connors, 20, and daughter Josie, 31, kept the men in squalid conditions at Greenacres caravan park in Bedfordshire, the trial heard.

Josie's husband James John Connors, 34, has also been found guilty.

Luton Crown Court heard victims were made to work for nothing and slept in sheds and horseboxes.

The jury heard that victims were verbally abused, beaten and exploited for financial gain at the Greenacres caravan site near Leighton Buzzard.

Cold water

The court was told the complainants, who cannot be named for legal reasons, were forced to work in the Connors' block paving business.

Over 13 weeks the trial was told the men were given next to no food, forced to wash in cold water and paid little or no money for working up to 19 hours a day, six days a week.

Some were alcoholics, drug addicts or had previously been in trouble with the law, and were picked up off the streets, at soup kitchens or in centres for the homeless.

Josie, who sobbed in the dock, and her husband James John were convicted of two counts each of servitude and forced labour, with James John also guilty of actual bodily harm.

Click In Text to See Annotation Detail

- Annotations
  - ApertiumToken
    - ApertiumToken ("were")
      - begin = 468
      - end = 472
      - fragments = null
      - metadata = null
      - pos = UnknownPOS
      - posString = vbser
      - base = be
      - morphology = <past>
  - SourceDocumentInformation

Legend

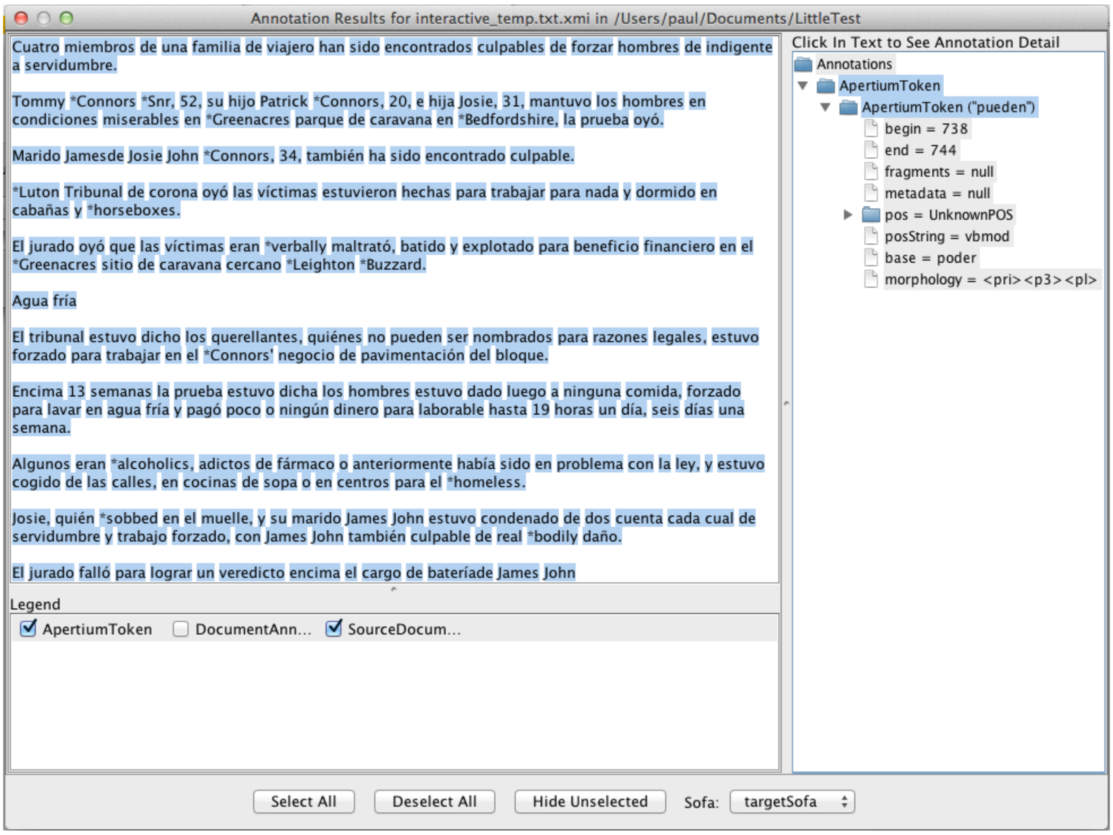
☑ ApertiumTo... ☐ DocumentA... ☑ SourceDocu...

Sofa: DEFAULT

Figure 5: UIMA Annotation Viewer: Source language display

The annotation editor in figure 5 shows a news article that has been taken from the BBC news website, http://www.bbc.co.uk/news/. In this case, this is this the source (i.e, untranslated text). As required by the Translator module, the source text has been morphologically analysed and POS-tagged, using the ApertiumMorpho and ApertiumPOS modules. Clicking on each token causes information about its associated "ApertiumToken" annotation to be displayed in the right hand pane. The information includes the begin and end offsets of the token, its part of speech (the posString attribute), the base form of the token and any morphological information that is associated with it. In the case that the token is constitutes a contraction, the values of the "posString" , "base" and "morphology" attributes will contain"+" signs, separating information about the individual words.

The translated text can be viewed by changing the value in the "Sofa" drop-down menu at the bottom of the annotation viewer to "targetSofa". Figure 6 shows the view of the same article, translated into Spanish.


Annotation Results for interactive_temp.txt.xmi in /Users/paul/Documents/LittleTest

Cuatro miembros de una familia de viajero han sido encontrados culpables de forzar hombres de indigente a servidumbre.

Tommy *Connors *Snr, 52, su hijo Patrick *Connors, 20, e hija Josie, 31, mantuvo los hombres en condiciones miserables en *Greenacres parque de caravana en *Bedfordshire, la prueba oyó.

Marido Jamesde Josie John *Connors, 34, también ha sido encontrado culpable.

*Luton Tribunal de corona oyó las víctimas estuvieron hechas para trabajar para nada y dormido en cabañas y *horseboxes.

El jurado oyó que las víctimas eran *verbally maltrató, batido y explotado para beneficio financiero en el *Greenacres sitio de caravana cercano *Leighton *Buzzard.

Agua fría

El tribunal estuvo dicho los querellantes, quiénes no pueden ser nombrados para razones legales, estuvo forzado para trabajar en el *Connors' negocio de pavimentación del bloque.

Encima 13 semanas la prueba estuvo dicha los hombres estuvo dado luego a ninguna comida, forzado para lavar en agua fría y pagó poco o ningún dinero para laborable hasta 19 horas un día, seis días una semana.

Algunos eran *alcoholics, adictos de fármaco o anteriormente había sido en problema con la ley, y estuvo cogido de las calles, en cocinas de sopa o en centros para el *homeless.

Josie, quién *sobbed en el muelle, y su marido James John estuvo condenado de dos cuenta cada cual de servidumbre y trabajo forzado, con James John también culpable de real *bodily daño.

El jurado falló para lograr un veredicto encima el cargo de bateríade James John

Legend

☑ ApertiumToken □ DocumentAnn… ☑ SourceDocum…

Click In Text to See Annotation Detail

Annotations
ApertiumToken
ApertiumToken ("pueden")
begin = 738
end = 744
fragments = null
metadata = null
pos = UnknownPOS
posString = vbmod
base = poder
morphology = <pri><p3><pl>

Figure 5: UIMA Annotation Viewer: Target language display

In the representation of the target language, each translated token also has an associated "ApertiumToken" annotation, which provides the same types of information as the source language, i.e. POS, base form and morphological information. Words that cannot be translated are marked with an asterisk (*) and remain in the source language form.

When run as part of a workflow with the Apertium morphological analyser and Apertium POS tagger on a single core machine with 8 GB RAM, the Apertium Translator takes approximately 11.3 seconds to run on the 1KB text, with the complete workflow taking around 11.74 seconds.

4. LICENCES

a) The ApertiumTranslator UIMA wrapper is licensed using the GNU General Public License version 2.0 (GPLv2). Please see "COPYING.txt" in the "Licences" directory. Please acknowledge the National Centre for Text Mining, University of Manchester if you use the ApertiumTranslator UIMA component

b) The underlying Apertium software is licensed using the GNU General Public License version 2.0 (GPLv2). Please see "COPYING.txt" in the "Licences" directory.

c) The UIMA framework is licenced using the Apache licence. Please see "Apache.txt" in the licenses directory.

5. ADMINISTRATIVE INFORMATION

Contact

Contacts for the Apertium system can be found here:
http://wiki.apertium.org/wiki/Contact

For further information regarding this UIMA wrapper for the Apertium morphological analyser module, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

6. REFERENCES

Armentano-Oller, C., Carrasco, R., Corbí-Bellot, A.,Forcada, M., Ginestí-Rosell, M., Ortiz-Rojas, S. (2006). Open-source Portuguese–Spanish machine translation. Computational Processing of the Portuguese Language ,50-59

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.