atilf ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

Analyse10001001010001001100011010101010001111
010011et010100110001110011010100101011
1Traitement0101000110001010110011101
01001Informatique010100101100100
de0101la0100011101010001
0101Langue010111001
Française01010011
0101010Analyse
1001100011010

Frantext POS annotation

bertrand.gaiffe@atilf.fr

march 16th, 2021

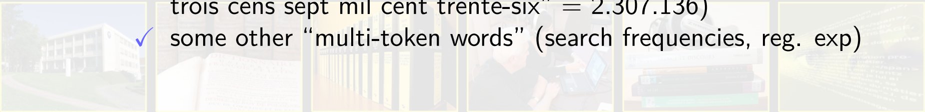
Frantext ?

- Corpus of around 5500 works (read “book”)
- mainly 20th and 19th century
- but, all periods of french language covered (not evenly)
- the whole corpus is tagged in Parts Of Speech and lemmas
- interrogation through a CQL based query language

Frantext

POS annotation

- a group of persons (including S. Ollinger, C. Benzitoun and L. Berrard)
  - ✓ defined a set of categories
  - ✓ built a gold corpus
  - ✓ learned a model with Talismane
- some particularities :
  - ✓ numbers are represented as a single token (even “deux millions trois cens sept mil cent trente-six” = 2.307.136)
  - ✓ some other “multi-token words” (search frequencies, reg. exp)

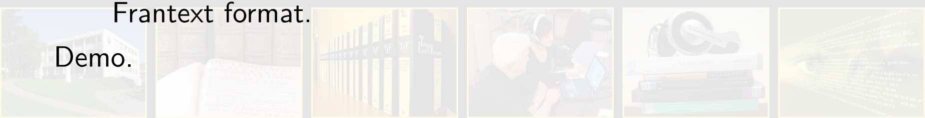


Talismane

- We use a rather old version of Talismane (A. Uriely)
- takes “sentences” as entry and produces a conll style output
- takes some time...
- in fact, Talismane is a dependency parser, we use only the POS tagging stage.

How texts enter Frantext

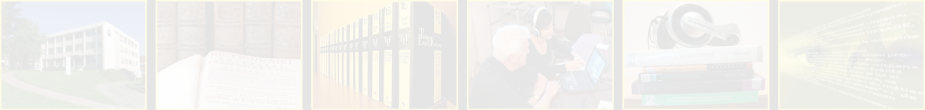
- I. Clément (and G. Toubiana) scan and OCR the texts
- they encode them in TEI (Text Encoding Initiative) using an XML editor
- they use a script that tag the texts using Talismane in the background
- and finally they transform the resulting annotated TEI into Frantext format.



Demo.

The problems

- Talismane (the version we use) relies too much on its sentence splitting
- Natural Language Processing progressed a lot
- the annotation process was designed for texts pots 1850 only



An attempt at Bert'izing the annotation

- the gold corpus is divided (as usual) in train, dev and test
- learned a "camembert" based POS tagger
- 97% on test (talismane does slightly better but was trained on test also)
- comparison between the two annotations
  - ✓ R467
  - ✓ S437

What to do next ?

- tagging persNames and placeNames ?
- design models for other periods of the french language
  - ✓ {vieux Camem | Hau}bert a bert model for middle age french.

Contraints :

- have to scale to Frantext (5500 texts),
- have to provide a tool for the people that enter new texts into Frantext.