

# U-Compare Part-of-Speech Tagging Service

## 1. BASIC INFORMATION

### *Service name*

U-Compare Part-of-Speech Tagging Service

### *Overview and purpose of the tool*

This is a web service that identifies tokens in Maltese text and assigns parts-of-speech to the tokens.

### *A short description of the algorithm*

This web service is based on a UIMA-based workflow, created using the U-Compare text mining system<sup>1</sup>. The workflow was exported from U-Compare as a web service using the built in functionality (Kontonastios et al., In Press). The workflow was created as part of the work to increase the number of interoperable tools operating on different European languages (Ananiadou et al, 2011).

The workflow consists of a single UIMA-compliant tool, i.e.:

1) MLRS Part-of-Speech tagger (University of Malta)<sup>2</sup>

## 2. TECHNICAL INFORMATION

### *Software dependencies and system requirements*

This is a web service that can be run from a browser or accessed programmatically. The only basic requirement is an internet connection.

### *Installation*

There is no installation. The web service can be accessed at the following URL:

[http://nactem001.mib.man.ac.uk:8080/UCompareWebServices/POS\\_Tagging\\_MLRS](http://nactem001.mib.man.ac.uk:8080/UCompareWebServices/POS_Tagging_MLRS)

The web form available at this URL is shown in Figure 1

---

<sup>1</sup> <http://nactem.ac.uk/ucompare/>

<sup>2</sup> <http://mlrs.research.um.edu.mt/>

Examples

XML document  inline XML  stand-off annotation

Ta' sikwit nisinghu l-kelma 'diskriminazzjoni'. Forsi qajla naghtu kas ta' kemm is-soċjetà Maltija qed tkun hi stess li tweggħa lil numru ta' persuni. Hemm tfa' li minn età żgħira qed iħossuhom esklużi u mwarra. Fis-satra tal-lejl hemm żgħażaġh u adulti li qed ikollhom ibatu minhabba toroq diffiċli li qabdu għalix m'għandhomx għażla oħra.

Il-karba tal-persuni transesswali qed tistenna li tkun indirizzata hekk kif ma' kull jum li jgħaddi qed ibatu frott l-istigma li għadha tnaqqar lil pajjiż. Dan huwa wiehed mill-ftit każi li jittlob li kulhadd jerfa' parti mir-responsabilità. Int li qed taqra l-artiklu, jekk int ġenitur, jekk int haddiem jew negozjant kif ukoll int bħala politiku li tagħgen il-ligijiet, tista' twassal għal differenza f'hajjet dawn in-nies.

Li tkun transesswali mhix marda. Persuna transgendered hija persuna li tidentifika lilha nnifsiha mas-sess oppost minn dak li għet assenjata fit-twelid. Din il-persuna jista' jkun li tiddel is-sess tagħha b'operazzjoni biex tkun taqbel ma' l-identità maskili/femmini li tiegħu jew tagħha. Fid-dinja huwa kkalkulat li wiehed minn kull 500 persuna huma transgendered. Malta mhix eċċezzjoni. Għad li ma jeżistix rekord uffiċjali bin-numru preċiż ta' dawn il-persuni, l-indikaturi juru li Malta tqarreb lejn din l-istatistika dinjija.

Run

**Service Description**

Web service created by exporting UIMA-based workflow from the U-Compare text mining system.  
Functionality: Identifies tokens in plain text and assigns parts-of-speech  
Trade in workflow: MIBS POS Tagging web service (University of Malta)

**Usage**

POST request should be sent to use the service

1) text -- the value of this parameter is the text to analyze. Expected encoding is UTF-8

**References**

**Application programming interface**

```
String text = "Hello Mr. John Smith I!";
String parameters = "text=" + URLEncoder.encode(text, "UTF-8") +
"&mode=inline";
URL url = new URL(url of the service);
```

Figure 1: Web form for the web service

### Execution instructions

The web service can be executed by typing or pasting text into the online form and clicking on the “Run” button.

Alternatively, the web service can be executed from within program code, as explained in the “Usage” and “Application programming interface” boxes of the web form.

A POST request should be used to call the service. The following parameters may be used in the request:

- **text** - the value of this parameter is the text to analyze. Expected encoding is UTF-8. This parameter is obligatory.
- **lang** - This parameter sets the language of the text. If this parameter is not provided, then the value "en" will be used
- **mode** - This parameter sets the format of the annotated information returned by the service. If this parameter is not set, XML output will be produced. The two possible types of output are as follows:
  - **inline** – annotations are encoded as inline XML.

- **xml** – results are output as an XML document containing the annotations added

The following code example shows how the web service can be called from Java code:

```
//Set the input text
String text = "<Text_to_be_analysed>";
//Set the parameter string
String parameters = "text=" + URLEncoder.encode(text,
"UTF-8") + "&mode=inline";
//Create the URL connection
URL url = new
URL("http://nactem001.mib.man.ac.uk:8080/UCompareWebServi
ces/POS_Tagging_MLRS");
URLConnection connection = url.openConnection();
connection.setDoOutput(true);
//Create Output stream
OutputStreamWriter writer = new
OutputStreamWriter(connection.getOutputStream());
//write parameters to output stream
writer.write(parameters);
writer.flush();

//Read the results returned by the service
BufferedReader reader = new BufferedReader(new
InputStreamReader(connection.getInputStream(), "UTF-8"));
String line;
while ((line = reader.readLine()) != null) {
    System.out.println(line);
}
```

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

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

The input is plain text, UTF-encoded.

## Output data format

If the service is run from the web interface, then the output is visualized in the interface using colored highlights in the text to show the individual annotations, and one or more tables of information below, each corresponding to a particular type of annotation.

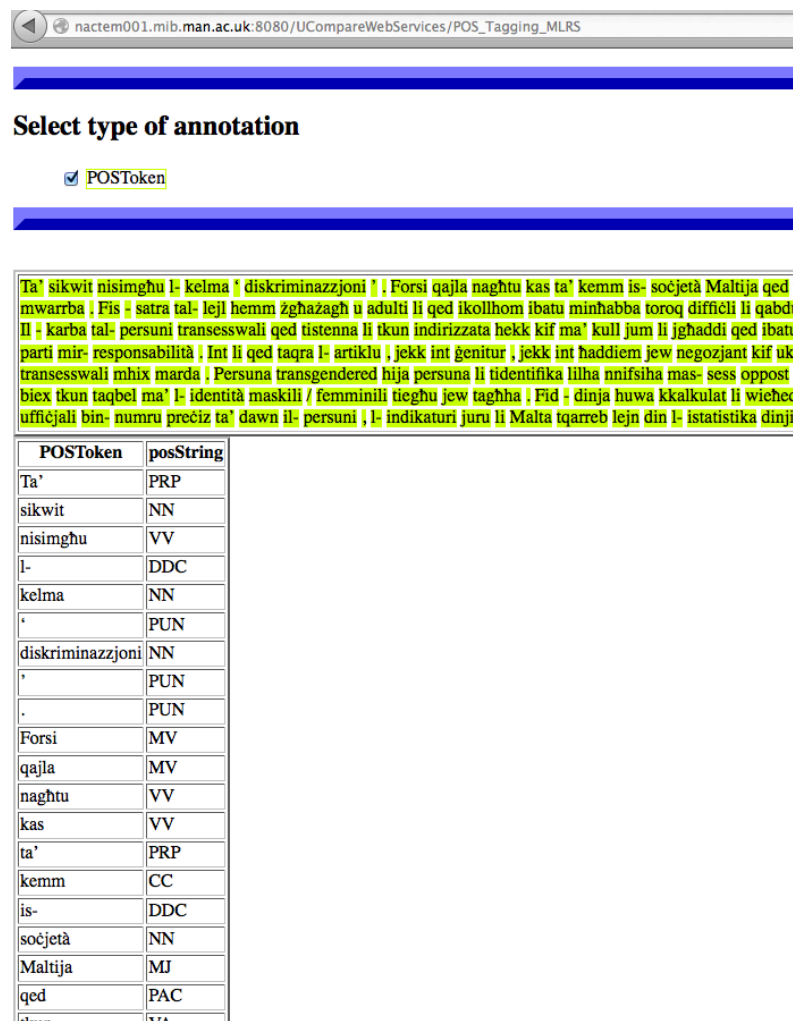
If the service is run programmatically, then the output is provided in XML format. See section 3 for an example.

## Integration with external tools

The API allows the functionality of the web service to be embedded in any application.

## 3. CONTENT INFORMATION

Using the web interface, the output of the service is visualised as shown in Figure 2.



The screenshot shows a web browser window with the URL `nactem001.mib.man.ac.uk:8080/UCCompareWebServices/POS_Tagging_MLRS`. Below the browser, there is a section titled "Select type of annotation" with a checked checkbox for "POSToken". Below this, a text snippet is displayed with various words highlighted in yellow. Below the text snippet, there is a table with two columns: "POSToken" and "posString".

POSToken	posString
Ta'	PRP
sikwit	NN
nisinghu	VV
l-	DDC
kelma	NN
'	PUN
diskriminazzjoni	NN
'	PUN
'	PUN
Forsi	MV
qajla	MV
naghtu	VV
kas	VV
ta'	PRP
kemm	CC
is-	DDC
soċjetà	NN
Maltija	MJ
qed	PAC
l-	VV

Figure 2: Visualisation of web service output

In Figure 2, the top of the screen has check boxes corresponding to each type of annotation produced by the workflow – in this case only “POSToken” annotations are present. Checking the box determines whether the token annotations are highlighted in the text below.

Below the text, the annotations added by the workflow are shown in tabular format. In Figure 2, there is a single table, which indicates the span of text covered by each “POSToken” annotation, together with the part-of-speech tag assigned (one sentence per row of the table).

An example of the XML output format, which is more suited to programmatic use, is shown in Figure 3. In the XML, the start and end offsets of each annotation in the text are encoded in the “begin” and “end” attributes, whilst the part-of-speech tag assigned is stored in the “posString” attribute.

```
- <result>
  <POSToken begin="0" end="3" posString="PRP">Ta' </POSToken>
  <POSToken begin="4" end="10" posString="NN">sikwit</POSToken>
  <POSToken begin="11" end="19" posString="VV">nisinghu</POSToken>
  <POSToken begin="20" end="22" posString="DDC">1-</POSToken>
  <POSToken begin="22" end="27" posString="NN">kelma</POSToken>
  <POSToken begin="28" end="29" posString="PUN">'</POSToken>
  <POSToken begin="29" end="45" posString="NN">diskriminazzjoni</POSToken>
  <POSToken begin="45" end="46" posString="PUN">'</POSToken>
  <POSToken begin="46" end="47" posString="PUN">.</POSToken>
  <POSToken begin="48" end="53" posString="MV">Forsi</POSToken>
  <POSToken begin="54" end="59" posString="MV">qajla</POSToken>
  <POSToken begin="60" end="66" posString="VV">naghtu</POSToken>
  <POSToken begin="67" end="70" posString="VV">kas</POSToken>
  <POSToken begin="71" end="74" posString="PRP">ta'</POSToken>
  <POSToken begin="75" end="79" posString="CC">kemm</POSToken>
  <POSToken begin="80" end="83" posString="DDC">is-</POSToken>
  <POSToken begin="83" end="90" posString="NN">soċjetà</POSToken>
  <POSToken begin="91" end="98" posString="MJ">Maltija</POSToken>
  <POSToken begin="99" end="102" posString="PAC">qed</POSToken>
  <POSToken begin="103" end="107" posString="VA">tkun</POSToken>
  <POSToken begin="108" end="110" posString="PP">hi</POSToken>
  <POSToken begin="111" end="116" posString="MJ">stess</POSToken>
  <POSToken begin="117" end="119" posString="CMP">li</POSToken>
  <POSToken begin="120" end="127" posString="VV">tweġġa'</POSToken>
  <POSToken begin="128" end="131" posString="PRP">lil</POSToken>
  <POSToken begin="132" end="137" posString="NN">numru</POSToken>
  <POSToken begin="138" end="141" posString="PRP">ta'</POSToken>
  <POSToken begin="142" end="149" posString="NN">persuni</POSToken>
  <POSToken begin="149" end="150" posString="PUN">.</POSToken>
  <POSToken begin="151" end="155" posString="EX">Hemm</POSToken>
  <POSToken begin="156" end="160" posString="NN">tfal</POSToken>
  <POSToken begin="161" end="163" posString="CMP">li</POSToken>
  <POSToken begin="164" end="168" posString="PRP">minn</POSToken>
  <POSToken begin="169" end="172" posString="NN">età</POSToken>
```

Figure 3: XML output example

### 3. LICENCE

a) The web service only is licenced NaCTeM Web Service Licence Agreement (standard non-commercial use) – see “U-Compare-Part-of-Speech-Tagging-Service-Licence.pdf” in the “licences” directory. Please contact us using the details below if you require a commercial licence.

b) The tools used in the workflow on which the web service is based may have their own licences. The NaCTeM Web Service Licence Agreement does NOT apply to these tools.

## **4. ADMINISTRATIVE INFORMATION**

### **Contact**

For further information, please contact Sophia Ananiadou:

[sophia.ananiadou@manchester.ac.uk](mailto:sophia.ananiadou@manchester.ac.uk)

## **5. REFERENCES**

Ananiadou, S., Thompson, P., Kano, Y., McNaught, J., Attwood, T. K., Day, P. J. R., Keane, J., Jackson, D. and Pettifer, S.. (2011). Towards Interoperability of European Language Resources. *Ariadne*, 67.

Kontonatsios, G., Korkontzelos, I., Kolluru, B., Thompson, P. and Ananiadou, S. (In Press). Deploying and Sharing U-Compare Workflows as Web Services. *Journal of Biomedical Semantics*.