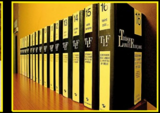
atilf ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

A Tool for adding word Annotations into TEI Files

Bertrand Gaiffe



cnrs

UNIVERSITÉ DE LORRAINE

A small example

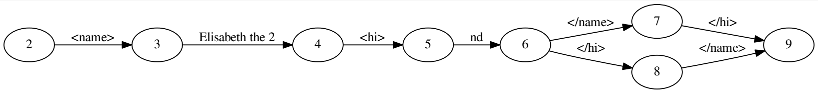
<**p**>The queen Elisabeth the 2<**hi** rend=”sup”>nd</**hi**>
celebrates her 93<**hi** rend=”sup”>rd</**hi**> birthday.</**p**>

red : <name xmlns=”http://www.tei-c.org/ns/1.0”>

blue : <num xmlns=”http://www.tei-c.org/ns/1.0”>

The problem as a graph

An extract of the graph:



Parsing with the grammar of the xml language, we get:

<**p**>The queen <name>Elisabeth the
2<**hi** rend="sup">nd</**hi**></name> celebrates her
<num>93<**hi** rend="sup">rd</**hi**></num> birthday.</**p**>

Mixer

The program (Mixer) takes as arguments:

- an XML file
- a "companion file" of the form:

index1	index2	XML elements
10	27	<name xmlns="http://www.tei-c.org/ns/1.0">
43	47	<num xmlns="http://www.tei-c.org/ns/1.0">

It proved reliable enough to POS-tags a corpus of about 5000 french novels.

Annotations alignment

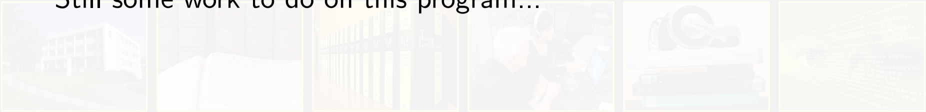
The annotation data in practice:

The	O
queen	O
Elisabeth	B-<name xmlns=”http://www.tei-c.org/ns/1.0”>
the	I-<name>
2	I-<name>
nd	I-<name>
celebrates	O
her	O
93	B-<num xmlns=”http://www.tei-c.org/ns/1.0”>
rd	I-<num>
birthday	O

The alignment program

- Takes as input an XML file and a tabular file and adds two columns corresponding to the indexes of the companion file.
- Of course, the text contents of the XML file has to match the text contents of the tabular file...

Still some work to do on this program...



The whole process

- fold the non annotated parts (ex <teiHeader xml:id=”something”/>)
- extract the text and feed some annotation process (pos, ner...). If the result is not tabular, transform it into a tabular file.
- align the tabular result with the “folded xml”
- produce the “companion file”
- apply mixer
- unfold the xml result

Conclusion and perspectives

- we do not modify the text contents of the XML file
- no schema so far (so the result is not necesserally TEI) (⟶ to do, but...
  - some more work on corrections...
  - it may happend that no valid result

)
- so far, gives "a" result somewhat randomly (could output the intersection grammar...)
- the alignment process should be improved...

to test:

https://github.com/bgaiffe/Annotations