UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Extraction and annotation of “location names”

Tita Kyriacopoulou, Claude Martineau, Markarit Vartampetian

Mars 2019

LIGM has been conducting fundamental and applied research in the field of Natural Language Processing

- Since 2001, LIGM has developed the international open-source platform Unitex/Gramlab, a tool used for corpus processing and construction of linguistic resources. It constitutes a multilingual (22 languages: French, English, Greek, Serbian, Russian, Arabic, Thai, Korean, and soon Chinese in collaboration with the UPEM's DUT Informatique) multiplatform (Windows, Mac, Linux). **http://unitexgramlab.org**

Downstream applications

- Noise-tolerant extraction (patent pending)
- Citations in scientific and technical texts (Cixplorer project)
- Semantic annotation in literary and scientific text (Animalhumanité and Métamorphosis projects)
- Interoperability between Unitex/GramLab and TreeCloud (Eclavit project)
- Opinion expressions (Doxa project)

Our goal is:

- Extraction of **relevant** information from unstructured data (corpus), namely answering to:

  Who did What When Where How =

  Named Entities

- Methods : Symbolic and statistical
- Software : Unitex

Within this context, my student C. Martinez produced in his thesis results that are currently patent pending: *Fault-tolerant information extraction*

Named Entity Extraction

This task consists of automatically recognising the NE in corpus, extracting and classifying them into categories such as Person, Location, Organization.

As indicated by Sagot (2012), we can distinguish two ways of identifying an entity, either intrinsic (*"France denotes a place"*), or in context (*"France signed the treaty"*, where *France* can be recognised as an organization).

Since 2014…

We have been working on named entities by creating:

- ressources (dictionaries and dictionary graphs),
- recognition and verification grammars (automata).

It is time to move on to supervised learning for named entity recognition

- Named entities annotation in literary and scientific texts
- Location names

It is worth noting that this choice was made thanks to the University Gustave Eiffel (the University of Marne-la-Vallée is part of it) whose topic will be "The City" and will be created by 1st january of 2020.

But what is exactly a Location?

Paris1

Marie va à **Paris**

*Mary is going to **Paris***

Paris va organiser les jeux olympiques en 2024

***Paris** will organise the Olympic Games in 2024*

But what is exactly a Location?

La **maire de Paris** va se présenter aux prochaines élections

*The **mayoress of Paris** will be running in the next elections*

Marie habite **Rue de Paris**

*Marie lives in Rue de **Paris***

Paris2

Paris a battu Lille 2-0

***Paris** beat Lille 2-0*

Challenges

1. Do we intend to recognise a Named Entity or an Extended Named Entity ?

Gaio and Moncla (2017) have used the concept of Extended Named Entity (ENE). Based on the Jonasson's definition, an ENE refers to an entity built with a proper name (***Rue de Paris***) and may be composed of one or more concepts (***La maire de Paris)***

Challenges

1. NE or ENE ?

We believe that when we have to extract and annotate locations, it is evident that we have to recognise the ENE rather than the NE since:

- Paris and Rue de Paris are two different places, in particular Rue de Paris can be located in Lille and Paris can also indicate the football team
- there are many ambiguities *La maire de Paris (is a PERson)*

Challenges

2. Do we intend to recognise a location or a locative complement ?

Paris est une belle ville
Paris *is a beautiful city*

Claire va **à Paris**
Claire is going to ***Paris***

Claire dort **à Paris**
Claire sleeps in ***Paris***

Challenges

A locative complement is an argument or a modifier?

- Claire se repose **à Paris**/ Claire va **à Paris**
  Claire is resting in Paris / Claire is going to Paris
  Claire se repose. / * Claire va.
  *Claire is resting. / *Claire is going.*

- **À Paris**, Claire se repose. / ***À Paris**, Claire va.
  In Paris, Claire is resting/ To Paris, Claire is going

Challenges

- **Argument**→ Depends on the predicate
- **Modifier** → Does not depend on the predicate

➔ in order to make this distinction a syntactic analyser is necessary

Challenges

- **All these problems arise due to the large number of ambiguities:**

  *C'est là que Simone Evangelista intervient dans son concept avec un iPhone 5,5 pouces et 4GB de **RAM**, ainsi qu'une batterie plus performante de 3000 mAh pour faire tourner le processeur A8.*

  *En 2007, Aigle Azur proposait 30 destinations régulières au départ de plusieurs villes de France, et lançait des vols réguliers **Paris-Orly**-Djerba, **Paris-Orly**-Rimini et Marseille-Sal (Cap-Vert).*

  ***Allons** en avant !*
  ***Viens** !*

Methods

- Graphs in Unitex → Automatic Annotation
- Results of the automatic annotation → Brat
- Manual Corrections / Manual annotation in Brat

Corpus

- Ambiguous Test Corpus

  *La porte-parole **d'Aéroports de Montréal** assure, pour sa part, ...*

  The spokeswoman for Aéroports de Montréal ensures, …

  *Quand on arrive à **l'Aéroport de Montréal**, ..*

  When we arrive at the Aéroport de Montréal…

  *Trouvez votre vol **vers Araxa** au meilleur prix sur Skyscanner!*

  Find your flight to Araxa at the best price with Skyscanner!

- « Le Tour du monde en 80 jours »

  Around the World in Eighty Days

- « 20.000 lieues sous les mers »

  Twenty Thousand Leagues Under the Sea

Test Corpus

Corpus:Corr_corpus_organisations_Amina_FR_utf8.txt

- 1756 sentence delimiters
- 100642 (8467 diff) tokens
- 43664 (8404) simple forms
- 3742 (10) digits
- DLF: 10082 simple-word lexical entries
- DLC: 2039 compound lexical entries
- ERR: 1405 unknown simple words

20.000 lieues sous les mers

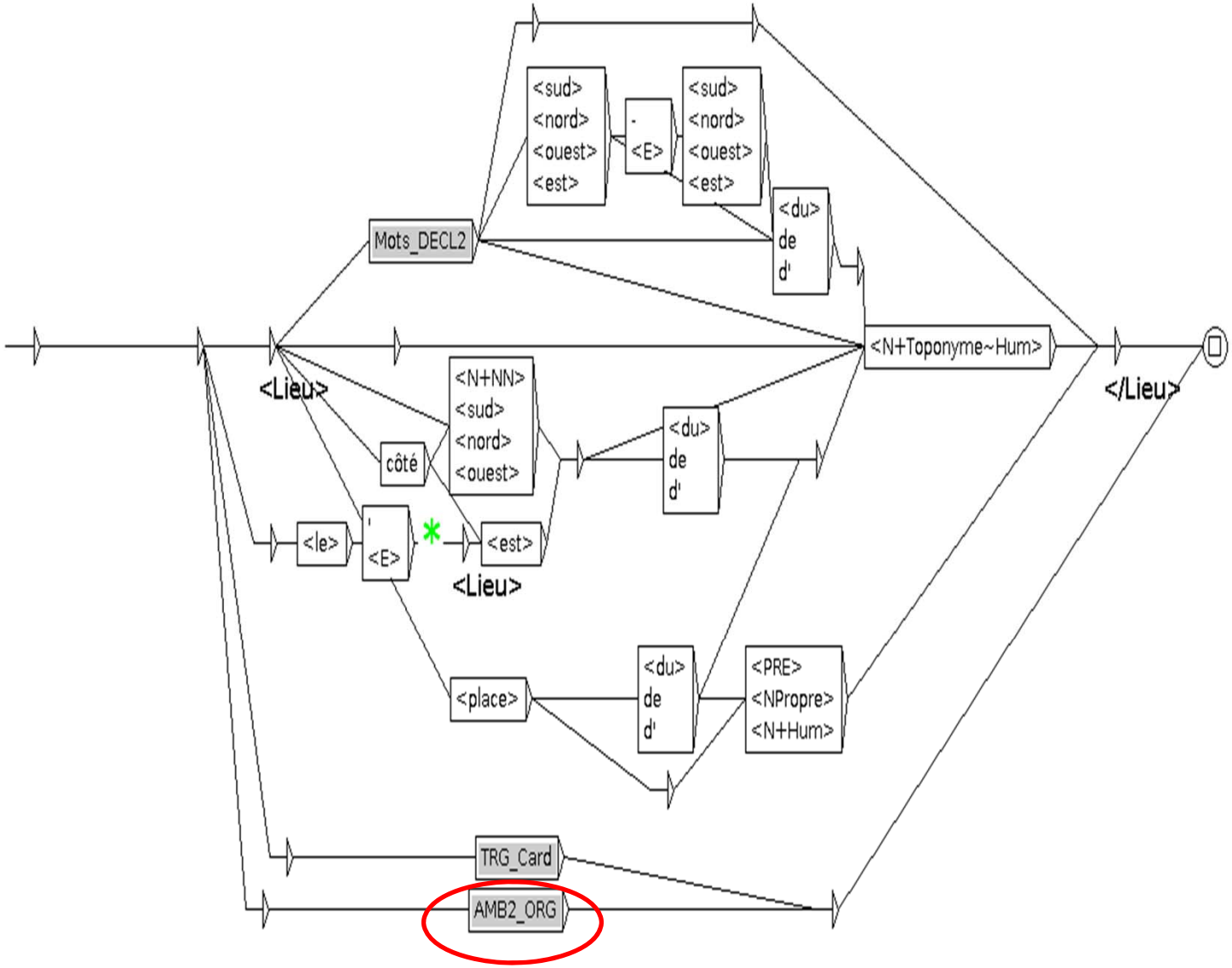
Corpus:20000_lieues_sous_les_mers_FR_utf8.txt

- 7349 sentence delimiters
- 339425 (15301 diff) tokens
- 149007 (15268) simple forms
- 1047 (10) digits
- DLF: 20654 simple-word lexical entries
- DLC: 3123 compound lexical entries
- ERR: 2161 unknown simple words

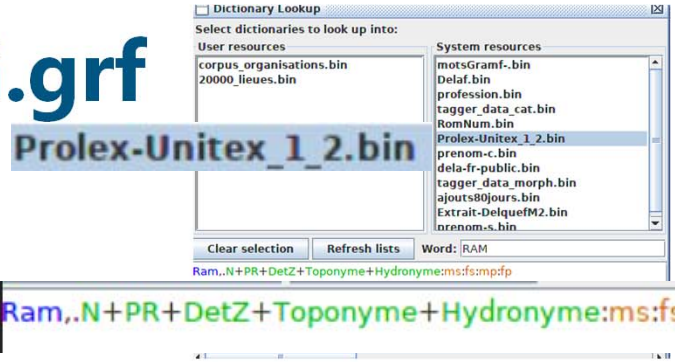
Le Tour du monde en 80 jours

Corpus: 80_jours.txt

- 3652 sentence delimiters
- 165239 (9452 diff) tokens
- 71859 (9422) simple forms
- 438 (10) digits
- DLF: 13229 simple-word lexical entries
- DLC: 2099 compound lexical entries
- ERR: 3156 unknown simple words

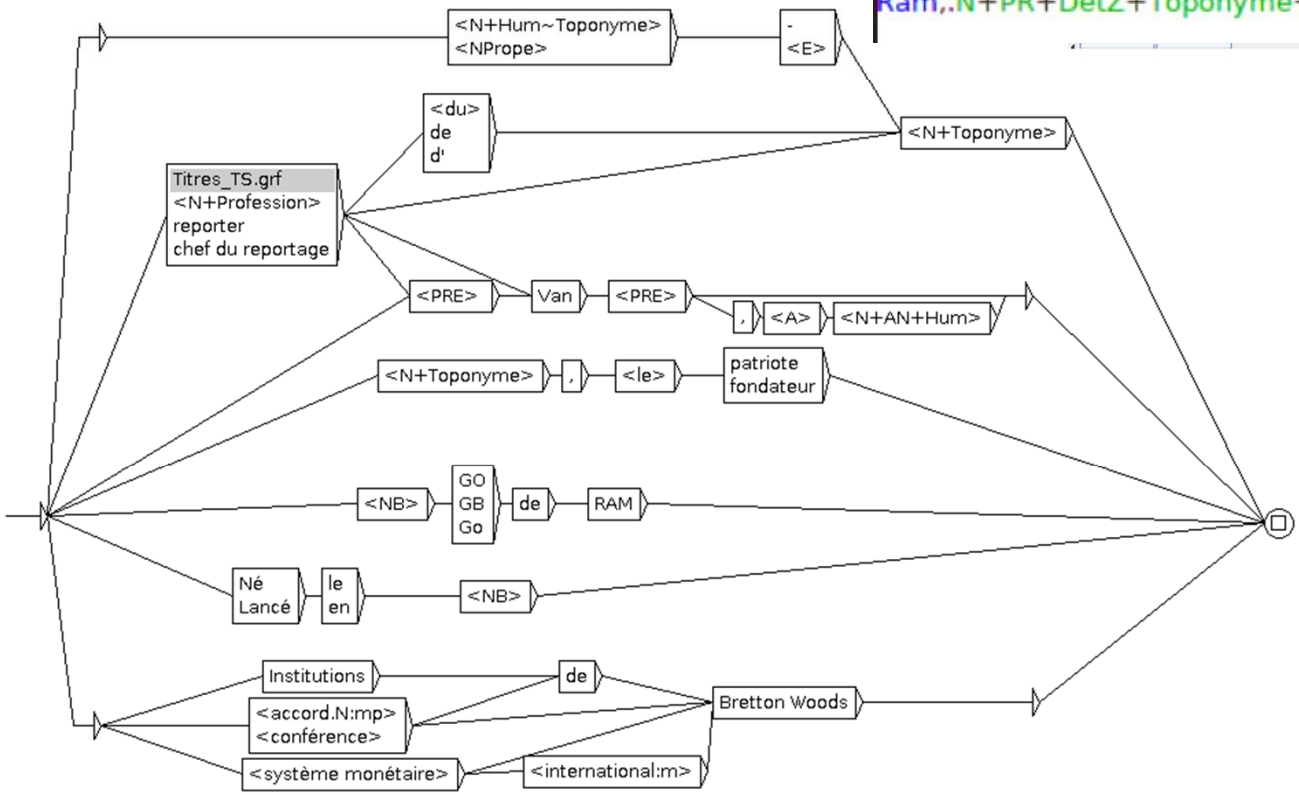

<sud>
<nord>
<ouest>
<est>
-
<E>
<sud>
<nord>
<ouest>
<est>
<du>
de
d'
Mots_DECL2
<N+Toponyme~Hum>
<Lieu>
</Lieu>
<N+NN>
<sud>
<nord>
<ouest>
côté
<du>
de
d'
<le>
'
<E>
*
<est>
<Lieu>
<place>
<du>
de
d'
<PRE>
<NPropre>
<N+Hum>
TRG_Card
AMB2_ORG

Sub-Graph AMB2_ORG.grf

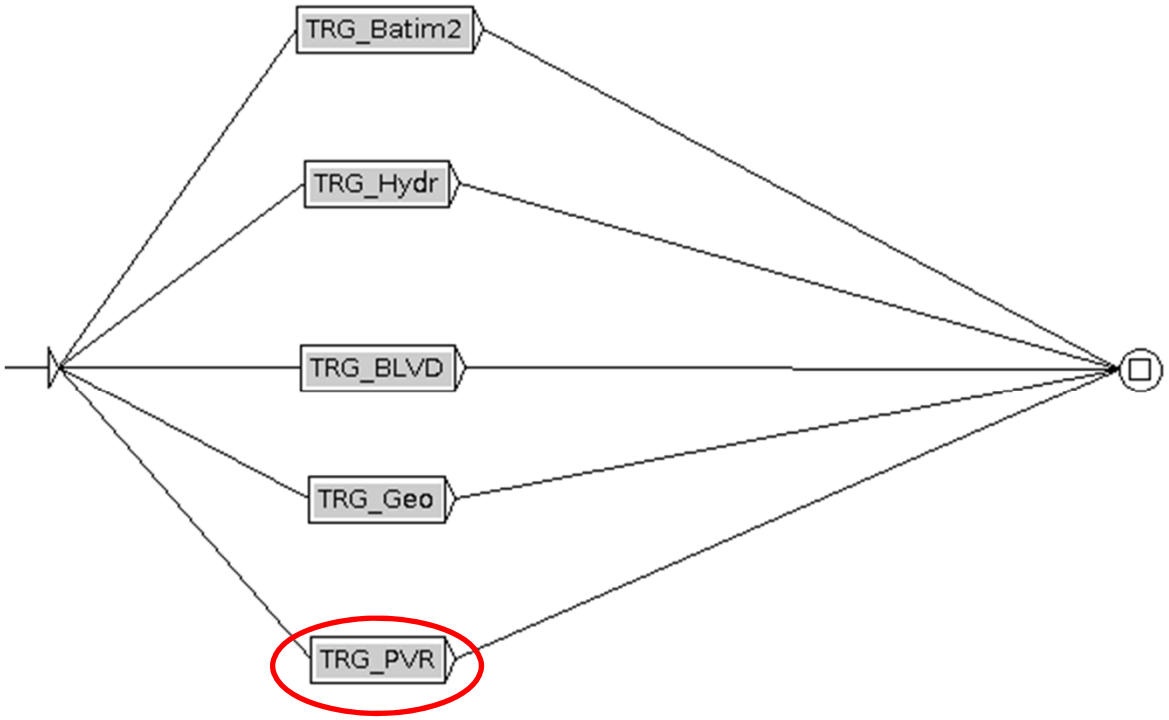


Prolex-Unitex_1_2.bin

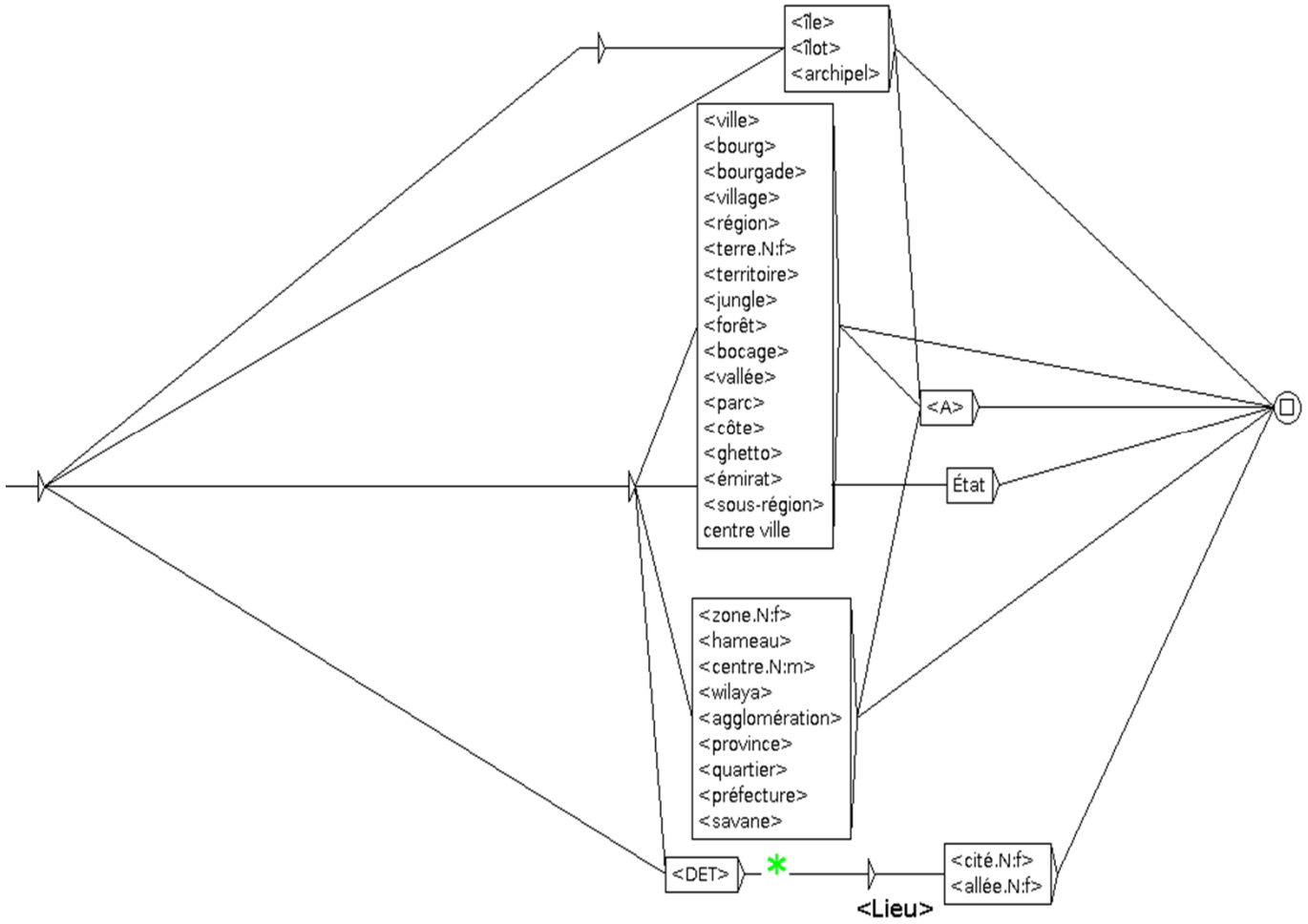
Ram,.N+PR+DetZ+Toponyme+Hydronyme:ms:fs:mp:fp



General Graph



Sub-Graph TRG_PVR.grf



Procedure

- Annotation of arguments only

(2 annotated corpora) -> Evaluation

- Annotation of all the locatives (arguments and modifiers)

  (2 annotated corpora) -> Evaluation

Evaluation – Gemini Tool

RESULTAT_AUT1_MAN1A

```
fengyuhengdeMacBook-Pro:result_10JUIL FENGYuheng$ java -jar geminiV2.1.jar -xmlf
ile1 ORG_AUT1.xml -xmlfile2 ORG_MAN1A.xml -visualize=Lieu -CSV


Similarity score (weak precision) : 0.24898694
Similarity score (strict precision) : 0.23097704
Similarity score (weighted precision) : 0.240175
Similarity score (weak recall) : 0.9600694
Similarity score (strict recall) : 0.890625
Similarity score (weighted recall) : 0.92609143
Similarity score (weak F-measure) : 0.39542368
Similarity score (strict F-measure) : 0.3668216
Similarity score (weighted F-measure) : 0.38142914
```

RESULTAT_AUT2_MAN2

```
fengyuhengdeMacBook-Pro:result_10JUIL FENGYuheng$ java -jar geminiV2.1.jar -xmlf
ile1 ORG_AUT2.xml -xmlfile2 ORG_MAN2.xml -visualize=Lieu -CSV


Similarity score (weak precision) : 0.82061404
Similarity score (strict precision) : 0.70307016
Similarity score (weighted precision) : 0.76315707
Similarity score (weak recall) : 0.86300737
Similarity score (strict recall) : 0.73939115
Similarity score (weighted recall) : 0.8025822
Similarity score (weak F-measure) : 0.841277
Similarity score (strict F-measure) : 0.72077346
Similarity score (weighted F-measure) : 0.78237325
Time to calculate the score: 748 ms
fengyuhengdeMacBook-Pro:result_10JUIL FENGYuheng$
```

Conclusion

Recognition and annotation of locations

- Need for a syntantic analyser → distinction → modifiers & arguments

+

« *La Centrafrique traverse une grave crise politique et humanitaire depuis mars 2013.* »

& locative complements

- Satisfactory results, however we need to annotate more corpora and process more languages.