A Data-driven Approach to Named Entity Recognition for Early Modern French

Pedro Ortiz Suarez
Data and Web Science Group,
University of Mannheim
B 6, 26
68159 Mannheim, Germany
pedro.ortiz@uni-mannheim.de

Simon Gabay
Université de Genève
Rue du Général-Dufour 24,
1211 Genève, Switzerland
simon.gabay@unige.ch

Abstract

Named entity recognition has become an increasingly useful tool for digital humanities research, specially when it comes to historical texts. However, historical texts pose a wide range of challenges to both named entity recognition and natural language processing in general that are still difficult to address even with modern neural methods. In this article we focus in named entity recognition for historical French, and in particular for Early Modern French (16th-18th c.), i.e. *Ancien Régime* French. However, instead of developing a specialised architecture to tackle the particularities of this state of language, we opt for a data-driven approach by developing a new corpus with fine-grained entity annotation, covering three centuries of literature corresponding to the early modern period; we try to annotate as much data as possible producing a corpus that is many times bigger than the most popular NER evaluation corpora for both Contemporary English and French. We then fine-tune existing state-of-the-art architectures for Early Modern and Contemporary French, obtaining results that are on par with those of the current state-of-the-art NER systems for Contemporary English. Both the corpus and the fine-tuned models are released.

1 Introduction

Named entity recognition (NER) is an extensively studied task in natural language processing (NLP) that consists in identifying and classifying *named entities* mentions in unstructured text. These *named entities* often are real-world objects such as a *person*, a *location*, an *organisation* name or even a *product*. NER has been an important task in natural language processing for some time now. It was the focus of the MUC conferences and associated shared tasks (Marsh and Perzanowski, 1998), and later that of the CoNLL 2003 and the ACE shared tasks (Tjong Kim Sang and De Meulder, 2003; Doddington et al., 2004).

NER has quickly established itself as a pillar of the new methods of reading texts promoted by the digital humanities (DH), based on the analysis of large sets of literary or historical data via computational methods (Moretti, 2005). These sources being not only contemporary, the need for tools dealing with medieval or early modern states of language is now increasing. NER interests researchers in DH for numerous reasons since the application can be quite broad, from genealogy or history studies for which finding mentions of persons and places in texts is very useful; to applications in digital literature where researchers can use NER to highlight the path of different characters in a book or in a series of publications. Both the research in NER and DH can benefit from one another as it has already been suggested particular properties of literature can help to build better NER systems (Brooke et al., 2016) and even study how much diachronic variation influences NER systems (Ehrmann et al., 2016).

For the present study, we will focus on developing both an annotated corpus as well as a NER system for *Early Modern French*. We loosely define *Early Modern French* as a state of language following Middle French in 1500—adopting here the *terminus ad quem* used by the *Dictionnaire de Moyen Français* (Martin, 2020)—and ending with the French Revolution in 1789. In consequence, it encompasses three centuries (16th, 17th and 18th c.), or two linguistic periods: the *français préclassique* or "preclassical French" (1500–1630) and the *français classique* or "classical French" (1630–1689); both periodisations which are currently used in French linguistics (e.g. by Vachon 2010 and Amatuzzi et al. 2019). Early Modern French poses some particular challenges for NER systems, and mainly two. First, the spelling was not fixed and place names could be written differently from one text to another, but also in the same text. In Early Modern French, the name of the city of *Lyon* could

Proceedings of the 29th International Conference on Computational Linguistics, pages 3722–3730
October 12–17, 2022.

be written *Lyon*, but also *Lion*, creating in this case a homograph that has today disappeared (the *lion* being, like in English, an animal). Second, cities have changed their names, states have appeared, empires have disappeared, etc. and it is therefore impossible to use tools available for Contemporary French.

In this paper we develop a system that tries to tackle these specific challenges posed by Early Modern French, however, instead of developing a specialised architecture for this, we opt for a data-driven approach in which we try to annotate as much text as possible of an heterogeneous corpus covering several centuries and a vast range of genres and styles. We produce a fine-grained NER annotated corpus for Early Modern French that is many times bigger than some of the most popular NER annotated corpora for Contemporary English and French (Tjong Kim Sang and De Meulder, 2003; Sagot et al., 2012). We then fine-tune existing state-of-the-art architectures D'AlemBERT (Gabay et al., 2022) and CamemBERT (Martin et al., 2020) for Early Modern and Contemporary French respectively obtaining results that surpass the current state of the art NER systems for Contemporary French (Ortiz Suárez et al., 2020a), and that are on par with NER systems for Contemporary English (Straková et al., 2019; Yamada et al., 2020; Wang et al., 2021). We release both the corpus and the fine-tuned model in order to insure reproducibility of our experiments.[1]

2 Related work

If many evaluation campaigns for the recognition of named entities have been carried out since the end of the nineties [2], most of the corpora produced have until recently dealt with contemporary documents, particularly taken in the press (articles, dispatches...). In recent years, however, research has begun to focus on "historical" documents, but the diachronic depth of the language remains imperfectly treated, with a very clear concentration on the most recent textual sources: the 19th c. and 20th c. are by far over-represented (Ehrmann et al., 2021).

If the older states of language, linguistically more complex because of the instability of their spelling, remain left aside, we do note some attempts to extract entities from texts written before the 19th c. Previous research concerns 17th c. English (OCRised versions of the *Journals of the House of Lords*, cf. Grover et al. 2008), medieval latin (charters, cf. Torres Aguilar et al. 2016), German and French (legal documents written between the 14th and the 18th c., cf. Gwerder 2017). With the emergence of data-driven approaches, new corpora keep emerging for niche languages such as Middle High German and Old Norse (Besnier and Mattingly, 2021).

French is a typical case regarding NER, with resources and solutions focusing on documents written after the French Revolution. One of the oldest dataset is the one produced during the ESTER-2 evaluation campaign (Galliano et al., 2009), dealing with of radio broadcast transcripts. For the older documents, we have the *Quaero* (Rosset et al., 2012), *Europeana* (Neudecker, 2016) and *Impresso* (Ehrmann et al., 2020) corpora, going back the 19th c., but again with an almost unique focus on the press. Non-journalistic and/or non-recent French, however, seem to have attracted researchers in recent years. We have already mentioned the study of Gwerder (2017), whose data has unfortunately not been manually annotated and is therefore far from being optimal, and is limited to place and person names. If older rule-based approach keep being used (for place names, cf. Kogkitsidou and Gambette 2020), only one project has produced a manually annotated corpus, but limited to toponyms and using normalised versions (i.e. aligned with Contemporary French) of 17th c. plays (Gabay and Vitali, 2019).

An ambitious manually annotated corpus for pre-Revolutionary non-normalised French is still needed to give the means to researchers in history, literature or linguistics to offer new interpretation, relying on quantitative approaches such as "distant reading" (Moretti, 2013). If possible, this would corpus would need cover several centuries, and to offer more entities than just place and person names, such as quantities or events.

3 Corpus

Rather than designing a new corpus, we have decided to use a subpart of the "core corpus" of the *Presto* project (Blumenthal et al., 2017), namely the text written during the French *Ancien Régime*

[1] URL retained for anonymity.

[2] We spare the reader this story, which has already been perfectly told elsewhere cf. Ehrmann (2008); Nouvel et al. (2015).

Person			Function		
`pers.ind`	`pers.coll`		`func.ind`	`func.coll`	
Location			**Production**		
`loc.adm.town`	`loc.phys.geo`	`loc.fac`	`prod.art`	`prod.rule`	`prod.object`
`loc.adm.reg`	`loc.phys.hydro`	`loc.oro`			
`loc.adm.nat`					
`loc.adm.sup`					
Organization		**Time**		**Event**	**Quantity**
`org.adm`	`org.ent`	`time.date.abs`		`event`	`amount`
		`time.date.rel`			

Table 1: Types (in gray) and subtypes retained from the *Quaero* typology.
f

(c.15^{th}-18^{th} c., i.e. 34 texts)[3]. This choice is driven by our will to limit the number of annotated corpora for historical French, the same set of documents having already been abundantly corrected to train a lemmatizer (Gabay et al., 2020), but also to avoid a complex selection of works supposed to ensure a relative representativeness of literary documents from the *Ancien Régime*, already perfectly done by our colleagues.

The number of genres covered is extremely large: poetry, drama, novel, correspondence, grammar, philosophy, short stories, encyclopedic literature, etc. and guarantees, here again, a reasonable representativeness of the range of possibilities of *Belles-Lettres*[4]. The corpus is balanced regarding the distribution per century (c. 10/century) but not regarding the length of the texts, which increases over time (cf. fig. 1), following a possible trend in literature.

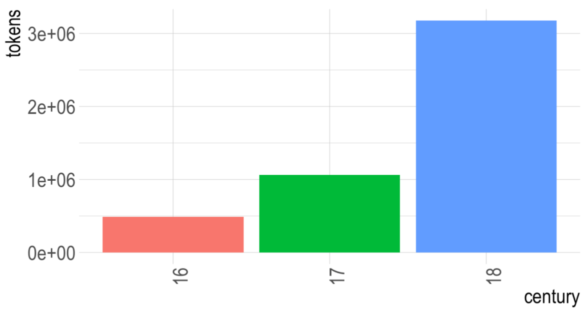

Figure 1: Number of tokens per century.

3.1 Annotation

It seemed logical to follow the *Quaero* annotation guide (Rosset et al., 2011), that is used by two important historical corpora presented *supra* (*Quaero* and *Impresso*). Because our texts and interests diverge from those of the aforementioned corpora, only some types and subtypes have been kept (cf. tab. 1) from the *Quaero* typology. The details of our choices can be found in a dedicated annotation manual (Gabay et al., 2022).

The annotated texts are available in multi-columns `tsv` files (cf. tab. 1). Each token has a lemma (manually corrected) and a POS (produced by the *Presto* project, non-systematically corrected but fairly reliable) using the MULTEXT tag set. We propose a coarse-grained annotation for high-level entity types and fine-grained annotation using subtypes using the following syntax:

`BIO-TYPE.SUBTYPE`
For instance: `B-loc.adm.town`

Subtypes are sometimes simple (`B-org.town`) sometimes double (`B-loc.phys.geo`), depending of the complexity of the entity to annotate. Nested entities (i.e. an entity in an entity, such as a place name in a person name in *Henri d'**Angleterre***, "Henry of England") follow exactly the same syntax, and components a similar one, using six transverse elements:

- `name` to annotate tokens that are names (*Louis*, *Philippe*...)
- `title` to annotate tokens that are titles (*sieur*, *duc*, *abbé*...)
- `qualifier` to annotate tokens that are adjectives (*l'Inde **orientale***, *l'Arabie **heureuse***, *la mer **athlantique***, *l'**ancienne** Colchide*)... but also the generation (*Henri **IV***) or a cardinal position
- `kind` to annotate tokens that are hyperonyms (*l'**Empire** de Constantinople*, *la **mer** du Japon*

[3] A text has been withdrawn: the *Histoire d'un voyage faict en la terre du Brésil* by Jean de Léry, the transcription being too faulty to be able to correctly annotate the document.

[4] We do not offer a detailed description of the genres covered, these overlapping easily: poetry can be theological, political correspondence...

Token	Lemma	POS	COARSE	FINE	FINE-COMP	NESTED	Wikidata ID
Les	le	Da	O	O	O	O	_
allemands	allemand	Nc	O	O	O	O	_
élurent	élire	Vvc	O	O	O	O	_
pour	pour	S	O	O	O	O	_
empereur	empereur	Nc	B-pers	B-pers.ind	B-comp.title	O	Q438435
Rodolphe	Rodolphe	Np	I-pers	I-pers.ind	B-comp.name	O	Q438435
duc	duc	Nc	I-pers	I-pers.ind	B-comp.title	O	Q438435
de	de	S	I-pers	I-pers.ind	I-comp.title	O	Q438435
Suabe	Souabe	Np	I-pers	I-pers.ind	I-comp.title	B-loc.adm.reg	Q438435

Table 2: NERC Fine-Grained annotation with EL

- `unit` to annotate tokens that are units (meters, league, inches, pounds. . .)
- `val` to annotate tokens that are values (a number) that is linked to a unit to annotate an `amount`.

We have decided not to annotate metaphorical uses differently or in a separate column: everything is annotated in a literal sense. Thus, in ***France** goes to war*, *France* is labelled `loc.adm.nat` (i.e. the country) and not `org.adm` (i.e. the French government).

We have also started a first phase of semantic annotation, using Wikidata (Vrandečić and Krötzsch, 2014) identifiers, which remains imperfect. Due to the complexity of analysing certain entities, in particular personal names (e.g. *Pope John*), it was decided to annotate them only very marginally, only in the event of the absence of ambiguity (e.g. *Pope John V*). The annotation of place names, on the other hand, is more advanced and almost functional.

A first layer of annotation was made using regular expressions, before moving on to a manual correction phase. Given the size of the corpus, it is obvious that each token has not been checked, and that the final result does not claim to be perfect. Regular and thorough checks, however, concluded that the annotation was of the best possible quality and allow to move on to the training phase. All of the annotation work was carried out by a single person, in order to ensure the consistency of the data. The structure of the file and the form of the tags was controlled by a specific parser, designed specifically for this corpus.

3.2 A Note on Size

Our final annotated corpus has around 5 million annotated tokens, this makes it around 18 times bigger than the French treebank (Abeillé et al., 2003; Sagot et al., 2012; Ortiz Suárez et al., 2020a) and almost 23 times as big as the CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) corpus. Figures 1 and 2 show both the distribution of tokens by century and by *coarse* entity type. We can see that even though our corpus is far from balanced, even the 16th century portion of the corpus, which is our smallest, is still slightly larger than both the ConNLL 2003 and the FT corpora. We therefore believe that this annotated corpus gives us a great opportunity to study how state-of-the-art NER architectures behave when confronted with large amounts of annotated heterogeneous text.

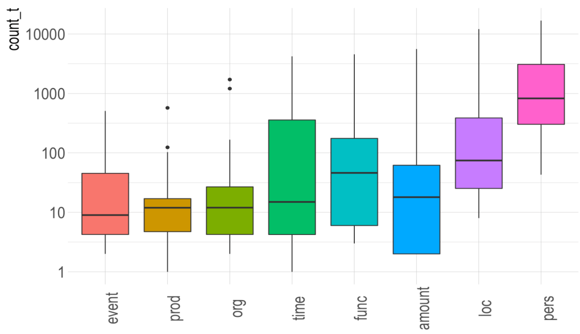

Figure 2: Number of entities ($\log_{10}$ scale) per category by text.

Given the size of our corpus, we opt for a 90-5-5 type split, that is, 90% of the text goes to the training set, 5% to the development set and 5% to the test. Otherwise the test and development sets would have been too big and training would have taken too long. The split is done at a document level and the sentences that go into the development and test sets are chosen at random, ensuring that both sets contain a representative portion of each of the documents in our corpus.

4 NER Evaluation

Having produced this annotated corpus, we now proceed with an evaluation using the *coarse* level of annotation. We only use this level of annotation and not the other columns depicted on on table 2 as the training of some of as architectures turned out to be quite expensive due to the size of the corpus, with a single run of some of our models taking more than 24 hours on a machine equipped with an Nvidia V100 with 32 GB of memory. We also believe that the development of an architecture able to predict all levels of annotations at once merits a study of its own.

4.1 Models

We train three different models, a BiLSTM-CRF (Lample et al., 2016), CamemBERT (Martin et al., 2020) and D'AlemBERT (Gabay et al., 2022). All the training and fine-tuning is conducted using the `flair` framework[5] for sequence tagging (Akbik et al., 2019). To fine-tune D'AlemBERT and CamemBERT POS we follow the same approach as Schweter and Akbik (2020) with some modifications: we append a linear layer of size 256 that takes as input the last hidden representation of the <s> special token and the mean of the last hidden representation of the subword units of each token, that is, we use a *"mean"* subword pooling strategy. For the BiLSTM-CRF we use the implementation provided by the `flair` library, and we couple it with character embeddings as well as the Common Crawl-based FastText embeddings (Grave et al., 2018) originally trained by Facebook. Here is a small description of each of the models:

BiLSTM-CRF A classical neural architecture originally proposed by Lample et al. (2016) that combines a pre-trained fixed word embeddings with character embeddings, that are then feeded into a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) encoder and a CRF (Lafferty et al., 2001) decoder. This model will serve as our baseline.

CamemBERT A Contemporary French language model originally proposed by Martin et al. (2020), is a Bidirectional Transformer-based model (Devlin et al., 2019; Vaswani et al., 2017) more precisely based on the RoBERTa (Liu et al., 2019) architecture, but using SentencePiece (Kudo and Richardson, 2018) instead of the original Byte-Pair Encoding (BPE) (Sennrich et al., 2016). CamemBERT uses a *base*-type architecture, which consists of 12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters. CamemBERT was pre-trained using the French subcorpus of OSCAR 2019 (Ortiz Suárez et al., 2019; Ortiz Suárez et al., 2020b) which is extracted from Common Crawl snapshots, specifically from the plain text *WET* format distributed by Common Crawl which removes all the HTML tags and converts the text formatting to UTF-8. It follows the same approach as Grave et al. (2018) by using a language classification model based on the fastText linear classifier (Joulin et al., 2016; Joulin et al., 2017).

D'AlemBERT An Early Modern French language model originally pre-trained by Gabay et al. (2022) using a 1.2 GB corpus of Early Modern French called FREEM$_{max}$ (Gabay et al., 2022). D'AlemBERT uses the exact same *base*-type architecture as CamemBERT but for the tokenizer it uses the original BPE (Sennrich et al., 2016) of RoBERTa's (Liu et al., 2019) instead of SentencePiece (Kudo and Richardson, 2018). As opposed to CamemBERT or RoBERTa, D'AlemBERT was only trained for 31k steps.

4.2 Results and discussion

Model	Precision	Recall	F1-Score
BiLSTM-CRF	0.8640	0.8533	0.8586
CamemBERT	*0.9303*	*0.9309*	*0.9306*
D'AlemBERT	**0.9329**	**0.9323**	**0.9326**

Table 3: Comparison between D'AlemBERT, CamemBERT and an LSTM-CRF-based model performance on the test set of our corpus, results are averaged over 10 runs with different seeds.

Table 3 shows a brief overview of our results, we can see that our BiLSTM-CRF already produces quite strong results, attaining an f1-score of 0.8586 which is quite remarkable taking into account how heterogeneous our corpus is and how different the data itself is from the pre-training data used in the FastText word embeddings of the Bi-LSTM model.

On the other hand for both CamemBERT and D'AlemBERT we obtain quite high results above the 0.93 in f1-score. These results are quite remarkable because in spite of how heterogeneous our corpus is, and despite of the challenges posed by an historical language previously discussed, we obtain results that are almost on par with the current state of the art architectures for Contemporary English (Straková et al., 2019; Yamada et al., 2020; Wang et al., 2021).

[5] https://github.com/flairNLP/flair

CamemBERT				
Entity Type	Precision	Recall	F1-Score	Support
pers	0.9373	0.9236	0.9304	2734
loc	0.9140	0.9371	0.9254	1384
amount	0.9840	0.9840	0.9840	250
time	0.9447	0.9407	0.9427	236
func	0.9209	0.9143	0.9176	140
org	0.8364	0.9388	0.8846	49
prod	0.7742	0.8889	0.8276	27
event	0.8333	0.8333	0.8333	12
micro avg	0.9303	0.9309	0.9306	4832
macro avg	0.8931	0.9201	0.9057	4832
weighted avg	0.9307	0.9309	0.9307	4832
samples avg	0.8856	0.8856	0.8856	4832

D'AlemBERT				
Entity Type	Precision	Recall	F1-Score	Support
pers	0.9355	0.9279	0.9317	2734
loc	0.9242	0.9335	0.9288	1384
amount	0.9800	0.9800	0.9800	250
time	0.9456	0.9576	0.9516	236
func	0.9333	0.9000	0.9164	140
org	0.8148	0.8980	0.8544	49
prod	0.8621	0.9259	0.8929	27
event	0.8333	0.8333	0.8333	12
micro avg	0.9329	0.9323	0.9326	4832
macro avg	0.9036	0.9195	0.9111	4832
weighted avg	0.9331	0.9323	0.9327	4832
samples avg	0.8893	0.8893	0.8893	4832

Table 4: Results of CamemBERT and D'AlemBERT on the test set of our corpus by entity type. Results are averaged over 10 runs with different seeds.

Strikingly, we do not see the same phenomenon as Gabay et al. (2022) who fine-tuned both CamemBERT and D'AlemBERT in POS tagging for Early Modern French, and that obtained remarkably good results with D'AlemBERT but subpar results with CamemBERT. We believe that this is due to the striking size of our corpus which has more than 5 million annotated tokens, that is, we believe that in this case CamemBERT has enough training data in order to properly fine-tune to this task in Early Modern French and in particular to potentially overcome the poor representations given by the SentencePiece (Kudo and Richardson, 2018) trained on Contemporary French for the out-of-vocabulary words found in the Early Modern French data. [6] We believe that to a certain extent, given the size of our corpus, CamemBERT might be "*forgetting*" its pre-training contemporary data and "*re-learning*" the Early Modern French data in our corpus. In any case, these high score proves the effectiveness of our data-driven approach as we didn't use any dedicated architecture for NER, yet we obtain state-of-the art results for a very challenging state of the French language.

In tables 5 and 4 we see the results of the BiLSTM-CRF, CamemBERT and D'AlemBERT models by entity type. All results are averaged over 10 runs using different seeds. For the BiLSTM-CRF model we see that in general it performs the best for the most common entity types and the worst for the least common types. It has particular trouble with the production category which might be due to the lack of these entities in the web-based pre-training corpus of the FastText fixed word embeddings. Strikingly, we see very good results for the amount entity type with our LSTM-based model, this is actually remarkable as this has historically been a rather difficult entity type to annotate for NER systems.

BiLSTM-CRF				
Entity Type	Precision	Recall	F1-Score	Support
pers	0.8808	0.8435	0.8617	2734
loc	0.8109	0.8707	0.8397	1384
amount	0.9040	0.9040	0.9040	250
time	0.9604	0.9237	0.9417	236
func	0.8872	0.8429	0.8645	140
org	0.8824	0.6122	0.7229	49
prod	0.9231	0.4444	0.6000	27
event	0.7273	0.6667	0.6957	12
micro avg	0.8640	0.8533	0.8586	4832
macro avg	0.8720	0.7635	0.8038	4832
weighted avg	0.8659	0.8533	0.8583	4832
samples avg	0.7737	0.7737	0.7737	4832

Table 5: Results of the BiLSTM-CRF model on the test set of our corpus by entity type. Results are averaged over 10 runs with different seeds.

For the CamemBERT and D'AlemBERT results by entity type, we see almost the exact same results for both models which actually supports our hypothesis that due to the size of our corpus, the Transformer-based models might be "*forgetting*" some of their pre-training contemporary data and "*re-learning*" the training data of our corpus seen during fine-tuning. There is a small exception to this and it again the *production* entity type, we can see that D'AlemBERT performs a bit better for this particular type which might be explained by the presence of these in D'AlemBERT's pre-training data as opposed to the lack of it in Camem-

[6] We observe that SentencePiece tends to split OOV words by characters which might not be ideal for sequence-tagging tasks, specially for NER.

BERT's web-based pre-training corpus, suggesting that while these models might be "*forgetting*" while exposed to corpora of the size of our corpus, they can still leverage their pre-training data to a certain extent.

5 Conclusion

In this paper we have produced a significantly big, fine-grained NER annotated corpus for Early Modern French, as well as state-of-the-art models for coarse NER annotation in Early Modern French. We showed that adopting a data-driven approach in which one focuses on producing as much annotated data as possible as opposed to producing highly specialised machine learning architectures for NER, is a quite successful approach as we have obtained results for Early Modern French that far surpass the current state of the art for Contemporary French and that are on par with the current state-of-the-art specialised architectures for Contemporary English. The corpus that we have produced also opens many future perspectives of research, for instance, we hope that in the future we will be able to study the impact of the size of the fine-tuning data in the fine-tuning of Transformer-based models, something that could be easily achieved by iteratively fine-tuning different Transformer-based with subsets of our corpus of incremental size. Furthermore, one could also use all the other levels of annotation of our corpus to develop a specialised architecture capable of predicting all annotation layers at once. In the end, we hope that both our corpus and our fine-tuned models will be useful to researchers in both Natural Language Processing and Digital Humanities.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. *Building a Treebank for French*, pages 165–187. Springer Netherlands, Dordrecht.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Antonella Amatuzzi, Carine Skupien Dekens, Wendy Ayres-Bennett, Annette Gerstenberg, and Lene Schoesler. 2019. Améliorer et appliquer les outils numériques. ressources et approches pour l'étude du changement linguistique en français préclassique et classique. In *Le français en Diachronie*, Travaux de Linguistique Romane, pages 337–364. Editions de linguistique et de philologie.

Clément Besnier and William Mattingly. 2021. Named-entity dataset for medieval latin, middle high german and old norse. *Journal of Open Humanities Data*, 7:23. Publisher: Ubiquity Press.

Peter Blumenthal, Sascha Diwersy, Achille Falaise, Marie-Hélène Lay, Gilles Souvay, and Denis Vigier. 2017. Presto, un corpus diachronique pour le français des XVIe-XXe siècles. In *Actes de la 24ème conférence sur le Traitement Automatique des Langues Naturelles - TALN'17*. Association pour le traitement automatique des langues.

Julian Brooke, Adam Hammond, and Timothy Baldwin. 2016. Bootstrapped text-level named entity recognition for literature. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 344–350, Berlin, Germany.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Maud Ehrmann. 2008. *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Theses, Paris Diderot University.

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. Diachronic evaluation of NER systems on old newspapers. *Proc. of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 97–107.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification on historical documents: A survey. *arXiv:2109.11406 [cs]*.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers. *Experimental IR meets multilinguality, multimodality, and interaction. 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*.

Simon Gabay, Jean-Baptiste Camps, and Thibault Clérice. 2022. Manuel d'annotation linguistique pour le français moderne (XVIe -XVIIIe siècles). Version B (« Béate Béatrice »).

Simon Gabay, Thibault Clérice, Jean-Baptiste Camps, Jean-Baptiste Tanguy, and Matthias Gille-Levenson. 2020. Standardizing linguistic data: method and tools for annotating (pre-orthographic) french. In *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC '20)*.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From FreEM to D'AlemBERT: a Large Corpus and a Language Model for Early Modern French. *arXiv e-prints*, page arXiv:2202.09452.

Simon Gabay and Giovanni Pietro Vitali. 2019. A theatre of places: Mapping 17th c. french theatre. In *Proceedings of the 13th Workshop on Geographic Information Retrieval (GIR'18)*. ACM.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech 2009 - Proceedings of the 10th Annual Conference of the International Speech Communication Association*, pages 2583–2586.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. 2008. Named entity recognition for digitised historical texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA).

Yvonne Gwerder. 2017. *Named Entity Recognition in Digitized Historical Texts*. University of Zurich.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv e-prints*, page arXiv:1612.03651.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Eleni Kogkitsidou and Philippe Gambette. 2020. Normalisation of 16th and 17th century texts in French and geographical named entity recognition. In *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, pages 28–34, Seattle, Washington, USA.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, page arXiv:1907.11692.

Elaine Marsh and Dennis Perzanowski. 1998. MUC-7 evaluation of IE technology: Overview of results. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Robert Martin. 2020. *Dictionnaire du Moyen Français*. ATILF - CNRS & Université de Lorraine.

Franco Moretti. 2005. *Graphs, maps, trees : abstract models for a literary history*. Verso.

Franco Moretti. 2013. *Distant reading*. Verso.

Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4348–4352. European Language Resources Association (ELRA).

Damien Nouvel, Maud Ehrmann, and Sophie Rosset. 2015. *Les entités nommées pour le traitement automatique des langues*. ISTE Group.

Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary, and Benoît Sagot. 2020a. Establishing a new state-of-the-art for French named entity recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4631–4638, Marseille, France. European Language Resources Association.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020b. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Sophie Rosset, Cyril Grouin, Karën Fort, Olivier Galibert, Juliette Kahn, and Pierre Zweigenbaum. 2012. Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 40–48. Association for Computational Linguistics.

Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum. 2011. *Entités nommées structurées : guide d'annotation Quaero*. LIMSI-CNRS.

Benoît Sagot, Marion Richard, and Rosa Stern. 2012. Annotation référentielle du corpus arboré de Paris 7 en entités nommées (referential named entity annotation of the Paris 7 French TreeBank) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 535–542, Grenoble, France. ATALA/AFCP.

Stefan Schweter and Alan Akbik. 2020. FLERT: Document-Level Features for Named Entity Recognition. *arXiv e-prints*, page arXiv:2011.06993.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Sergio Torres Aguilar, Xavier Tannier, and Pierre Chastang. 2016. Named entity recognition applied on a data base of medieval latin charters. the case of chartae burgundiae. In *3rd International Workshop on Computational History (HistoInformatics 2016)*.

Claire Hélène Vachon. 2010. *Le Changement linguistique au XVIe siècle: une étude basée sur des textes littéraires français*. ELiPhi, Éditions de linguistique et de philologie.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.