History Fishing When engineering meets History

Charles Riondet, Luca Foppiano

History Fishing

When engineering meets History

Charles Riondet, Luca Foppiano
ALMAnaCH, Inria Paris

Abstract

While doing research in digital based context, historians often face the same kind of problem: how to translate their research questions in machine readable format?
Find out how to annotate efficiently their material and which techniques/methods to apply are technical steps that tend to become everyday tasks for many Humanities researchers. The main issue is that their questions are too specific to simply apply generic tools and, on the other hand, adapting such tools may require technical abilities that humanists hardly ever have. In such cases, interacting with experts of different backgrounds (e.g. information technologies, developers) could be difficult due to different views, approach to the problem, way of reasoning, etc.
In this paper is proposed a methodology, and the research use case on which it is tested, where the close collaboration between historians and engineers allows for a better understanding of the needs of each party, and helps the creation of customizable tools capable of being used in many different contexts and domains.
The use-case is to study and compare entities corresponding to the actors of conflicts in a corpus of personal diaries written during World War II. The discourses analyzed are characterized by multiple and very peculiar terminologies, often very ambiguous (like pejorative nicknames). The challenge is to apply generic tools (like named-entity recognition tool – NERD, a POS tagger) and a domain specific dictionary to this corpus, trying not to cross the thin line between generic customization and ad hoc development.

Acknowledgement

We would like to thank Professor Anne Baillot (Université du Maine, Le Mans) and Hector Martinez Alonso (ALMAnaCH, Inria Paris) for their help.

Full text

Introduction

While doing research in digital based context, historians often face the same kind of problem: how to translate their research questions in machine readable format?
Find out how to annotate efficiently their material and which techniques/methods to apply are technical steps that tend to become everyday tasks for many Humanities researchers. The main issue is that their questions are too specific to simply apply generic tools and, on the other hand, adapting such tools may require technical abilities that humanists hardly ever have. In such cases, interacting with experts of different backgrounds (e.g. information technologies, developers) could be difficult due to different views, approach to the problem, way of reasoning, etc.
In this paper is proposed a methodology, and the research use case on which it is tested, where the close collaboration between historians and engineers allows for a better understanding of the needs of each party, and helps the creation of customizable tools capable of being used in many different contexts and domains.
The use-case is to study and compare entities corresponding to the actors of conflicts in a corpus of personal diaries written during World War II. The discourses analyzed are characterized by multiple and very peculiar terminologies, often very ambiguous (like pejorative nicknames). The challenge is to apply generic tools (like named-entity recognition tool – NERD, a POS tagger) and a domain specific dictionary to this corpus, trying not to cross the thin line between generic customization and ad hoc development.

Context

The work presented in this paper is a case study made possible by the very transdisciplinary nature of the Inria ALMAnaCH team. ALMAnaCH means Automatic Language Modelling and Analysis & Computational Humanities and is a joint team between Inria and EPHE (École pratique des hautes études). As a matter of fact, it is a strange mixture of people with different skills working side by side: computer engineering people specialized in data mining and information extraction, NLP experts in semantic and syntactic analysis, parsing, etc., Digital Humanities people (with History and Literature background) focusing more on data modelling, textual analysis, digital philology.
Why not combining all the skills available around? The original situation of the ALMAnaCH team was seen as an opportunity to learn new things and be creative, even if the different partners involved in this project were not at first meant to work together, and were involved in different projects.
The first tries were very empirical and rather unfruitful. Usually, the person with the data is the trigger, and the risk is that the computer science expert might be considered as a simple service provider. In the present case, working alongside on a daily basis allowed the team to reshape this collaboration as a reciprocal enrichment. For historians, it is the opportunity to improve the data and consider new questions. For data mining specialists, it is a way to improve tools, workflows and find new use cases. As a final result, the ambition is to provide a sustainable data mining toolbox generic enough to answer a great variety of Humanities research questions.

Research question

The context of war surely has a strong influence on the language (Antoine, Martin, 1999, p. 3) and sometimes the occasion for lexical creations (Gérard, Lacoste, 2014). Personal writings testify this influence maybe better than any other type of text, because they offer less filtered corpora and cover, at least in the case of WWI and WWII, a diverse range of socio-economic situations, locations, age, gender, etc. The diaristic form also allows for very precise diachronic analysis. In such texts, determining how the Self and the Other are being represented, and how this representation evolve, can offer interesting lines for thoughts. These mentions reveal often structural elements of an individual (or a group) in a given time, how they see themselves, as individuals, as a group, as opponents, winners, losers, etc. When compared to the course of the war, it also allows for the identification of external elements that influence the language.
In the present case, the analysis is limited to the mentions of actors of WWII in French personal writings, more precisely, diaries. What is the Other in this context? First, it is the enemy, the other nation/nations the writers are fighting against (themselves of their group, their nation). The most obvious case is to find how French see the German. But the idea is to broaden the study to all the mentions of nations, organisations, persons, person types, trying to determine the complete panorama of these witnesses and/or actors.

An important parallel goal is to bring together different tools, set up a solid editorial and hermeneutical workflow and make it available for further research. Ideally, this workflow would allow researchers to process historical texts from digitization to publication, via OCR, annotation, edition, etc.
The digitization and OCR parts are left aside, to focus on the annotation task. All instances appearing in the discourse are modelized and brought in relation to one another. This corpus of annotations is the foundation for both synchronic and diachronic analyses. In other words, the future development of this framework will make possible to analyse the variations of the mentions over the corpus and, because the material - diaries - are chronological in essence, over time. Variations can be semantic evolutions, variations in spelling, or appearance and disappearance of a mention.

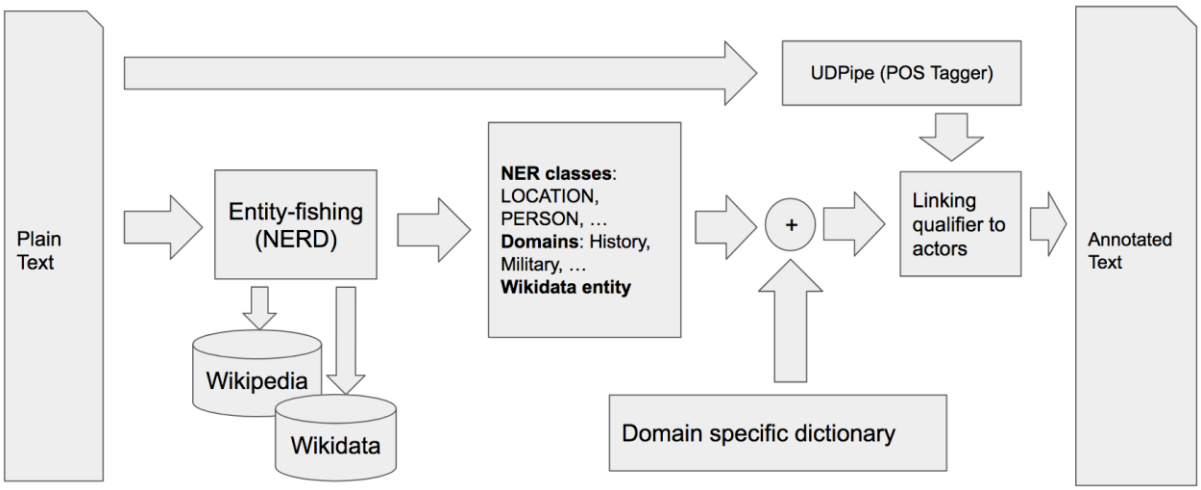
Methodology

In order to tackle the research question in a structured way, two main tasks were defined: the recognition of the actors and the extraction of the related structural discourse (actions or qualifier).
Actors, either “the others”, or “ourselves” are usually people, locations or institution. As these tokens appear in the text as Named Entities the more reasonable approach is the use of a Named Entity Recognition (NER) tool. The NER task “*is a subtask of information extraction that seeks to locate and classify named entities (a real-world object, such as persons, locations, organizations, products, etc, that can be denoted with a proper name) in text (..)"* (Jurafsky *et al.*) These Named entities are classified in predefined classes, such as the names of persons, organizations, locations, etc.

For each actor mention identified in the text, the secondary task is to collect all meaningful tokens, usually verbs, adjectives and nominal modifiers, linked to it.. These are the tokens carrying information (sentiment, judgement, point of view of the writer, etc.) about the actors

and they often spanning towards extremes depending period, location and social origin of the author. The approach for the recognition of structural discourse is based on the part of speech tagging (POST), which is "*is the process of assigning a part-of-speech or other lexical class marker to each word in a corpus*". Traditional grammar classifies words based on eight **parts of speech**: verb, noun, adjective, pronoun, preposition, adverb, conjunction and the interjection. According to [Jurafsky *et al.*] the part of speech for a word gives a significant amount of information about the word and its neighbors.

These two tasks are combined in the following workflow:



The main components of this schema are:

- entity-fishing[1] (NERD): tool performing the entity recognition (NER) and disambiguation against wikipedia and wikidata
- UDPipe[2] (POST): tool trained to perform Universal Dependency parsing and part of speech tagging.

Assuming to have input as text (but it could be a PDF document as well) it is crunched in parallel by POST and NERD, the result is further post-processed and combined to obtain the annotated text. The annotations obtained in output are composed by the actor and the qualifiers.

Entity fishing

Entity fishing is a tool written by Patrice Lopez[3] and released open source under the licence Apache 2.0.
It is also more generically named (N)ERD that stands for (Named) Entities Recognition and Disambiguation, *Named* is between parentheses because the obtained entities might not necessarily *Named Entities* but just mentions. In addition the found mentions are matched against Wikipedia and Wikidata to find the article related to the real entity. Wikipedia/Wikidata are used for disambiguation because they are the largest and most complete knowledge bases available nowadays.

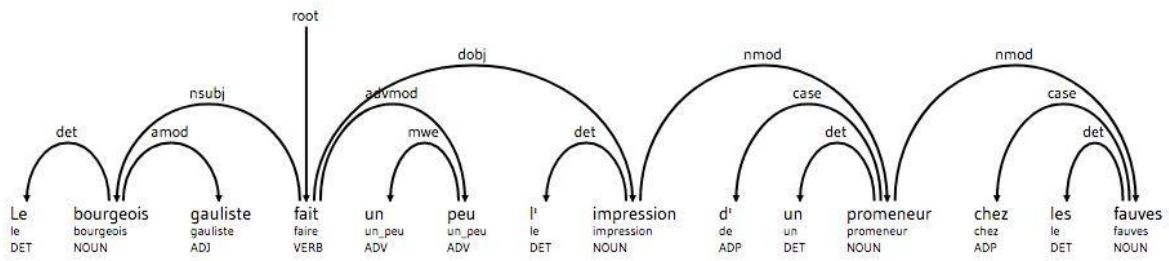
[1] http://www.github.com/kermitt2/nerd
[2] http://ufal.mff.cuni.cz/udpipe
[3] http://github.com/kermitt2

This task is called disambiguation or entity-linking and for certain cases can be quite subtle, as illustrated in the figure below where the ambiguities can be only resolved by the context itself:

In this case, it is possible to choose the right article for **German Army** only by reading the whole text and understand whether it's talking about the first or second World War.

UDPipe

UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing. This tool is used to obtain the dependency graph of the sentence combined with the Part-of-Speech tagging of each token. The dependency graph represent each sentence as hierarchical tree having the verb as root. Each token is connected to one head (a token closer to the root) and one or more more descendents tokens.



Thanks to this approach, it is possible to discover relations between various components of the sentence. In the example in figure, the entity *bourgeois* is modified by the descendant adjective *gauliste*.

The corpus

This workflow is tested with a rather small corpus: two diaries written in French by persons living in Paris during WW2. They report two different personal experiences.

The main text of the corpus is the *Journal de Léo Hamon* (Archives nationales, 72AJ42). Léo Hamon (1908-1993) was a French lawyer of Russian origin, and one of the leaders of the Parisian Resistance. His diary relates his underground daily life, reports on his comments on the course of the war, meetings he attended, the organization of the Resistance and on the preparation of the seizure of power in Paris (Riondet, 2016).
Another diary used is the *Journal d'Henri Chabasse* (Musée de la Résistance nationale, 13/3907b). Chabasse is a nationalist middle-class Parisian, not involved in the Resistance nor in the Collaboration. His diary reports daily life but contain mostly comments on the course of the war and the French political situation, from the D-Day to autumn 1944 and an additional entry related to the Hiroshima bombing.

Specific dictionary

These two diaries contains a peculiar vocabulary, that was known at the time it was written. Therefore, numerous entities are difficult to recognize for a generic Named-entity recognition system. Many context based expressions are present in both texts. For example, "souris grises" means "grey mice", but in the context of WWII is a nickname to designate the German army female auxiliaries. Such expression was commonly used by parisians talking about the situation during WWII and might appear in other sources.
Another problem is the use of a different internal terminology for each diary. For instance, Léo Hamon uses a very specific term when mentioning the members of the French communist party: "les cocs", from the pejorative diminutive "coco", itself shorten.
There are other cases that may seem less important but required to be taken into account to avoid losing important information. For example unnormalized spelling (e.g.: *Gaulisme* for *Gaullisme*) or metonymies (e.g.: *Vichy* for *Régime de Vichy* and not Vichy town).

Due to these particularities, even a complete source like Wikipedia doesn't contains everything. Therefore it is necessary to create a specific domain dictionary to record all the possible terminology variations and link them the concept they represent. This dictionary is based on other domain dictionaries like the Dictionnaire historique de la Résistance (Marcot et al., 2006), other scholarly work (Vast, 2008) and on manual curation.
The dictionary is modeled with the XML format TEI-TBX, a combination of two standards, Text Encoding Initiative (TEI)[4] and the TermBase eXchange (TBX)[5]. This format offers the possibility to cluster all the mentions related to a concept and link concepts between one another (Pernes, Romary, Warburton, 2017). As it is a standard, the resource created is sustainable.

```
<conceptEntry xml:id="c_1067">
	<langSec xml:lang="fr">
		<form type="lemma">
			<orth>Gaullisme</orth>
		</form>
		<form type="variant">
			<orth>Gaulisme</orth>
		</form>
	</langSec>
```

[4] http://www.tei-c.org/About/mission.xml
[5] https://www.iso.org/standard/45797.html

```
</conceptEntry>
```

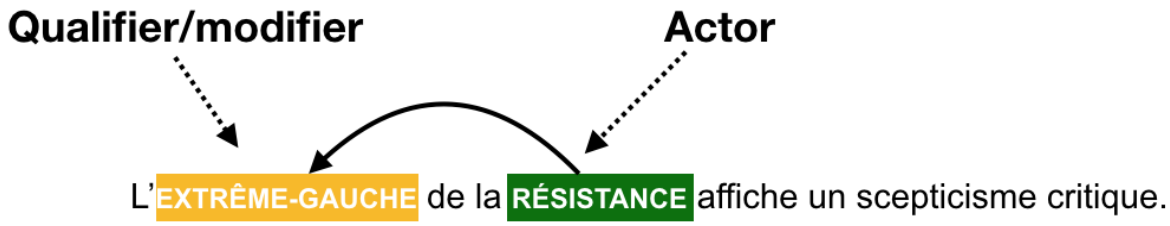
The concept "Gaullisme" and its variant "Gaulisme" expressed in TEI-TBX.

Workflow

In this section is described the implemented workflow that combines all the tasks described above:

- mentions extraction and entity disambiguation against Wikipedia/Wikidata
- specific domain entities extraction using the dictionary
- dependency parsing and POS tagging

The result is composed of the entities representing the actors of the text linked to their respective qualifiers. For example:



The first step is the analysis of the text and the extraction of named entities and mentions. The extracted entities are then filtered by Named Entity classes (*Entity Fishing* cover up to 27 classes[6]) as mentions of actors of the conflict are likely to bo entities of the following classes:

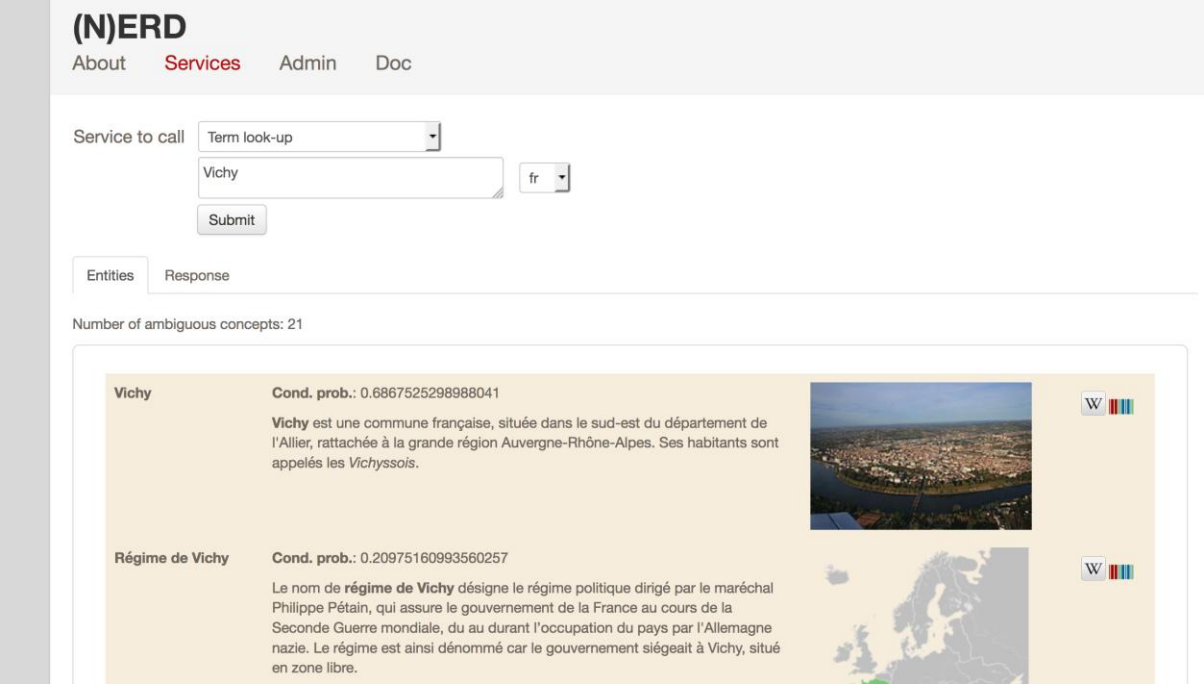
- LOCATION
- PERSON
- TITLE
- ACRONYM
- PERSON TYPE
- ORGANISATION
- INSTITUTION
- ANIMAL

In the following example, taken from the diary of Léo Hamon, can be found entities particularly relevant within the classes *PERSON*, *ORGANISATION* and *LOCATION*:

"Nous parlons du procès Pucheu. La question est plus actuelle. (...) J'indique qu'á mon avis tout ce proces a ete 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 traitre. (...) Je vois Yves qui m'avait cité l'impatience d'Alger devant le cas Pucheu comme une illustration de la crise de Gaulisme."

[6] http://grobid-ner.readthedocs.io/en/latest/class-and-senses/

In addition to the NE classes, Entity Fishing provides a link to the corresponding article in Wikipedia/Wikidata by looking into all possible mentions, that means that if a token (word) is not recognised as a Named Entity but has a corresponding article in Wikipedia/Wikidata, it will be found, increasing the overall recall.
The entity *Vichy*, present in the above example, have 21 corresponding articles in Wikipedia. The first two meanings are *Vichy*, the city in France, and *Régime de Vichy*, the government allied with the Germans, led by Philippe Pétain during WW2.



Whether the mention *Vichy* refers to one or another of the 21 articles is inferred by the context of the whole text. For instance, processing a text describing the Ironman race of Vichy would more likely select the city of Vichy as the real entity corresponding to the mention. This example is not particularly fortunate as Vichy can well appear in WW2 texts as the city, hence the results have always a margin of error.

This disambiguation process return the following results:

Vichy	Régime de Vichy, État Français
Washington	United States of America
Alger	Comité Français de Libération nationale
Gaulisme	Movement inspired by general de Gaulle

The disambiguation process add precious information in the treatment of this text, once the correct article is found, the access to the corresponding wikidata information rich set of properties, statements and translations that can be further exploited (outside the scope of this article).

Adjusting entity resolution

During the tests with this specific corpus were found many expressions either widely used at the time of the war or specific of the author. Léo Hamon uses a lot of nicknames to identify other members of the Resistance (many of them can be mapped to the real person), he uses also expressions which are not found in Wikipedia nor in any other knowledge base. As discussed before, since these expressions are countable and very specific, it is inevitable to use a dictionary in order to identify them and extend the pool of generic entities found with the NERD. The dictionary is also used to override the resolution phase in order to force the correct definition:

Place de la République, Hotel Moderne, vaste bâtisse où étaient logées les petites souris grises, d'autres disent « les Salamandres », jeunes allemandes en uniforme.

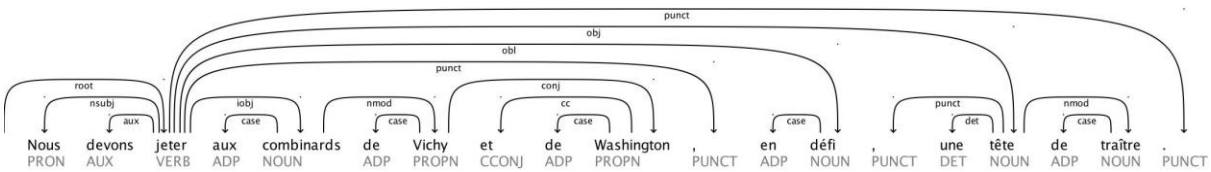
In this case, the tokens *petites souris grises*, *Salamandres* and *jeunes allemandes en uniforme* all refer to the same entity: German Army female auxiliaries. Two of these mentions are nicknames based on animals. This can only be infer by the use of the specific dictionary. There is, however, a limitation in its use: if the author uses the real expression with the nickname, the disambiguation is impossible. The first approximation is to assume that the author doesn't talk about grey mice (*souris grises*) or salamanders (*Salamandres*) in his World War 2 diary. A second approximation (suitable for a future paper) would be to use the probability returned by Entity Fishing to make a decision whether to use the dictionary to override the real entity found from Wikipedia.

POS tagging and parsing

The last building block of this process is to link the found entities to their modifiers, using the dependency graph provided by UDPipe.
For each entity, the goal is to find the node marked as HEAD and the nodes marked as descendants. The HEAD is the node upstream in the graph, the one whose the entity directly depends on. On the other hand the DESCENDANTS are nodes depending on the entity itself.

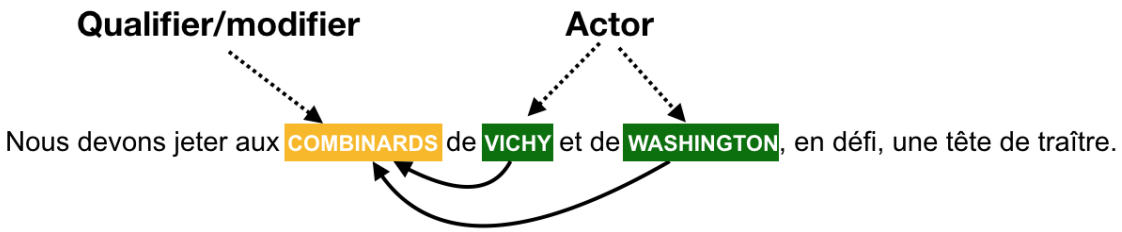
For example the sentence:
"*Nous devons jeter aux combinards de Vichy et de Washington, en défi, une tête de traître.*"

From the dependency graph showing the whole syntactic structure (see below), it is necessary to select the relevant elements.



From a theoretical perspective the named entities "Vichy" (i.e. Régime de Vichy) and Washington (i.e. the US government) share a nominal modifier *combinards* (schemer)

In the parsing results, the modifier *combinards* is identified as the HEAD of *Vichy*, it is also linked to *Washington* via the conjunction *et* (and) and the fact the *Vichy* is the HEAD of *Washington.*



Expected results

This paper represents a first step for this broad research subject. The main expected results are twofold: clustering entities and their modifiers and the generation of multilingual terminology focused on our topic (French-German modern wars).
These results can be seen as a graph where relevant and disambiguated entities are connected to all their modifiers, their adjectives, etc. This graph would show clusters of entities, sorted according to certain criterias: their polarity, by the persons who uses them, by the chronological extent of their use, etc. The same approach should also be applied to the extracted qualifiers.
The onomasiological terminology, that records all the terms used to designate the concept can also be considered as a layer of the entities graph.

Bibliography

- Gérard Antoine, Robert Martin (ed.), Histoire de la langue française des origines à nos jours (1880 - 1914), Paris, CNRS, [1975] 1999
- Christophe Gérard, Charlotte Lacoste. La création lexicale dans les écrits de combattants de la Première Guerre mondiale. *La Première Guerre mondiale et la langue*, Jun 2014, Paris, France. 〈halshs-01093817〉
- Charles Riondet. Journaux Intimes de Clandestinité: Le cas de Léo Hamon (19040-1944). *Vingtième siècle*, Fondation Nationale des Sciences Politiques, 2016. 〈hal-01416988〉
- Jurafsky, Dan. *Speech & language processing*. Pearson Education India, 2000.
- Cécile Vast, *Une histoire des Mouvements Unis de Résistance (de 1941 à l'après-guerre) : Essai sur l'expérience de la Résistance et l'identité résistante*, Besançon, 2008. <tel-00596588>
- François Marcot, Bruno Leroux et Christine Levisse-Touzé, *Dictionnaire historique de la Résistance: résistance intérieure et France libre*, Paris, R. Laffont, 2006.
- Stefan Pernes, Laurent Romary, Kara Warburton. TBX in ODD: Schema-agnostic specification and documentation for TermBase eXchange. *LOTKS 2017- Workshop on Language, Ontology, Terminology and Knowledge Structures*, Sep 2017, Montpellier, France. 〈https://langandonto.github.io/LangOnto-TermiKS-2017/〉. 〈hal-01581440v2〉
- Straka Milan, Hajič Jan, Straková Jana. *UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing*. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, May 2016.
- Milan Straka and Jana Straková. *Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe*. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017.