

# UIMA/U-Compare Apertium POS Tagger

## 1. BASIC INFORMATION

### *Tool name*

UIMA/U-Compare Apertium POS Tagger

### *Overview and purpose of the tool*

This tool assigns a part-of-speech tag and base form to each token in a text. It operates on text that has previously been tokenised and morphologically analysed. The POS tagger is a module of Apertium machine translation system<sup>1</sup> (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, Catalan, Galician, Portuguese and Basque.

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

The tool is provided as a UIMA<sup>2</sup> (Ferrucci et al., 2006) component, specifically as Java archive (jar) file, which can be incorporated within any UIMA workflow. However, it is particularly designed use in 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.

The Apertium morphological analyser is also available as a UIMA component (called ApertiumMorpho).

### *A short description of the algorithm*

The part-of-speech tagger determines the appropriate morphological analysis from amongst those produced by the morphological analyser model. The tagger is based on first-order hidden Markov models. The states of the Markov model represent parts of speech, and the observable parameters are ambiguity classes formed by groups of parts of speech. For the purposes of part-of-speech tagging, the fine-grained tags produced by the morphological analyser are mapped to more coarse-grained categories. See the Apertium documentation (<http://wiki.apertium.org/wiki/Documentation>) for more information.

---

<sup>1</sup> <http://www.apertium.org/>

<sup>2</sup> <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.

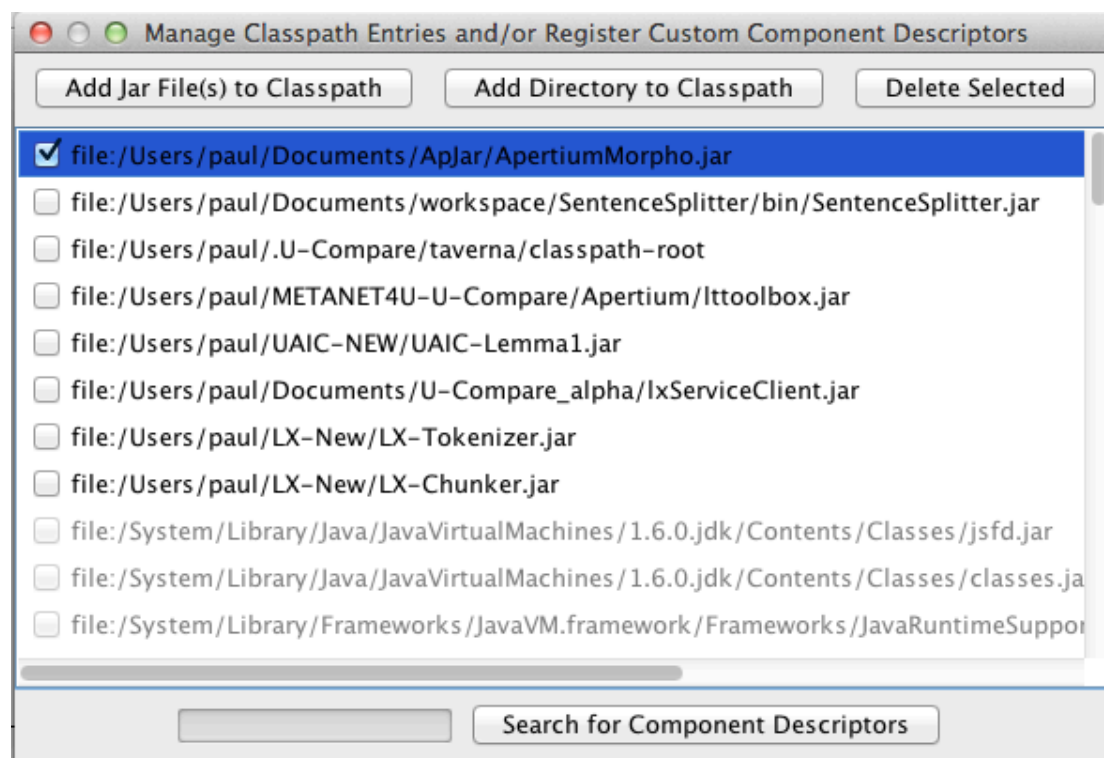
To run the tool as a UIMA component independently of U-Compare, Apache UIMA must be installed (see <http://uima.apache.org/>).

### ***Installation***

In order to run the Apertium POS Tagger 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:

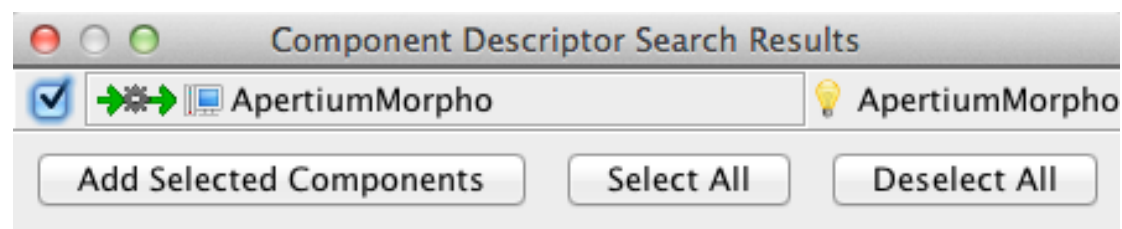
1) From the “Library” menu in the U-Compare Workbench, choose the item “Register External Components (Edit Classpath)”. This will cause a window to appear that allows external components to be managed. It is shown in Figure 1.



**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 “ApertiumPOS.jar” has been saved, and click on “Open”. This 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 “ApertiumPOS.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).



**Figure 2: Component Descriptor Search Results window**

4) Check the box next to “ApertiumPOS”, 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/>

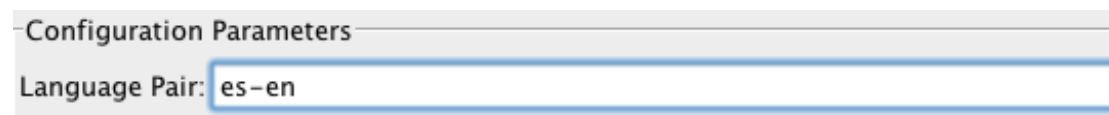
The component must be configured before use, to tell the component the languages in which the text(s) to be morphologically analysed are written. This is set by specifying a value for the “languagePair” parameter. Since this tool is a module of a machine translation system, the language data is stored in pairs, i.e., the source language and the target language. 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.

If this POS Tagger is run in a workflow without the translation module, only the source language is relevant, but the a complete language pair string must still be

specified as the value of the “languagePair” parameter, where the language to be analysed appears first in the pair string.

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.



**Figure 3: Configuration of LanguagePair parameter**

### ***Input/Output data formats***

#### ***Input data formats***

The tool operates on text that has previously been tokenised and morphologically analysed. The type of morphological analysis expected is the one produced by the Apertium morphological analyser module. Hence, the UIMA component corresponding to the Apertium morphological analyser (ApertiumMorpho) MUST be run prior to this POS tagger. This will ensure that annotations of type “ApertiumToken” will be added to the UIMA Common Analysis Structure (CAS), which are required as input to the POS tagger.

#### ***Output data format***

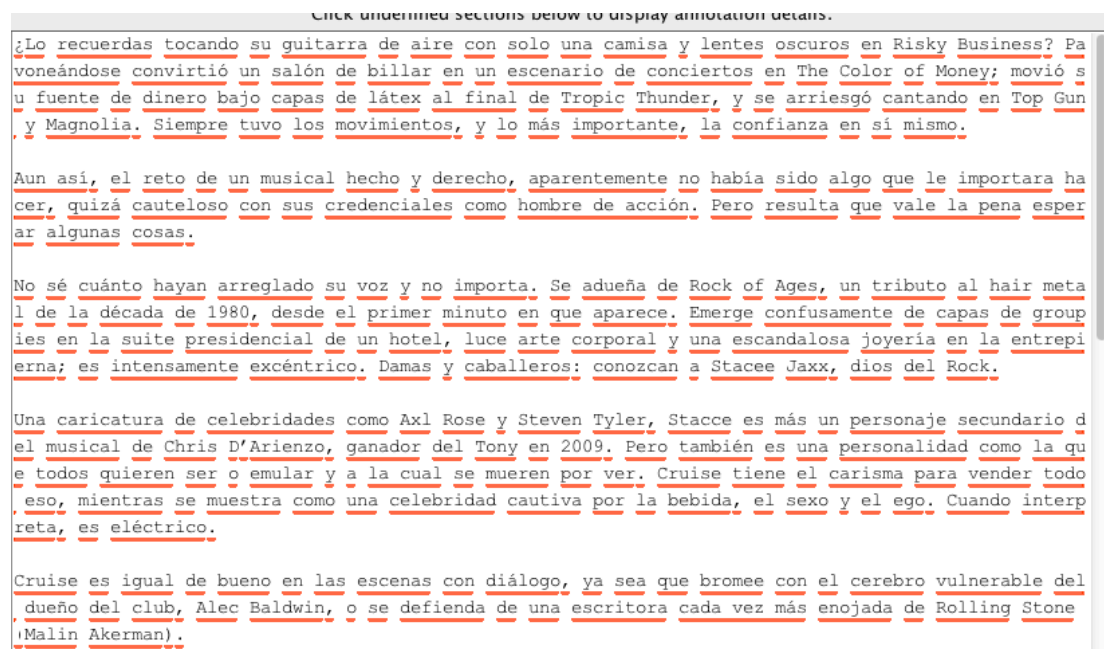
The result of running the Apertium POS tagger is that the “morphology” attribute of the annotation is updated to contain only the morphological analysis selected by the tagger (the morphological analyser may have produced several possible morphological analyses). The part-of-speech and base form of the selected morphological analysis are extracted and set as the values of the “posString” and “base” attributes of the annotation, respectively.

### ***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

Figure 4 shows the output of the tool in the U-Compare workbench. Each token (corresponding to an “ApertiumToken” annotation) is separately underlined. The sample text is taken from the CNN Español site (<http://www.cnnespanol.com>).



**Figure 4: Output of a workflow in which the Apertium morphological analyser is run prior to the Apertium POS tagger, in the U-Compare workbench.**

Figure 5 shows another part of analysis displayed in U-Compare, with the attributes of each annotation. These consist of the start and end offsets of each token, together with the part-of-speech tag assigned and the appropriate base form of the word.

Covered Text	begin	end	pos posType	posString	base
¿	0	1	lquest	lquest	¿
Lo	1	3	prn	prn	Lo
recuerdas	4	13	vblex	vblex	recordar
tocando	14	21	vblex	vblex	tocar
su	22	24	det	det	suyo
guitarra	25	33	n	n	guitarra
de	34	36	pr	pr	de
aire	37	41	n	n	aire
con	42	45	pr	pr	con
solo	46	50	adv	adv	solo
una	51	54	det	det	uno
camisa	55	61	n	n	camisa
y	62	63	cnjcoo	cnjcoo	y
lentes	64	70	n	n	lente
oscuros	71	78	adj	adj	oscuro
en	79	81	pr	pr	en
Risky	82	87			
Business	88	96			
?	96	97	sent	sent	?
Pavoneándose	98	110			
convirtió	111	120	vblex	vblex	convertir
un	121	123	det	det	uno
salón	124	129	n	n	salón
de	130	132	pr	pr	de
billar	133	139	n	n	billar
en	140	142	pr	pr	en
un	143	145	det	det	uno
escenario	146	155	n	n	escenario
de	156	158	pr	pr	de
conciertos	159	169	n	n	concierto
en	170	172	pr	pr	en

**Figure 5: Attributes for the RichToken annotations, displaying the beginning and end offsets of each token, plus the assigned POS tag and base form**

When run as part of a workflow with the Apertium morphological analyser on a single core machine with 8 GB RAM, the Apertium Tagger takes approximately .87 seconds to run, with the complete workflow taking around 1.13 seconds.

## 4. LICENCES

- a) The ApertiumPOS 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 ApertiumPOS 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](mailto: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. *Bioinformatics*, 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.