GROBID for Humanities When engineering meets History

Charles Riondet, Luca Foppiano

GROBID for Humanities

When engineering meets History

Charles Riondet, Luca Foppiano
ALMAnaCH, Inria Paris

Acknowledgement

Anne Baillot (Centre Marc Bloch, Berlin) → German literature and History, Mastermind

Hector Martinez Alonso (ALMAnaCH, Inria Paris) → POS tagging, NLP stuff

Context

ALMAnaCH = Automatic Language Modelling and Analysis & Computational Humanities, joint EPHE-Inria team

A strange mixture of people with different skills working side by side:

- Computer engineering → data mining, information extraction
- NLP experts → Semantic and Syntactic analysis, Parsing, etc..
- Digital Humanities (History and Literature background) → data modelling, textual analysis, digital philology

Context

ALMAnaCH = Automatic Language Modelling and Analysis & Computational Humanities, joint EPHE-Inria team

A strange mixture of people with different skills working side by side:

- Computer engineering → data mining, information extraction
- NLP experts → Semantic and Syntactic analysis, Parsing, etc..
- Digital Humanities (History and Literature background) → data modelling, textual analysis, digital philology

Why not combining all the skills available around?

Who are we?

Charles Riondet - Historian

WW2
data modeling

Luca Foppiano - Engineer

Software engineering, Data mining,
Machine learning, Knowledge discovery

Original hermeneutical process

In this situation, we started with some empirical tries

Person 1: "Hey, I have data"
Person 2: "Hey, I have tools"

Original hermeneutical process

In this situation, we started with some empirical tries

Person 1: "Hey, I have data"
Person 2: "Hey, I have tools"

Person 1 and 2 (ecstatic): ***"Let's try to use tools on some random data (YOLO)"***

Original hermeneutical process

But with a research question, it worked a little better:

Person 1: "Hey, I have data."
Person 2: "Hey, I have tools."

Person 1: ***"Waow, I also have an idea of what I want to do"***
Person 2: ***"Maybe my tools could help you shedding some light"***

Original hermeneutical process

Next step, we started to think about a win-win strategy"

Person 1: "Hey, I have data and a research question"
Person 2: "Hey, I have tools"

Person 1: ***"Your tools might need new data and use cases to improve (ensure genericity and cross-domain applicability)"***

Original hermeneutical process

Finally…

Person 1: "Hey, I have data."
Person 2: "Hey, I have tools."
Person 3: "Hey guys, I also have data, tools and some nice research questions"
Person 4: "I want to be part of the group, please"

*Choir : "**Let's create a research project all together"***

ECRPER project (ANR/DFG, 2018-2021)

Franco-german personal writings in wartime (19th-20th century)

With ALMAnaCH (Inria Paris) - DHI (Paris) - Centre Marc Bloch (Berlin) - IEP Lille

Analysing German and French diaries and letters written during the French-German wars since the 19th century:

- Napoleonic Wars
- War of 1870
- First World War
- Second World War

Funding is not granted yet, but we're in a hurry, so we already started to work.

ECRPER in short

1) Diachronic and synchronic analysis:
- how the Self and the Other are being represented
- how relationship to events is being elaborated through personal writings

2) Bring together different tools, sets up a solid editorial and hermeneutical workflow and make it available for further research.
- OCR
- Annotation
- Edition
- publication

ECRPER in short

1) Diachronic and synchronic analysis:
- **how (the Self and) the Other are being represented**
- how relationship to events is being elaborated through personal writings

2) Bring together different tools, sets up a solid editorial and hermeneutical workflow and make it available for further research.
- OCR
- **Annotation**
- Edition

Representation of the other

The representation of the other in the context of war discourses

- How Germans see the French and reciprocally.
- Broaden to all the mentions of the conflicts actors (Nations, organisations, persons, person types, …)

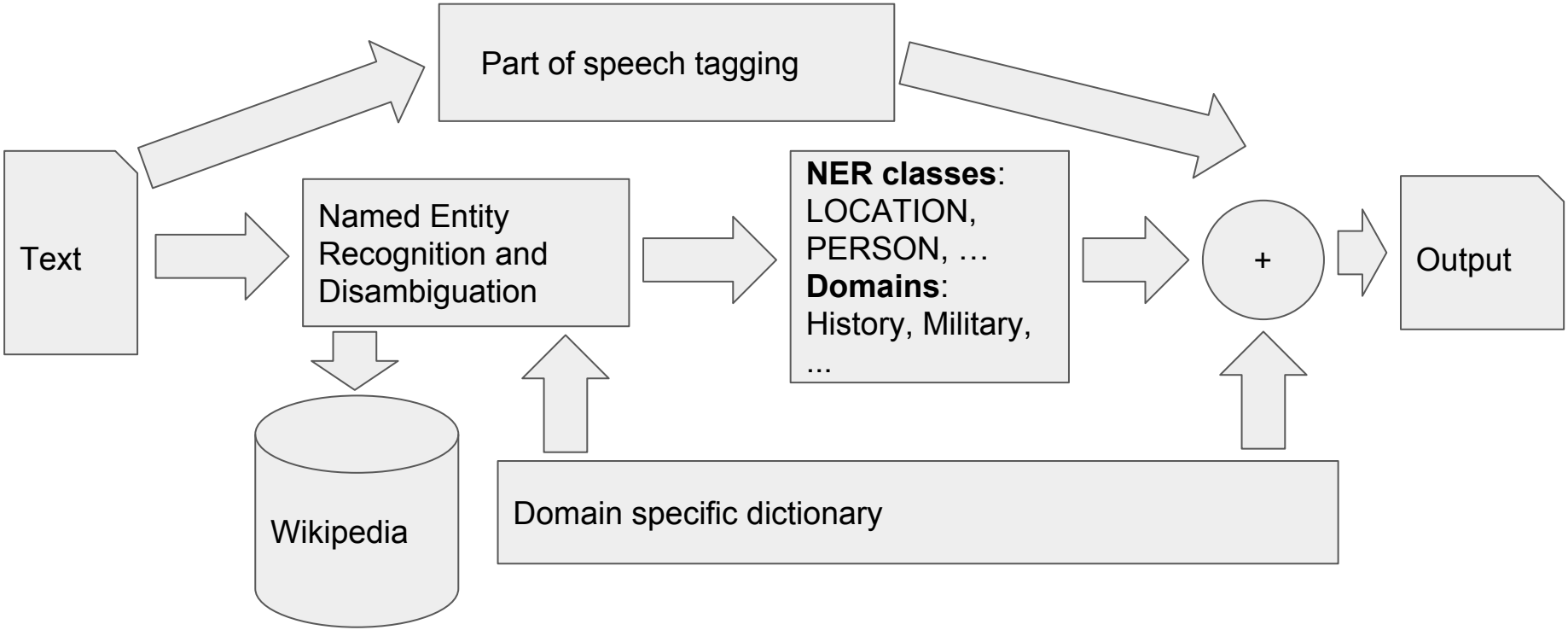
- All instances appearing in the discourse will be modelized and brought in relation to one another.
- Diachronic analysis: appearance and disappearance of a mention, semantic evolution, spelling variations, ...

Representation of the other: NLP approach

- The other/the actors of the conflicts
  → a set of Named-entities

- Recognize structural discourse elements in French and in German
  → Part of speech tagging
  - Opinion mining → lexical analysis of the context of the mentions
  - likely to vary according to time and nation + social origin of the writer.

The big picture



(N)ERD, a tool for Named Entity Recognition and Disambiguation

GROBID (N)ERD is a tool for recognise and extract named entities from text or PDF documents. They are then resolved (disambiguated) against Wikipedia.

E.g. The president Washington went to Washington to celebrate his birthday.

Normalized: **George Washington**

conf: *0.7604841161333109*

George Washington (; – , 1799) was an American politician and soldier who served as the first President of the United States from 1789 to 1797 and was one of the Founding Fathers of the United States. He served as Commander-in-Chief of the Continental Army during the American Revolutionary War, and later presided over the 1787 convention that drafted the United States Constitution. He is popularly

Type: **LOCATION**

Sense: **municipality/N1**

Normalized: **Washington, D.C.**

Domains: **Administration**

conf: *0.9376965404365496*

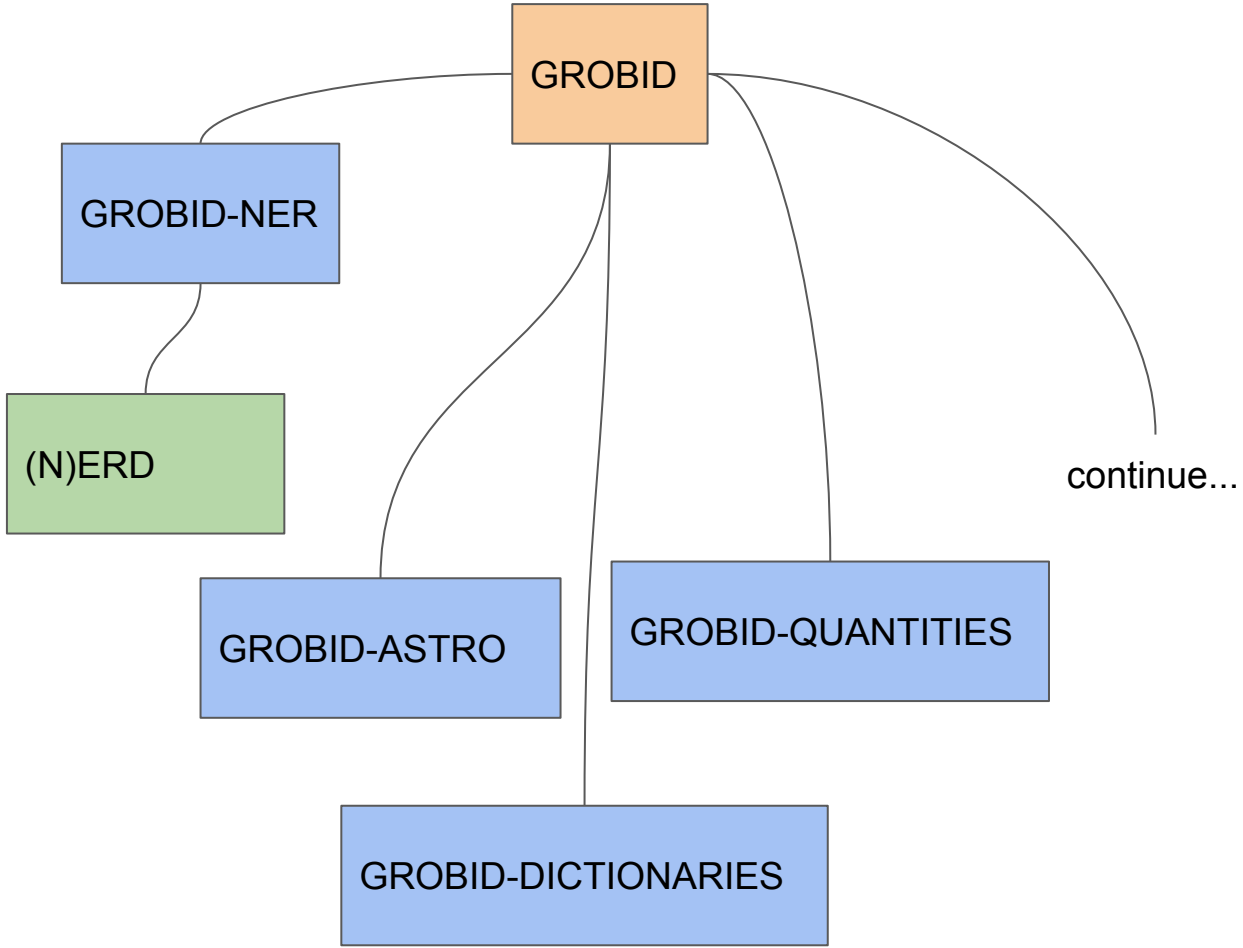
Washington, D.C., formally the **District of Columbia** and commonly referred to as **Washington**, **"the District"**, or simply **D.C.**, is the capital of the United States, founded on July 16, 1790. The U.S. Constitution allows for the creation of a special district to serve as the permanent national capital. The District is therefore not a part of any U.S. state and is instead directly overseen by the

Overview of the *GROBID Family*

GROBID (or Grobid) means GeneRation Of BIbliographic Data.

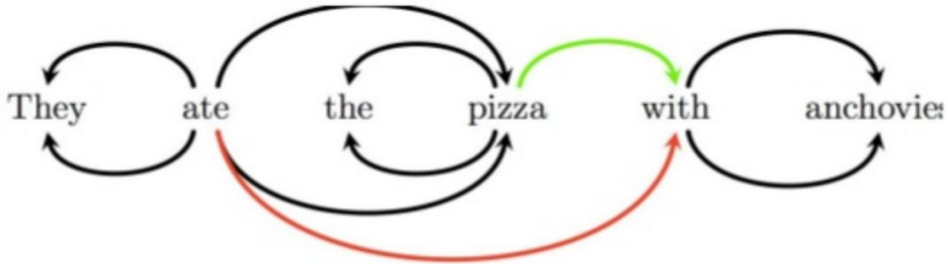
Written by Patrice Lopez and released open source (Licence Apache 2.0).

Available on Github:
- http://github.com/kermitt2/grobid
- http://github.com/kermitt2/grobid-ner
- http://github.com/kermitt2/nerd
- [...]

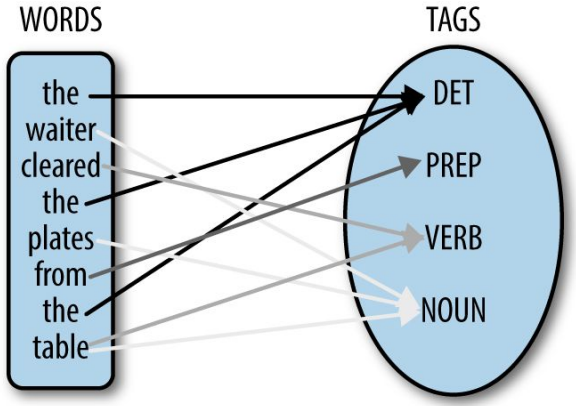


Part of speech tagging (in one slide)

Syntactic analysis of sentences, produces dependency tree and tags (noun, verb, adj, etc.)




```
nsubj(ate-2, They-1)
root(ROOT-0, ate-2)
det(pizza-4, the-3)
dobj(ate-2, pizza-4)
prep(ate-2, with-5)
pobj(with-5, anchovies-6)
```



[1] https://www.slideshare.net/vseloved/crash-course-in-natural-language-processing-2016
[2] http://naviglinlp.blogspot.fr/2017/04/lecture-7-part-of-speech-tagging.html

The corpus: WW2 diaries written in French

Journal de Léo Hamon (Archives nationales, 72AJ42)
→ French Lawyer of Russian origin, one of the leaders of Parisian Resistance.
→ underground daily life, reports on meetings, comments on the course of the war and on the organization of the Resistance and on the preparation of the seizure of power in Paris.

Journal d'Henri Chabasse (Musée de la Résistance nationale, 13/3907b)
→ Nationalist middle-class Parisian, not involved in the Resistance nor in the Collaboration.
→ Daily life but mostly comments on the course of the war and French political situation, from D-Day to fall 1944 (+ an entry related to Hiroshima bombing)

Need a specific dictionary

Context based expressions

Ex: "Souris grises" → German army female auxiliaries (not the animal)

Diary internal terminology

Ex: "les cocs" → Members of the communist party (not the animal badly written)

Unnormalized spelling

Ex: Gaulisme → Gaullisme

Metonymies

Ex: Vichy → Régime de Vichy (not Vichy town)

Need a specific dictionary

Because Wikipedia doesn't know everything.

Sources:

- Marcot et al., Dictionnaire historique de la Résistance, Paris, R. Laffont, 2006.
- Rue de la Mémoire, Volksbund Deutsche Kriegsgräberfürsorge e.V., 2016, Parler de l'Histoire et de la Mémoire. Première et Deuxième guerre mondiale. Glossaire Franco-allemand.
- Specific terms found in the diaries

Need a specific dictionary

Modeled in TEI-TBX
(thanks to Stefan Pernes)

- Machine-readable
- Standard

```
<conceptEntry xml:id="c_217">
   <langSec xml:lang="fr">
      <form type="lemma">
         <orth>Parti communiste français</orth>
      </form>
   </langSec>
</conceptEntry>

<conceptEntry xml:id="c_218">
   <descrip type="subordinateConceptPartitive" target="#c_217"/>
   <langSec xml:lang="fr">
      <form type="lemma">
         <ref target="#Hamon_AN72AJ42"/>
         <orth>communiste</orth>
      </form>
      <form type="variant">
         <ref target="#Hamon_AN72AJ42"/>
         <orth>coc</orth>
      </form>
   </langSec>
</conceptEntry>
```

4 main steps

1) Extract Named Entities
2) Apply domain specific dictionary
3) POS tagging
4) Mixing up everything (POS and NER)

Workflow (1) - Extract Named Entities

Nous parlons du procès Pucheu. La question est plus actuelle. (...)

J'indique qu'à mon avis tout ce procès a été mal conduit - il fallait (...) proclamer que devant des crimes inouïs, (...) la nation prenait une décision politique, immoler les hommes de la haute trahison : "Jetons à l'Europe, en défi, une tête de roi", jetons aux combinards de Vichy et de Washington, en défi, une tête de traître. (...) Je vois Yves qui m'avait cité l'impatience d'Alger devant le cas Pucheu comme une illustration de la crise du Gaulisme.

Journal de Léo Hamon, March 14th 1944 entry

Workflow (2) - Apply domain specific dictionary

Nous parlons du procès Pucheu. La question est plus actuelle. (...)

Régime de Vichy

État Français

United States of America

aux combinards de Vichy et de Washington, en défi, une tête de traître. (...)

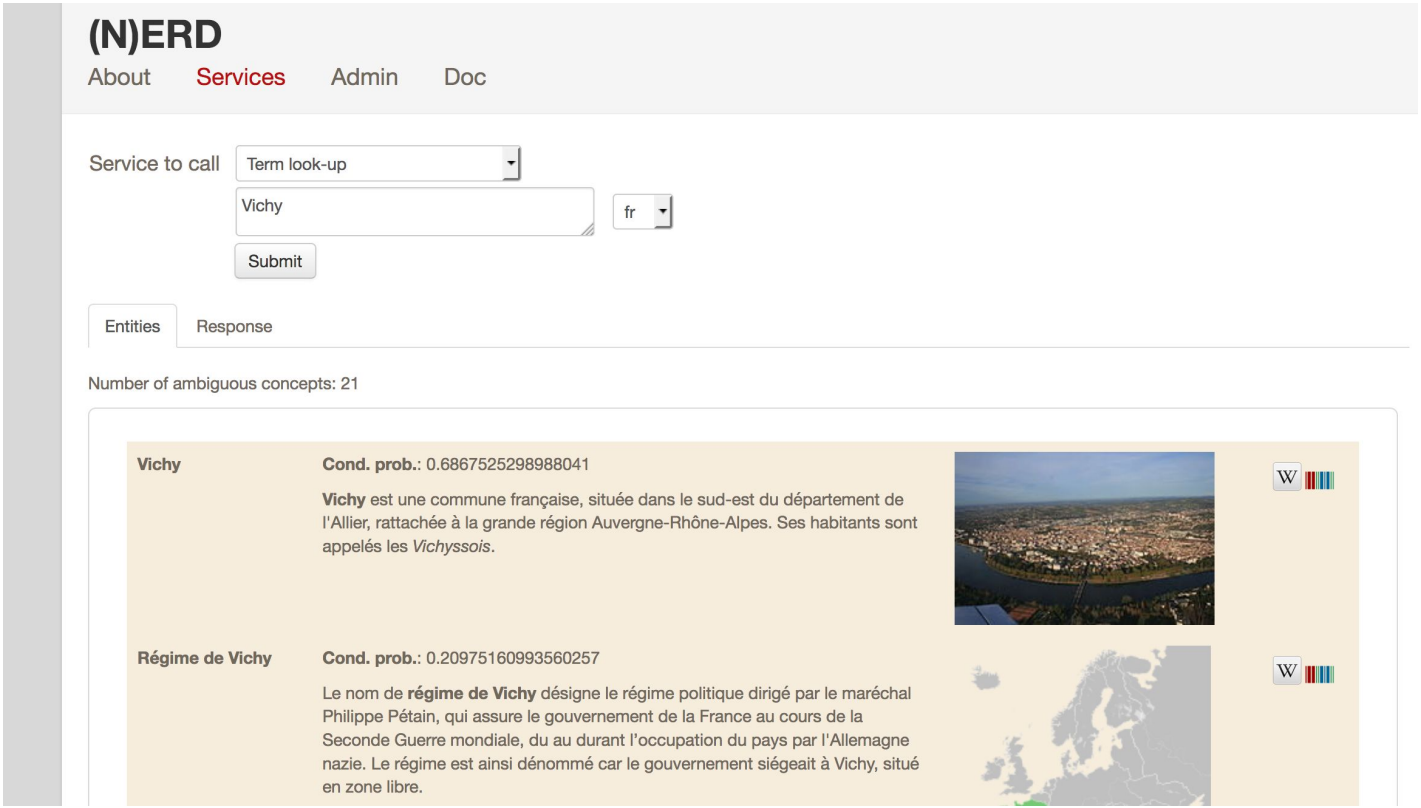
Comité français de Libération nationale

Je vois Yves qui m'avait cité l'impatience d'Alger devant le cas Pucheu comme une illustration de la crise du Gaulisme.

Movement inspired by De Gaulle

Journal de Léo Hamon, March 14th 1944

Resolving ambiguities - Vichy



Resolving ambiguities

GROBID (N)ERD supports context customisation

The customisation is a way to specialize the entity resolution for a particular domain/profile, for example selecting the correct Vichy wikipedia entry related to the ww2 context.

Specific dictionaries can then be integrated seamlessly in the tool.

NOTE: When wikipedia doesn’t provide a page for the *alternative* meaning, this approach cannot be applied.

Workflow (1) - Extract Named Entities

LOCATION LOCATION

Place de la République, Hotel Moderne, vaste bâtisse où étaient logées

ANIMAL ANIMAL NATIONAL

les petites souris grises, d'autres disent « les Salamandres », jeunes allemandes en uniforme. Elles partent et elles ne pouvaient emporter qu'un léger bagage à la main. (...)

Journal d'Henri Chabasse, August 13th 1944

Adjusting entity resolution

No wikipedia page, no party.

The dictionary is then used to override the resolution phase to force the correct definition.

This approach is very specific, less frequent and should be only a fallback solution

E.g. presence of nicknames in the text

Workflow (2) - Combining dictionaries

Place de la République, Hotel Moderne, vaste bâtisse où étaient logées les petites souris grises, d'autres disent « les Salamandres », jeunes allemandes en uniforme. Elles partent et elles ne pouvaient emporter qu'un léger bagage à la main. (...)

German army female auxiliaries

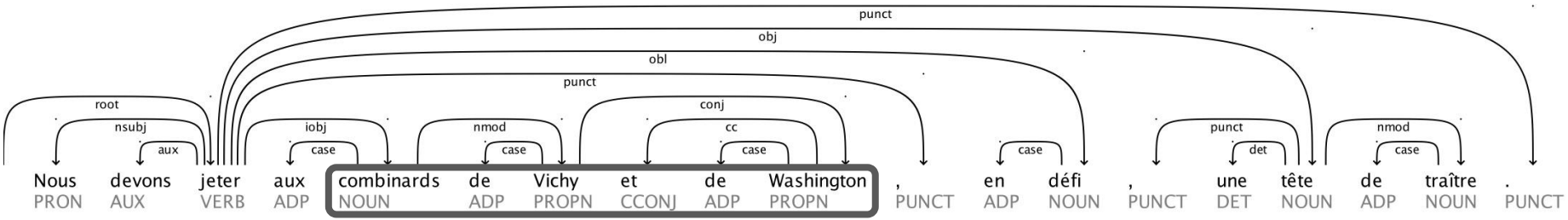
Journal d'Henri Chabasse, August 13th 1944

Workflow (3) - POS tagging and parsing

“Nous devons jeter aux combinards de Vichy et de Washington, en défi, une tête de traître.”

The dependency tree can be used to find expressions and modifiers relative to the entities of interest.

E.g. named entity "Vichy" (*i.e.* Régime de Vichy) and Washington (*i.e.* the US government) ⇔ nominal modifier "combinards"

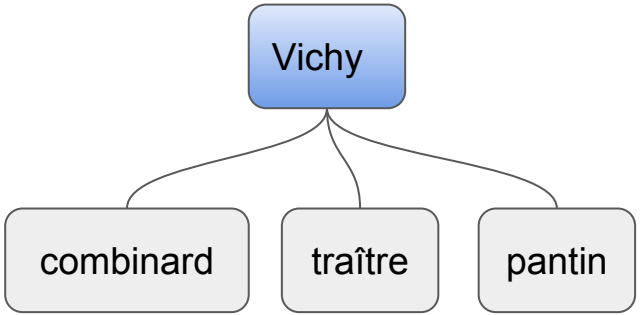


Workflow (4) - mixing everything

Nous parlons du procès Pucheu. La question est plus actuelle. (...)
J'indique qu'à mon avis tout ce procès a été mal conduit - il fallait (...) proclamer que devant des crimes inouïs, (...) la nation prenait une décision politique, immoler les hommes de la haute trahison : "Jetons à l'Europe, en défi, une tête de roi", jetons aux combinards de Vichy et de Washington, en défi, une tête de traître. (...)
Je vois Yves qui m'avait cité l'impatience d'Alger devant le cas Pucheu comme une illustration de la crise du Gaulisme.

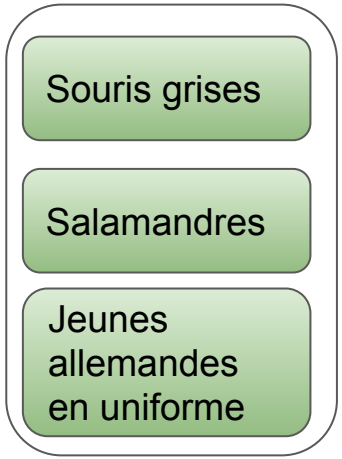
Journal de Léo Hamon, March 14th 1944 entry

Workflow (5) - expected results



List of Named entity modifiers
and expressions

German Army
female auxiliaries



Onomasiological terminology

Next steps

- Apply this workflow to a very large and diverse corpus
  → Multilingual terminology with the specific terms for each war

- Add sentiment polarity tagging layer

- Nice results visualisation

Thank you