DHQ: Digital Humanities Quarterly
2026
Volume 20 Number 3

Building open access, calibrated, syntactically annotated corpora over the history of French: A Case study of the application and adaptation of annotation tools for historical syntax

Pierre Larrivée <pierre_dot_larrivee_at_unicaen_dot_fr>, University of Caen https://orcid.org/0000-0001-8447-7102
Natasha Romanova <natalia_dot_romanova_at_unicaen_dot_fr>, University of Caen https://orcid.org/0000-0002-5723-9819
Mathieu Goux <mathieu_dot_goux_at_unicaen_dot_fr>, University of Caen https://orcid.org/0000-0003-4211-8309
Rayan Ziane <rayan_dot_ziane_at_univ-orleans_dot_fr>, University of Orléans https://orcid.org/0009-0000-4940-0169
Morgane Pica <morgane_dot_pica_at_unicaen_dot_fr>, University of Caen https://orcid.org/0000-0002-0981-4516

Abstract

This case study presents challenges and solutions for building calibrated syntactically annotated diachronic corpora of French. Through a series of integrated projects that developed methods for syntactic annotation of medium and large collections of non-literary texts over a long period of time (13th – 19th centuries), we consider how the availability of tools and the emergence of research questions shape and direct corpus-building activity. Supported by the experience of an earlier project, a semi-automated workflow for syntactic annotation was developed to allow processing of two corpora calibrated by provenance and genre, with a focus on parts-of-speech tagging in three annotation frameworks. The workflow combines the use of a neural dependency parser, a lexicon of lemmata and forms and a Python library. This experience, in turn, gave rise to a new initiative that aimed at formulating and testing best practices for creating syntactically annotated *treebanks* for different periods of the history of the language.

Introduction[1]

The study of the evolution of language over several centuries is facilitated by access to digital collections of syntactically annotated texts. Such *corpora* allow searching for syntactic structures automatically in order to identify particular configurations. Annotation of lemmata, parts of speech (PoS) and syntactic functions make it possible for researchers to compare and contrast the shape of language through time, explain the observable patterns and test general hypotheses. Such enrichment by lexical and grammatical metadata can be achieved by modern computer-assisted methods, including, most recently, deep-learning tools such as neural network parsers. For the history of French language, a number of annotated resources exist. However, they: 1

1. are usually limited to a particular period, which impedes comparative work across the history of the language; [2]
2. consist predominantly of literary material which exhibits conservative traits and is believed to be removed from the everyday language [Balon and Larrivée 2016] [Pinzin and Goux forthcoming];
3. lack accurate annotation of syntactic functions, which means that even after the extraction of desired configurations of parts of speech considerable investment is often needed to review the data manually.[3]

In the present contribution, we discuss how a series of integrated projects conducted in 2018-2024 at the CRISCO Laboratory at the University of Caen (France) set out to address these challenges making use of language processing technologies available to researchers at each stage.[4] 2

Corpora

Supporting research into the history of French syntax, the three projects presented here sought to integrate existing tools and develop methodologies for successful management and syntactic annotation of medium to large diachronic *corpora*. To limit potential documentation biases, all three share a common approach to the selection of data, opting for texts of a given non-literary genre and produced in the same geographical area (Normandy), to control for extraneous variation (Table 1). The ConDÉ project (2018-2021)[5] produced a large digitized corpus of Norman customal law texts (*coutumiers normands*) from the 13th to the 19th century, based on a method trialed on a set of historical letters (EPELE).[6] Building on the experience of ConDÉ, High-Tech project (2021-2023) gave rise to a collection of chronicles and historical treatises from the 12th to the 19th centuries, with one text per century.[7] In addition, as part of the Franco-German MICLE project (2021-2024), a further medium-sized corpus of legal texts was elaborated which includes trial accounts and *styles de procéder* from the 12th to the 17th centuries, with one text per half-century.[8] Below, we refer to the corpus of customal law as *ConDÉ*, to the corpus of chronicles produced as part of the High-Tech project as *Chroniques* and to the corpus of trial accounts and *styles de procéder*, created by the French team of the MICLE project, as *MICLE-Fr*. The semi-automated workflow for lemmatization and annotation developed by the High-Tech project and tested during the annotation of *Chroniques* and *MICLE-Fr corpora* is referred to as the HT-CRISCO workflow. The annotated collections can be consulted via the CRISCO Laboratory's TXM server and two dedicated websites.[9] Most texts are out of copyright and can be consulted either in full or via a concordance, whereas for a small minority of sources only concordance view is available.

Project/Funding Period/Funder	Corpus	Text Types	Period	Number of Tokens	Annotation
ConDÉ 4 2018-2021 5 (Réseau d'Intérêts Normands) 6	*ConDÉ*	Norman customals	13th-19th cent.	4,452,540 tokens	7 *lemmata *PoS (Presto) 8
MICLE 9 2021-2024 10 (ANR/DFG Franco-German grant) 11	*MICLE-Fr*	Legal texts (trials) 12 Legal treatises (*styles de procéder*) 13	13th-17th cent.	422,117 tokens	14 *lemmata *PoS (UD, UPenn, Presto) 15 16 *syntactic functions UD (automatic)

3

High-Tech 17 2021-2023 18 (Réseau d'Intérêts Normands) 19	*Chroniques*	Chronicles 20	12th-19th cent.	313,518 tokens	*lemmata 21 *PoS (UD, UPenn, Presto) 22 *syntactic functions UD (automatic) 23

Table 1. Diachronic corpora of Medieval French projects at the CRISCO Lab 2018-2024

Methodology

The different stages of processing of the data, from image to annotated text, are described below. The major methodological and practical challenges include transcription and digitization of large amounts of textual data from heterogeneous original sources, lemmatization and syntactic annotation. Appropriate tools need to be selected to process the data at each stage. Different types of software use different data formats; therefore, conversion scripts need to be provided to ensure interoperability. In addition, choices need to be made about when linguists intervene to correct outputs of automatic processing. 24

Transcription

All corpora were digitized from reproductions of manuscripts or scans of printed editions using OCR and HTR technologies made available via the Transkribus portal. [10] Transkribus allows the use of pre-trained transcription models and training of new models adapted to particular sources. Creating custom-made models was especially necessary in the case of the ConDÉ project, which processed a mass of material: the two volumes by Basnage (1678), for example, run up to a million characters. As part of *MICLE-Fr*, one particular challenge was the complete and fully verified transcription of a 16th-century manuscript of witness statements from the island of Guernsey Greffe Crime 1 (more than 41,000 tokens).[11] 25

The diversity of sources from across several centuries, ranging from medieval manuscripts to modern critical editions, not to mention early printed books, meant that the texts of the *corpora* presented considerable variation where spelling of individual words were concerned. The decision to remain as close to the original as possible and not to normalize the spelling was taken for all projects as the tools available permitted to accommodate variation. Lemmatization allowed grouping all grammatical forms and all spelling variants of the same word under the same headword. For example, in the *Chroniques* corpus, the French word *king* can be encountered under the forms *rei*, *reis*, *rey*, *reys*, *roi*, *rois*, all lemmatized under the headword *roi*. 26

File formats

Once digitized, the texts of *ConDÉ*, *Chroniques* and *MICLE-Fr corpora* were converted into the XML-TEI format [DeRose 1999]. This format permits encoding editorial interventions such as corrections of the original (e.g., in case of repetition of words). The use of this format facilitated manual correction of the data throughout the annotation pipeline described below and allowed integration of the final versions of the annotated files into the TXM-portal and websites to enable consultation and interrogation of the *corpora*. Texts were processed and corrected individually, with a different file corresponding to each. 27

The chosen format allowed reflecting the structure of the original (i.e., the manuscript or printed version that had been digitized) in the structure of the XML document. According to the XML-TEI hierarchy, a text can be divided into books, a book can be divided into sections, a section into chapters, and a chapter into paragraphs; the corresponding tags made it possible to indicate where paragraphs and larger sections of the text begin and end.[12] 28

For *ConDÉ*, only XML-TEI format was used throughout all stages of processing. For *Chroniques* and *MICLE-Fr*, files needed to be converted into the CONLL-U tabular format that is used by the syntactic parsers twice during the execution of the workflow (see 3d-2 below). To reintegrate the parsed data into the XML-TEI corpus, CONLL-U files were converted back into the XML-TEI format. The synchronization between the XML-TEI version and CONLL-U version of the files was done via the assignment of a unique ID to each sentence of the text. Thus, sentence 1 in paragraph 3 of chapter 34 in section 1 of book 2 would be given sentence ID “2_1_34_3_1”. This number would, in turn, allow finding the place of the sentence in the XML-TEI structure after parsing is completed. 29

Segmentation

Prior to any further treatment, the texts were tokenized and, in the case of *Chroniques* and *MICLE-Fr*, segmented into sentences (using strong punctuation as a prompt for automatic segmentation, followed by manual revision). For the ConDÉ project, annotation was done exclusively on the token level, whereas for High-Tech and MICLE-French sentence segmentation was an essential step towards preparing the data for syntactic parsing using an automatic tool. 30

Annotation

Considering the nature of the data and of the research needs as well as the availability of annotation tools, the sequence of projects followed two different yet complementary axes: 31

1) increased accuracy of the annotation and formalization of the correction process; 32

2) introduction of more layers of annotation. 33

Where *ConDÉ* is lemmatized and annotated in PoS using one set of tags, *Chroniques* and *MICLE-Fr* are lemmatized, PoS-annotated in three different systems, adding a layer of morphological annotation, and (automatically) annotated in syntactic functions. At every stage of this process, new tools, resources and methodological approaches were integrated as they became available to the linguistic community. 34

Lemmatization and PoS tagging (*ConDÉ*)

Following an extensive campaign of digitization, conversion to XML-TEI and tokenization, the texts of the *ConDÉ* corpus were lemmatized and PoS-annotated. These two operations were accomplished using the Presto lexicon. This lexicon takes the form of a list of possible (attested and computer-generated) word forms of modern and historical French *lemmata*, with an indication of the grammatical category (using the Presto *tagset*, a set of labels particularly adapted to French) of each word [Blumenthal et al 2017] [Lay and Pincemin 2010].[13] The information contained in the lexicon was compared to the data in the XML-TEI file using a dedicated Python script. 35

While the use of the Presto lexicon and the script allowed matching the form of the word with a possible lemma and PoS, it did not make use of contextual information. It neither disambiguated when there was more than one lemma or PoS speech possible, nor offer a solution when the form was not listed. In order to address such cases, a programme of semi-manual correction was put in place to reduce manual intervention. [14] For example, the French word form *en* can 36

have the role of a preposition (usually when followed by a common or proper noun or by a determiner) or that of a pronoun (usually preceding a verb), and disambiguating the two is important for identifying specific kinds of constructions (e.g., prepositional noun phrases or verbs preceded by a pronoun, see Figure 1). Similarly, the word *fait* may be a noun (*fact*) to be classified under the lemma *fait* or a form of the verb (*to do*, *to make*) to be lemmatized under *faire*.

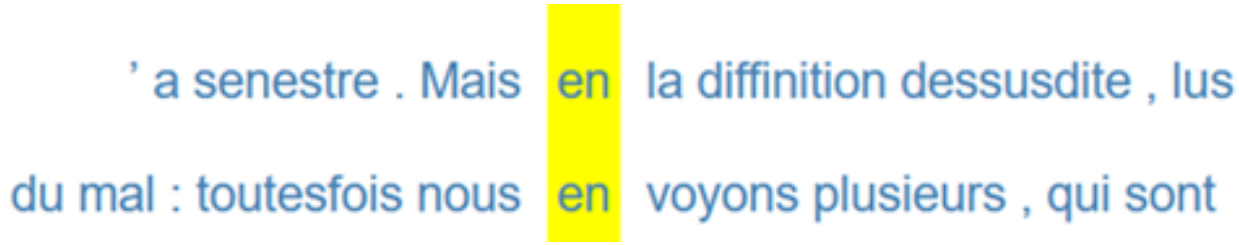


Figure 1. Example of a concordance search for *en* in ConDÉ (in the Coutumier by Terrien 1578). The first one is a preposition; the second one is a pronoun.

Even though, given the size of the corpus (just under four and a half million tokens), a comprehensive manual correction was never conducted, it proved to be a reliable resource for tracing grammatical change [Goux 2022]. 37

3d-2. Neural dependency parsing for annotation of PoS, lemmata, syntactic functions (*Chroniques* and *MICLE-Fr*)

The challenges identified during work on *ConDÉ* were addressed as part of the elaboration of the annotation pipeline for *Chroniques* and *MICLE-Fr corpora*, that resulted from two projects that run in parallel from 2021. Whereas the former focused on methodological issues (creation of an annotation workflow and a user-friendly consultation site), the latter was constituted around the research question that aimed at clarifying the evolution of the Verb Second (V2) word order in a diachronic perspective. The availability of tools and resources, on the one hand, and the research need for simplified access to annotated verb forms and subject and object nominal phrases, on the other, determined the choice of the annotation to be added and the accuracy level of correction. Where in *ConDÉ* annotation centered on linear succession of the tokens and on a set of rules and light-touch manual interventions to disambiguate ambivalent forms, the *HT-CRISCO* workflow, also used to annotate *MICLE-Fr*, put at its core a syntactic parser that operates both at the level of the word and sentence. Graph-based dependency parsers inspired by Dozat and Manning's architecture [Dozat and Manning 2017] are computer programmes that can use models trained on pre-annotated *corpora* to automatically annotate other texts in PoS (*tagging*) and syntactic functions (*parsing*) in the Universal Dependencies (UD) framework [de Marneffe et al 2021]. The accuracy of annotation depends on the proximity of the language on which the model was trained (*training corpus*) to the *target corpus* (corpus to be annotated). 38

The workflow developed and tested by the two project teams relied on a succession of automatic stages followed by manual checks (Figure 2).[15] The use of an automatic parser allowed to conduct PoS tagging using UD parts of speech to provide a basis and necessary contextual disambiguation for lemmatization and refinement of PoS and morphological information (using UPenn [Santorini 2007] and Presto [Blumenthal et al 2017] *tagsets*) at later stages of the processing. Manual revision of the data occasionally led to corrections (e.g., of OCR or HTR errors) or changes in sentences segmentation and tokenization, which meant that the automatic parsing (identification of the syntactic head and function of each token) needed to be performed again to take these changes into account (manually verified PoS tags and lemmata were preserved). 39

Among existing dependency parsers, HoPS parser was selected because in 2021 the developers of this tool had made available a model trained on a corpus of Old and Middle French literary texts, and a large proportion of the data to be annotated dated from the medieval period. Contextual analysis provided by HoPS allowed successfully distinguishing between forms that otherwise would have remained ambiguous. To take examples cited before, HoPS was very successful in distinguishing *en* prepositions (for example *en Normandie* "in Normandy") and pronouns (*et en receipt le roy l'hommaige* "and of-it received the king hommage", an example from a 1373 text from *Chroniques*). Similarly, in *Je* 40

dy qu'il a bien prouvé tel fait "I say that he proved this fact well" it was possible to correctly tag *fait* as a noun and lemmatize under *fait* whereas in *L'adjournement en cas de dolléance est fait en ceste manière* "in case of grievance the adjournment is done in this manner" the same word was tagged as a past participle of the verb *faire* (both examples from the 1425 text, *MICLE-Fr*). The overall performance of automatic parsing on the prediction of parts of speech (*tagging*) across a very diverse corpus has been very satisfactory.[16]

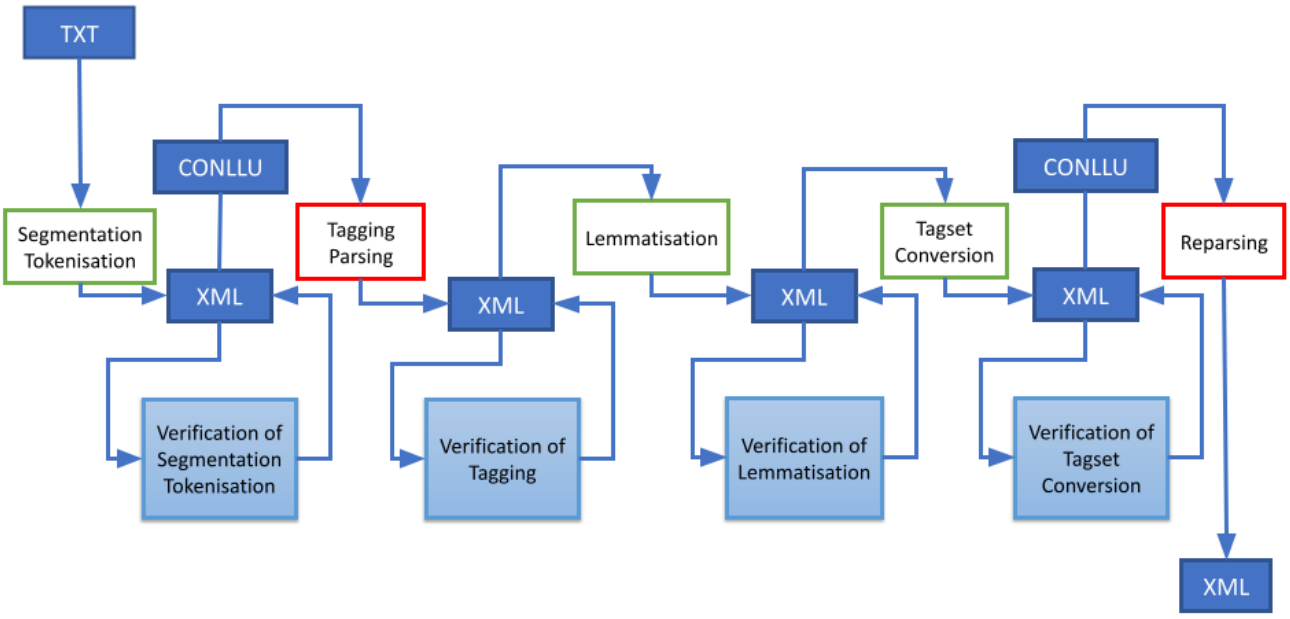


Figure 2. HT-CRISCO Semi-automated workflow for syntactic annotation of diachronic corpora of French (https://github.com/Corpus-Diachroniques-CRISCO/HT-CRISCO)

Results

The absence of annotation of syntactic functions in *ConDÉ* limited research potential of the corpus to an extent (in part compensated by the corpus size). Thus, the corpus was highly relevant for the search of particular items, such as relatively infrequent postverbal clitic pronouns [Olivier and Folli 2023]. Syntactic queries called for more elaborate strategies. An example is the analysis of the so-called *bare* nouns in the corpus [Larrivée and Goux 2024]. Whereas French nouns could be used without an article (also known as *determiner*) in its early history, as is the case in other Romance languages and in Latin, that possibility was severely reduced over time. Because the search could not refer to functions that would have allowed us to pick up the boundaries of the noun phrase, finding nouns without a determiner had to be realized using approximate heuristics. The strategy was to look for nouns that were not immediately preceded by a determiner. Nouns preceded by an adjective before the article were therefore counted as *bare* (i.e., nouns used without an article). We had to be content with the fact that the same search was applied equally to all texts, meaning that the same systematic errors would creep in, allowing meaningful comparison. The alternative would have been a manual examination of data, and, while this was impractical given the thousands of occurrences concerned, that could not ensure that the occurrences not returned did not contain some relevant examples. 41

Modern neural network parsers provide statistics-based predictions for annotation not only of PoS (*tagging)* but also of syntactic functions (*parsing*). In the UD approach, parsing consists in the identification of a series of asymmetric syntactic relations between two tokens with one being the *head* and the other the *dependent*; and in the nature of the relation, defined by labels such as *nsubj* (subject), *obj* (direct object) or *advcl* (adverbial clause). Such annotation is most susceptible to error, especially when the distance between the *training corpus* used to create the annotation model and the annotated corpus is important (e.g., when the *corpora* belong to different dialects, chronological periods and/or textual genres; for example, models for Modern French would be less successful on Medieval French material). One text-internal factor that we have observed to multiply mistaken syntactic annotation is sentence length [Ziane and 42

Romanova 2024] [Daoudi et al 2026]. Despite its considerable value for linguistic research, syntactic annotation would thus require most resources, time and expertise to be corrected. The creation of manually checked *treebanks* being outside the scope of *ConDÉ*, *Chroniques* and *MICLE-Fr* projects, we nevertheless empirically observed that automatic annotation of key functions such as subjects and objects was quite reliable and, combined with PoS annotation and lemmatization, allows structures such as postverbal subjects [Larrivée et al 2024] and preposed objects [Larrivée 2024] to be identified and to understand and explain their patterns of evolution.[17]

43 At the center of our preoccupations when setting up and implementing the diachronic *corpora* projects at the CRISCO lab has been the promotion of accessibility of processes and alignment to international practice. An example was the development as part of the High-TECH project of a visualization and search portal to make resources readable and searchable in an ergonomic format.[18] We adopted UD format and tools to contribute to the development of both the historical French corpus field, and the international treebank movement, while providing tags in other annotation systems.

44 Finally, the annotation process and the scripts developed as part of *HT-CRISCO* workflow can be used for lemmatization and PoS-tagging of new diachronic *corpora* in French to achieve annotation that can be used for exploring texts of any period of the history of the language.

Research perspectives

45 Broader implications of the projects discussed here are found in highlighting shortcomings in the current approaches to *corpora* building. The combined *Chroniques* and *MICLE-Fr corpora* result in a collection of just under a million tokens across eight centuries, reflecting diachronic variation in the calibrated conditions of controlled provenance and text type, lemmatized and annotated in PoS. The next stage is achieving reliable annotation of syntactic functions to enable cross-textual analysis that can be done *in one click* without a need for approximation or manual post-processing. In addition to the costly demands in human effort to correct large *corpora*, we are faced with two main interconnected challenges:

1. Lack of training *corpora* for creating models for all periods of the history of the French language.

   The texts of *Chroniques* and *MICLE-Fr* present considerable diachronic variation, ranging as they are from the 12th to the 19th century. As far as the choice of annotation model is concerned, in the UD collection, only the medieval and the contemporary period have reference *corpora* for French language.[19]
2. Lack of consistency in the annotation principles for Medieval and Modern French which impedes training of models and analysis. In the UD collection, Old/Middle French and Modern French are treated as different languages with slightly divergent annotation guidelines, e.g., where the annotation of modal verbs is concerned. For example, for Old/Middle French, modal verbs (*devoir*, *pouvoir*, *souloir*) [20] are treated as auxiliaries of the main lexical verb in the infinitive (e.g., *Il peut travailler* “He can work”). For modern French, on the other hand, the modal verb is the head of the clause, and the infinitive of the lexical verb is the head of an open clausal complement (*xcomp*). For reasons of consistency of annotation and given that the majority of our texts dates from before the 17th century, we used the pre-trained model for Old French supplied by the creators of the HoPS parser [Grobol et al 2022] and we have thus annotated modal verbs as auxiliaries across the corpus but more discussion about homogenizing guidelines for different periods of French (and other Romance languages) is needed.[21]

47 In addition to HoPS, several graph-based syntactic parsers are currently available to researchers, including UDify [Straka et al 2016] [Guiller 2020]. UDify can be used via a user-friendly portal UDPipe and BertForDeprel is integrated into an annotation and graph rewriting tools ArboratorGrew[22]. The latter also allows retraining models and adapting them to the text or corpus being annotated, by progressively adding annotated material to improve the quality of the annotation at each iteration, using a *bootstrapping* approach [Peng et al 2022] [Romanova et al 2025(a)]. [23]

Funding

Projects presented in this paper have received funding from Normandy Region (RIN: Réseau d'Intérêts Normands): ConDÉ, High-Tech, and from the ANR: Agence Nationale de Recherche (*MICLE-Fr* as part of the ANR-DFG-funded Franco-German scheme). Support was also brought by the Institut universitaire de France.

Notes

[1] An earlier version of this paper was presented at the "Journée d'études Retours d'expériences en édition numérique des textes", organised by Pierre-Jean Suriac, Angela Goebel and Morgane Pica on 22 June 2023 at Université Jean Moulin Lyon 3.

[2] One notable exception is the Frantext corpus that, as of June 2026, contained 275 million words in French from the 10th to the 21st centuries from a range of genres. The corpus is lemmatized and annotated in parts of speech (https://www.frantext.fr/, accessed 1 June 2026). Among existing resources for the medieval period, we can cite the Base du Français Médiéval (URL: http://bfm.ens-lyon.fr/, accessed 8 June 2026 [Guillot-Barbance et al 2017], MCVF-PPCHF [Martineau et al 2021] and PhraseoRoche [Denoyelle et al 2024]. For the 16th-17th centuries, SERMO *corpus* (URL: http://sermo.unine.ch/SERMO/, accessed 8 June 2026).

[3] For Old and Middle French two important exceptions are two *corpora* of literary texts, MCFV-PPCHF [Martineau et al 2021], annotated according the UPenn syntactic framework, and Profiterole (formerly known as SRCMF) corpus annotated in Universal Dependencies (UD) [Prévost and Stein 2013] [Prévost et al 2024].

[4] The methodological and practical challenges associated with the processing of the textual data from digitization to enrichment with annotation and the solutions proposed are presented in detail in the Methodology section.

[5] ConDÉ: CONstitution d'un Droit europÉen: six siècles de coutumiers normands [Goux and Pica 2019]. The project data and software have been deposited on GitHub https://github.com/RIN-ConDE/editions (accessed 8 June 2026). A revised version of the ConDÉ corpus is currently under construction as part of a PhD thesis [Pica in preparation].

[6] The EPELE (Écriture des peu lettrés: Français vernaculaire dans la Normandie médiévale) corpus can be consulted via the project website https://pdn-lingua.unicaen.fr/epele/epele/accueil (accessed 8 June 2026).

[7] High-Tech: High-level text annotation across historical texts: improving semi-automatization of big data management. URL: https://www.unicaen.fr/projet_de_recherche/high-tech/ (accessed 8 June 2026).

[8] The Franco-German project MICLE (MIcro-Cues of Language Evolution) addressed the question of the change of syntactic structure in two Romance language varieties, French of Normandy and Venetian from the origins to the 17th century (URL: https://www.unicaen.fr/projet_de_recherche/micle, accessed 8 June 2025). The German team at the University of Frankfurt (Cecilia Poletto and Francesco Pinzin) created an annotated corpus of medieval and early modern Venetian legal texts annotated in UPenn framework. This Venetian corpus (MICLE-VEC) is partially available via the CRISCO Lab TXM portal (URL: https://txm-crisco.huma-num.fr/txm/, accessed 8 June 2026). See [Goux and Pinzin 2024] [Goux 2024(b)].

[9] ConDÉ project website: https://mrsh.unicaen.fr/coutumiers/conde/accueil.html; consultation website for the *Chroniques* and *MICLE-Fr corpora* URL: https://criscoht.unicaen.fr/ (accessed 8 June 2026).

[10] URL: https://www.transkribus.org/ (accessed 8 June 2026). See Nockels et al 2022 [Nockels et al 2022].

[11] The transcription formed part of an internship project in the spring semester 2021-2022 to which student transcribers Agathe Aubert, Lucie-Marie Leblanc, Marie Picard and Valentin Simenel contributed. This source was the object of an MA dissertation in History at the University of Caen [Le Squer 2024].

[12] On the importance of not losing sight of the materiality of the original, including the structure of the text, when creating large linguistic *corpora* see [Goux 2024(a)].

[13] The Presto dictionary is currently available for download in its original version (https://unicloud.unicaen.fr/index.php/s/NSkPrcaZ3Rx2t9P) and the version revised by the team at CRISCO (https://unicloud.unicaen.fr/index.php/s/YgfYJenQMKD8bEC) (accessed 8 June 2026).

[14] The scripts are available at https://github.com/RIN-ConDE/tools/tree/main/corpus-construction/disambiguate-lemmatization-in-corrected-file (accessed 8 June 2026).

[15] For the full description and Python scripts see https://github.com/Corpus-Diachroniques-CRISCO/HT-CRISCO (accessed 8 June 2026). For further information on the workflow, see Ziane and Romanova, 2024 [Ziane and Romanova 2024].

[16] Recent quantitative tests conducted by the team on samples of two texts of the High-Tech corpus using HoPS, UDify and BertForDeprel parsers have shown performances between 87,12% and 90,24% depending on the tool and training conditions.

[17] In addition, during the annotation process, automatic annotation of functions was used to automatize disambiguation of some verb forms during the conversion from UD to UPENN and Presto *tagsets*: e.g., if the form *fait* that can be a finite form or a past participle of the verb *faire* has an auxiliary *avoir* or *être* we consider it to be a past participle [Ziane et al 2023].

[18] https://criscoht.unicaen.fr/ (accessed 1 June 2026).

[19] For the comparative statistics of the Modern French UD treebanks, see https://universaldependencies.org/treebanks/fr-comparison.html (accessed 8 June 2026). Profiterole Old French and Profiterole Middle French *corpora* (https://universaldependencies.org/treebanks/fro_profiterole/index.html and https://universaldependencies.org/frm/index.html, accessed 8 June 2026) were formerly known as SRCMF [Prévost and Stein 2013]. HoPS parser has a pre-trained model based on the previous version of the Old and Middle French corpus (https://github.com/hopsparser/hopsparser/blob/main/docs/models.md, accessed 8 June 2026).

[20] In the Profiterole corpus, *vouloir* is no longer considered as a modal verb/auxiliary.

[21] https://zenodo.org/records/7708976 (accessed 8 June 2026).

[22] https://grew.fr/ (accessed 8 June 2026).

[23] Outside the scope of the present contribution, building on the experience of previous corpus-building initiatives, Automated project (2023-2025), led by Pierre Larrivée and funded by Normandy Region, set out to test the bootstrapping approach to some of the texts of the *MICLE-Fr* corpus (on recent experiments on adapting models for one text of the *MICLE-Fr corpus*, see [Ziane and Romanova 2024]). Two small treebanks of Old and Middle French (URL: https://github.com/UniversalDependencies/UD_Old_French-ALTM and https://github.com/UniversalDependencies/UD_Middle_French-ALTM, accessed 8 June 2026) were produced as well as a larger treebank of 16th-century French (URL: https://github.com/UniversalDependencies/UD_French-ALTS, accessed 8 June 2026). On the experience of syntactic annotation of an Old Gascon treebank without a preexisting model, see [Romanova et al 2025(b)],

Works Cited

Balon and Larrivée 2016 Balon, L. and Larrivée, P. (2016) “L’ancien français n’est déjà plus une langue à sujet nul – nouveau témoignage des textes légaux.” *Journal of French Language Studies* 26(2): 221-237.

Blumenthal et al 2017 Blumenthal, P., Diwersy, S., Falaise, A., Lay, H., Souvay, G. and Vigier, D. (2017) “Presto, un corpus diachronique pour le français des XVIe-XXe siècle.” *Traitement Automatique des Langues Naturelles (TALN)*. June 2017, Orléans, France. 18-26. URL: https://taln2017.cnrs.fr/wp-content/uploads/2017/06/actes_ACor4French_2017.pdf (accessed 8 June 2026).

Daoudi et al 2026 Daoudi, K., Dehouck, M., Ziane, R. and Romanova, N. (2025) “Explicit Edge Length Coding to Improve Long Sentence Parsing Performance.” *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages* Varna, Bulgaria, pp. 102-110. URL: https://aclanthology.org/2025.lowresnlp-1.11 (accessed 8 June 2026).

DeRose 1999 DeRose, S. (1999) “XML and the TEI.” *Computers and the Humanities*, 33(1): 11-30. URL: https://doi.org/10.1023/A:1001771114509 (accessed 8 June 2026).

de Marneffe et al 2021 de Marneffe, M.-C., Manning, C. D., Nivre, J. and Zeman, D. (2021) “Universal Dependencies.” *Computational Linguistics*, 47(2): 255–308. URL: https://doi.org/10.1162/coli_a_00402 (accessed 8 June 2026).

Denoyelle et al 2024 Denoyelle, C., Kraif, O., Mounier, P., Renwick, A., Sorba, J. and Souvay, G. (2024) “Le corpus PhraséoRoChe: les défis de l’établissement des textes et de l’hétérogénéité des états de la langue.” *Corpus*, 25. 18 pp. URL: https://journals.openedition.org/corpus/8501 (accessed 8 June 2026).

Dozat and Manning 2017 Dozat T.C. and Manning, C. D. (2017) “Deep Biaffine Attention for Neural Dependency Parsing.” *International Conference on Learning Representations (ICLR) 2017.* 8 pp. URL: https://arxiv.org/abs/1611.01734 (accessed 8 June 2026).

Goux 2022 Goux, M. (2022) “Le temps long : l'évolution du français dans un corpus textuel calibré. Le témoignage de la coutume de Normandie.” *Studia Linguistica Romanica*, 8. 9 pp. URL: https://studialinguisticaromanica.org/index.php/slr/article/view/99 (accessed 8 June 2026).

Goux 2024(a) Goux, M. (2024) “De très grands corpus pour l’étude diachronique du français : annotations, informations métalinguistiques et paratextes.” *Humanités numériques*, 9. 21 pp. URL: https://journals.openedition.org/revuehn/3930 (accessed 8 June 2026).

Goux 2024(b) Goux, M. (2024) “Enjeux des corpus bilingues en diachronie longue : l’exemple du projet MICLE.” *Corpus*, 25. 13 pp. URL: https://journals.openedition.org/corpus/8468 (accessed 8 June 2026).

Goux and Pica 2019 Goux, M. and Pica, M. (2019) “Le projet ConDÉ : présentation. Les défis d’un corpus de textes en diachronie longue.” Conference slides. URL: https://hal.science/hal-02447030 (accessed 8 June 2026).

Goux and Pinzin 2024 Goux, M. and Pinzin, F. (2024) “Challenges of a Multilingual Corpus (Old French/Old Venetian): The Example of the MICLE project.” Fontana A. and Pezzini E. Franco Cesati (eds.) *Venezia e la Francia tra Medioevo ed età Moderna. Similitudini, specificità, interrelazioni*. Florence : Cesati Editore, pp. 153-175.

Grobol et al 2022 Grobol, L., Regnault, M., Ortiz Suarez, P., Sagot, B., Romary, L. and Crabbé, B. (2022) “BERTrade: Using Contextual Embeddings to Parse Old French.” Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, C., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J. and Piperidis, S. (eds.) *Proceedings of the Thirteenth Language Resources and Evaluation Conference* European Language Resources Association, pp. 1104–1113. URL: https://aclanthology.org/2022.lrec-1.119 (accessed 8 June 2026).

Guiller 2020 Guiller, K. (2020). *Analyse syntaxique automatique du pidgin-créole du Nigeria à l’aide d’un transformer (BERT): Méthodes et Résultats.* MA Dissertation, Sorbonne Nouvelle.

Guillot-Barbance et al 2017 Guillot-Barbance, C., Heiden, S. and Lavrentiev, A. (2017) “Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique.” *Diachroniques*, 7: 168-184. URL: https://shs.hal.science/halshs-01809581 (accessed 8 June 2026).

Larrivée 2024 Larrivée, P. (2024) “Deux déterminismes du déclin des objets antéposés dans l'histoire du français.” *9ème Congrès Mondial de Linguistique Française*, Université de Lausanne. URL: 10.1051/shsconf/202419103001 (accessed 8 June 2026).

Larrivée and Goux 2024 Larrivée, P. and Goux, M. (2024) “The evolution of bare nouns in the history of French. The view from calibrated corpora.” *Journal of French Language Studies*, 34(4): 323 - 350. URL: https://www.cambridge.org/core/journals/journal-of-french-language-studies/article/evolution-of-bare-nouns-in-the-history-of-french-the-view-from-calibrated-corpora/A809D923F7CD0854DBB9833DFAB4C428 (accessed 8 June 2026).

Larrivée et al 2024 Larrivée, P., Poletto, C., Pinzin, F. and Goux, M. (2024) “Asymmetry as a general cue for V2 (loss).” *Isogloss: Open Journal of Romance Linguistics*, 10(7). 22 p. URL: https://hal.science/hal-04839429 (accessed 8 June 2026).

Lay and Pincemin 2010 Lay, M.-H. and Pincemin, B. (2010) “Pour une exploration humaniste des textes: AnaLog.” *Proceedings of 10th International Conference Journée d’Analyse statistique des Données Textuelles 9-11 Juin 2010 – Sapienza University of Rome*. Bolasco, S., Chiari, I. and Giuliano, L. (eds.) V.2, pp. 1045-1056.

Le Squer 2024 Le Squer M. (2024) *Le registre « Crime » 1563-1569 au Greffe de Guernesey.* MA, University of Caen.

Martineau et al 2021 Martineau, F., Hirschbühler P., Kroch A. and Morin Y.Ch. (2021) MCVF Corpus, parsed, version 2.0. URL: https://github.com/beatrice57/mcvf-plus-ppchf (accessed 8 June 2026).

Nockels et al 2022 Nockels J., Gooding P., Ames, S. and Terras M. (2022) “Understanding the application of handwritten text recognition technology in heritage contexts: A systematic review of Transkribus in published research.” *Archival Science*, 22(3): 267-392.

Olivier and Folli 2023 Olivier, M., Sevdali, C. and Raffaella, F. (2023) “Clitic climbing and restructuring in the history of French.” *Glossa: a journal of general linguistics*, 8(1): 1–45.

Peng et al 2022 Peng, Z., Gerdes, K. and Guiller, K. (2022) “Pull your treebank up by its own bootstraps.” Becerra L., Favre B., Gardent C. and Parmentier Y. (eds.) *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*. Pp. 139–153. URL: https://hal.science/hal-03846834 (accessed 8 June 2026).

Pica in preparation Pica, M. (in preparation) *L'intertextualité dans les commentaires sur la coutume de Normandie, miroir de l'univers intellectuel des praticiens du droit à l'époque moderne. Apprentissage automatique et triplets RDF au service de la recherche en Histoire*. (Thesis in preparation at the Université Paris Sciences et Lettres under the supervision of Arabeyre P. and Hodel T.) URL: https://theses.fr/s378092 (accessed 8 June 2026).

Pinzin and Goux forthcoming Pinzin, F. and Goux, M. “How genre affects word order: a diachronic analysis of French.” Larrivée, P. and Pinzin, F. (eds.) *Syntactic change through text-types.* Berlin: de Gruyter.

Prévost and Stein 2013 Prévost, S. and Stein, A. (2013) “Syntactic annotation of medieval texts: the Syntactic Reference Corpus of Medieval French (SRCMF).” Bennett, P., Durrell, M., Scheible, S. and Whitt, R. (eds.) *New Methods in Historical Corpus Linguistics*, Corpus Linguistics and International Perspectives on Language. Tübingen: Narr Verlag, pp. 275-282.

Prévost et al 2024 Prévost, S., Grobol, L., Dehouck, M., Lavrentiev, A. and Heiden. S. (2024) “Profiterole: un corpus morpho-syntaxique et syntaxique de français médiéval.” *Corpus* 25. 25 pp. URL: https://journals.openedition.org/corpus/8538 (accessed 8 June 2026).

Romanova et al 2025(a) Romanova, N., Larrivée, P. and Ziane, R. (2025) *Procedure for semi-automatic parsing of Romance corpora (Version 1).* Zenodo. 28 pages. URL: https://zenodo.org/records/17737727. (accessed 8 June 2026).

Romanova et al 2025(b) Romanova, N., Ziane, R. and Francioni, B. (2025) “Adaptation of models for parsing of Old Gascon.” *Lift2-2025: Journées scientifiques du réseau thématique LIFT2 – linguistique informatique, formelle et de terrain, GDR LIFT.* 9 pp. URL: https://hal.science/hal-05338944 (accessed 8 June 2026).

Santorini 2007 Santorini, B. (2007) “Protocole d'étiquetage - Parties du discours (PDD).” URL: https://www.ling.upenn.edu/~beatrice/corpus-ling/annotation-french/pos/pos-index.html (accessed 8 June 2026).

Straka et al 2016 Straka, M., Hajič, J. and Straková, J. (2016) “UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing.” *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, Slovenia, pp. 4290–4297.

Ziane and Romanova 2023 Ziane, R. and Romanova, N. (2023) “Vers l’intégration des outils d’annotation syntaxique:

proposition d’une chaîne de traitement itérative pour faciliter l’adoption et l’accès aux technologies d’apprentissage automatique.” *Actes des 11èmes Journées Internationales de Linguistique de Corpus, 3-7 juillet 2023*, pp. 278-383. URL: https://jlc2023.sciencesconf.org/data/pages/abstracts_JLC_2024.pdf (accessed 8 June 2026).

Ziane and Romanova 2024 Ziane, R. and Romanova, N. (2024) “Pistes pour l'optimisation de modèles de parsing syntaxique.” *LIFT 2 - 2024: Journées de lancement*, Nov 2024, Orléans, France. 8 pp. URL: https://hal.science/hal-04800011v1 (accessed 8 June 2026).

Ziane et al 2023 Ziane, R., Romanova, N. and Lavergne, M. (2023) “En dialogue avec les outils d'apprentissage automatique: une chaîne de traitement pour l'annotation syntaxique.” *Journée d'études CRISCO axe 2, 8 june 2023.* Workshop slides. URL: https://crisco.unicaen.fr/wp-content/uploads/sites/43/2023/07/Journee-CRISCO_8.6.23-Ziane-Romanova-Lavergne.pdf (accessed 8 June 2026).



This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.