Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

NormDef-FR: an annotated corpus of normative definitions for the automatic processing of French legal codes

Aboudourazakou Tetereou

aboudourazakou.tetereou@etu.uae.ac.ma

Abdelmalek Essaadi University

Tarik Boudaa

Abdelmalek Essaadi University

Dadi El Wardani

Abdelmalek Essaadi University

Research Article

Keywords: normative definitions, French legal corpus, information extraction, French law, NormDef-FR, legal NLP

Posted Date: August 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9887034/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

NormDef-FR: an annotated corpus of normative definitions for the automatic processing of French legal codes

Aboudourazakou Tetereou[1*], Tarik Boudaa[2] and Dadi El Wardani[1]

[1*]SOVIA Team, LSA, Abdelmalek Essaadi University, Tetouan, Morocco.

[2]SDIC Team, LSA, Abdelmalek Essaadi University, Tetouan, Morocco.

*Corresponding author(s). E-mail(s): aboudourazakou.tetereou@etu.uae.ac.ma;
Contributing authors: t.boudaa@uae.ac.ma; w.dadi@uae.ac.ma;

Abstract

Normative definitions play a central role in the interpretation and application of the law: they establish the legal meaning of the terms, categories and regimes used in legislative and regulatory texts. However, these definitions are scattered across more than 161,000 articles currently in force, spread throughout the entire body of French legal codes, with no structured resource available to automatically identify or index them. To our knowledge, no annotated reference corpus dedicated to this task is available for French law.

This article introduces *NormDef-FR*, an annotated open-source corpus for the automated study of normative definitions in French legal codes. Version 1.0.0 contains 693 normative definitions of *gold* standard, extracted from 363 articles covering 50 distinct legal codes. Each instance links a defined term, a definition, legal metadata, a definition type and, where the text permits, information on the scope of application. The corpus was constructed using a hybrid pipeline combining detection via linguistic patterns, automatic extraction, manual annotation and systematic deduplication. Inter-annotator agreement achieves a Kappa coefficient (Cohen 1960) of *0.98* for the binary scheme and *0.97* for the three-class scheme, confirming the robustness of the annotation scheme. The corpus is accompanied by a manually verified article-level benchmark, comprising 300 definitional articles and 300 verified true negatives, as well as 449 annotated candidates for automatic validation. The dataset is distributed with comprehensive documentation (`DATA_CARD`, `MODEL_CARD`, bilingual annotation guides) and reproducible scripts. In order to characterise the difficulty of the tasks associated

with the resource, we provide illustrative baselines at three levels: rule-based term-definition extraction, the detection of definitional articles, and the automatic validation of candidates. These experiments show that French normative definitions exhibit a high degree of formal regularity that can be exploited by explicit patterns, that the detection of articles is well resolved against verified negatives, and that the fine-grained validation of candidates constitutes the main bottleneck, with performance stabilising around *0.71 macro-F1* in cross-validation. NormDef-FR thus constitutes a reusable reference resource for French-language legal NLP, designed to support the development of systems for knowledge extraction, legal information retrieval and retrieval-augmented generation.

Keywords: normative definitions, French legal corpus, information extraction, French law, NormDef-FR, legal NLP

1 Introduction

French positive law is based on a codified system in which many legal and technical terms are defined explicitly by legislative and regulatory texts. These normative definitions are not mere glosses: they establish the legally binding meaning of a term within a given scope of application, delineate the rights and obligations arising therefrom, and determine the interpretation of the provisions that refer to them.

A normative definition such as *"on entend par données à caractère personnel toute information se rapportant à une personne physique identifiée ou identifiable"* does not merely describe a concept: it creates a legally operative category whose scope determines the application of an entire legal regime.

Despite this central role, normative definitions remain scattered throughout the French legal codes published on Légifrance. In the snapshot of the LEGI database (DILA - Direction de l'information légale et administrative 2025) used in this work, the articles currently in force constitute a massive corpus in which the definitions are not directly available in a structured form that is searchable and usable by natural language processing systems. A lawyer, legal professional or legal information system seeking the legal definition of a given term must often search through multiple articles and codes, with no guarantee of comprehensiveness or clear distinction between explicit definition, legal characterisation, assimilation and simple cross-reference.

This lack of structuring poses practical problems: it limits the accuracy of legal information retrieval systems and RAG architectures (Lewis et al. 2021), complicates the training of computational terminology models, and makes it difficult to automatically track changes in definitions over time.

Automatic processing of legal language has seen significant developments in recent years, with the publication of corpora and benchmarks for document inference, statutory text retrieval or legal question-answering. Resources such as ContractNLI (Koreeda and Manning 2021), BSARD (Louis and Spanakis 2022) or LLeQA Louis et al. (2024) illustrate this trend.

However, these resources do not specifically target the extraction of normative definitions from French legal codes. Existing work on the extraction of definitions in natural language processing focuses mainly on English or on non-legal domains (Navigli and Velardi 2010), and does not provide an annotated resource dedicated to normative definitions in French law.

This article presents *NormDef-FR*, an annotated open-source corpus for the study and automatic extraction of normative definitions from French legal codes. Version 1.0.0 contains 693 gold definitions extracted from 363 articles covering 50 distinct legal codes. Each instance associates a defined term, a definition, a definition type, legal metadata, information on the extraction pattern and, where the text permits, information on the scope of application. The corpus is accompanied by a manually verified article-level benchmark (300 verified true negatives means not containing any definition term) and 449 annotated candidates for automatic validation. The whole package is distributed with comprehensive documentation and reproducible scripts under open licences.

The construction of NormDef-FR is based on a hybrid pipeline combining linguistic pattern detection, automatic extraction, systematic manual annotation and deduplication. A subsample of 150 examples was independently annotated by a second annotator, yielding a Kappa coefficient (Cohen 1960) of *0.98* for the binary scheme and *0.97* for the three-class scheme. The article also documents a methodological contribution to the construction of benchmarks in Legal NLP: the demonstration that negatives constructed by simple absence in the gold set introduce substantial contamination, and the manual verification procedure enabling the creation of verified true negatives.

In order to characterise the difficulty of the tasks associated with the resource, we provide illustrative baselines at three levels of increasing granularity: rule-based term–definition extraction, the detection of definitional articles on a verified benchmark, and the automatic validation of term–definition candidates. These experiments do not constitute standalone experimental contributions: they serve to document the properties of NormDef-FR, to establish benchmarks for future work, and to show that the main challenge lies not in the coarse detection of definitional articles but in the fine-grained validation of the extracted term–definition pairs.

The rest of the paper is organised as follows. Section 2 presents related work on definition extraction and natural language processing in legal texts. Section 3 describes the construction of the NormDef-FR corpus and the annotation pipeline. Section 4 presents the benchmark tasks and their evaluation protocol. Section 5 reports the baseline results. Section 6 discusses the implications for French-language legal NLP. Section 7 presents the limitations and future directions. Section 8 concludes the article.

2 Related work

2.1 Automatic definition extraction

Automatic definition extraction is a long-standing task in natural language processing, situated at the intersection of information extraction, terminology acquisition and lexical resource construction. Early approaches relied heavily on recurring lexico-syntactic

patterns. The work of Hearst (1992) on the automatic acquisition of hyponymy relations has shown that certain linguistic patterns, such as *NP such as NP*, allow semantic relations to be extracted on a large scale from unstructured text. Although these patterns do not directly target definitions, they have strongly influenced subsequent work on the identification of term–description or term–hypernym relationships. Building on this, several studies have focused more specifically on the extraction of definitions and definitional knowledge. Navigli and Velardi (2010) propose an approach based on generalised patterns and lexico-syntactic structures for the extraction of definitions and hypernyms from textual resources. More recently, the DeftEval task at SemEval-2020 Spala et al. (2020) provided a standardised evaluation framework for the extraction of definitions from free text, distinguishing in particular between the classification of definitional sentences, the sequential tagging of definitional components, and the identification of relationships between terms and definitions (Spala et al. 2020). This work shows that free-text definitions cannot be reduced to a few explicit patterns: they may be discontinuous, implicit, embedded within complex sentences, or distributed across several segments. NormDef-FR follows this line of research, but differs in its domain and the nature of the definitions targeted. The definitions studied here are not encyclopaedic or educational glosses: they are normative definitions, produced by legal texts and having a scope of application.

2.2 Resources and benchmarks in Legal NLP

Automatic processing of legal language has been heavily structured around annotated resources and benchmarks. In English-language law, Koreeda and Manning (2021) introduce ContractNLI, a corpus of annotated contracts for document inference with evidence identification. Chalkidis et al. (2022) propose LexGLUE, a multi-task benchmark for evaluating legal language understanding models in English. More recently, Guha et al. (2023) present LegalBench, a collaborative set of tasks designed to evaluate different forms of legal reasoning in large language models. Related work has also focused on identifying legally significant segments at a finer granularity than the document level. Lippi et al. (2019) propose CLAUDETTE, a system for the automatic detection of potentially unfair terms in terms and conditions. This type of work demonstrates that the extraction of specialised legal units requires expert annotation and a fine distinction between relevant and irrelevant segments. NormDef-FR shares this requirement for granularity, but targets a different unit: not the clause as a whole, but the term–definition pair and its scope of application. These resources have played an important role in standardising evaluation in Legal NLP, but they mainly focus on classification, inference, information retrieval, clause extraction or legal reasoning. They do not directly target the extraction of normative definitions from codified legislative texts. NormDef-FR therefore occupies a complementary position: it focuses on a finer unit of legal knowledge, the term–definition pair, the identification of which is essential for text interpretation, legal research, and question-answering or retrieval-augmented generation systems.

2.3 French-language legal resources and models for legal French

Annotated resources for French-language Legal NLP remain fewer in number than for English, but several recent works have helped to structure this field. Louis and Spanakis (2022) introduce BSARD, a dataset for searching statutory articles in French, constructed from Belgian legal queries annotated by legal experts with the relevant legal provisions. Louis et al. (2024) propose LLeQA, a corpus of long-form legal question-answering in French, pairing expert questions with detailed answers based on legal provisions. These resources demonstrate the growing interest in French-language legal benchmarks based on legislative texts, but they are primarily aimed at information retrieval and question-answering, not the structured extraction of definitions. From a multilingual perspective, Chalkidis et al. (2021) introduce MultiEURLEX, an annotated corpus of European legislation for multi-label classification in 23 languages, including French. This type of resource demonstrates that legal texts can serve as a basis for robust multilingual evaluations, but the task remains at the level of the document or legal topic. At a more lexical and semantic level, the extraction of normative definitions requires finer annotation of the internal structure of articles. In terms of models, Martin et al. (2020) introduce CamemBERT, a Transformer model pretrained on general French, whilst Douka et al. (2021) adapt masked language models to French legal text with JuriBERT. These models constitute important benchmarks for French-language legal tasks. However, their effective use for definition extraction requires the existence of specific annotated resources capable of providing supervision on defined terms, definitional segments and the validity of extracted candidates.

2.4 Positioning of NormDef-FR

NormDef-FR lies at the intersection of these three lines of research. To our knowledge, it is the first annotated resource dedicated to the extraction of normative definitions from French legal codes. Its contribution differs from existing work in three respects: it targets legally operative definitions rather than encyclopaedic ones; it structures each instance around a term–definition pair, enriched with legal metadata and, where the text permits, with scope information; it defines several levels of evaluation of increasing difficulty. These characteristics position it as a complement to existing French-language resources, which focus on article search, question-answering or classification, rather than the structured extraction of normative definitions.

3 Construction of the NormDef-FR corpus

This section describes the construction of NormDef-FR, from the selection of source legal texts to the creation of the gold corpus and associated experimental datasets. The aim of this stage is to produce a reliable resource for the automatic study of normative definitions in French legal codes, whilst retaining the metadata necessary for their legal interpretation.

3.1 Legal source and scope of the corpus

NormDef-FR is constructed from French legal codes published on Légifrance and available in the LEGI database (DILA - Direction de l'information légale et administrative 2025). We focus on articles in force at the time of extraction, in order to build a resource representative of French positive law. Each article is associated with legal metadata documenting its origin and status, including the source code, article number, `LEGIARTI` identifier, dates of entry into and expiry of force, the legal status of the article, and the path to the source file. This decision to work with the codes currently in force serves a dual purpose. On the one hand, it allows us to target definitions currently enforceable in French normative texts. On the other hand, it provides a directly usable basis for applications involving legal information retrieval, terminological analysis or retrieval-enhanced generation. The extracted definitions are therefore not considered as mere linguistic fragments, but as legal units anchored in an article, a code and a period of validity.

3.2 Definition of the annotated object

In this work, we refer to a *normative definition* as a textual relationship whereby a legal text assigns a legally operative meaning to a term, expression or category within a given scope. Such a definition may introduce a legal category, specify the conditions for falling within a regime, establish an assimilation, set a scope of application or delimit the content of a concept used by other provisions. A NormDef-FR instance associates at least one defined term, a definition and legal metadata. The defined term corresponds to the expression whose meaning is established by the text. The definition corresponds to the segment that specifies this meaning or qualification. Where the text permits, the instance may also contain scope information, for example where the definition is introduced by a phrase such as *pour l'application du présent article*, *au sens du présent chapitre* ou *aux fins de la présente section*. This scope is important, as the law often defines terms within a limited scope rather than in general terms. Not all sentences containing an apparent marker are accepted as normative definitions. Certain markers may introduce conditions, legal effects, cross-references or qualifications that do not constitute a usable definition. This distinction explains the central role of manual validation in the construction of the corpus. Instances are also associated with a definition type. In the current version, the main types identified are `explicit`, `explicit_subitem`, `explicit_qualification`, `explicit_internal` and `explicit_enumerative`. This typology allows us to distinguish direct definitions from more contextualised, internal, qualifying or enumerative forms.

3.3 Initial detection of articles and candidates

The construction of NormDef-FR is based on a hybrid pipeline combining automatic extraction and human validation. Initially, we identified articles likely to contain definitions based on recurring linguistic patterns in French legal texts. These patterns include, in particular, constructions such as *on entend par*, *s'entend de*, *est défini comme*, *est considéré comme*, *est réputé*, *constitue*, as well as forms of the type *terme : définition*. We also took into account expressions of scope such as *au sens de*, *pour*

l'application de ou *aux fins de*, which frequently indicate localised definitions. However, these patterns do not have the same level of precision. Explicit forms such as *s'entend de*, *on entend par* or definitions using a colon generally produce cleaner candidates. Conversely, broader patterns such as *constitue*,*est réputé* or *est considéré comme* may introduce qualifications, assimilations, legal effects or non-definitional fragments. This variation justifies the manual validation stage and motivates the experimental task of automatically validating candidates. Each candidate article was associated with score and noise risk information based on the detected patterns, the textual structure and the probability that the segment corresponds to a usable definition. The aim of this automatic stage was not to produce the final corpus directly, but to reduce the search space in order to provide human annotators with articles and candidates enriched with linguistic clues. In a second stage, more stringent extractors were applied to the candidate articles to produce potential term–definition pairs. These extractors primarily exploit the most explicit forms, such as colon-separated definitions, and constructions such as *s'entend de*, *on entend par* and *défini comme*. The resulting candidates were then subjected to manual annotation. The automated stage should therefore be understood as a candidate generation stage; 'gold' status is only assigned after manual validation, correction and normalisation.

3.4 Manual annotation and labelling scheme

Manual annotation aims to determine whether a term–definition candidate constitutes a usable normative definition. Three main labels were used. The label `valid_definition` indicates that the candidate contains a correct definition, with a sufficiently well-defined term and definition. The label `partial` indicates that the candidate contains a genuine definitional relationship, but requires correction, normalisation or cautious interpretation, for example when the term or definition is partially truncated. The label `invalid` indicates that the candidate does not constitute a usable definition, despite the possible presence of a linguistic marker. For automatic candidate validation experiments, the labels `valid_definition` and `partial` are grouped into a binary class `accepted`, whilst the label `invalid` constitutes the class `invalid`. This grouping reflects the practical operation of the pipeline: a `partial` candidate may be useful for enriching the corpus after correction, whereas a `invalid` candidate must be rejected. The annotation is not limited to an overall judgement of the candidate. Where necessary, the fields `term`, `definition`, `definition_type`, `pattern`, `scope_text` and `scope_label` were corrected or standardised. This correction work explains why NormDef-FR should not be regarded as a simple output from an automatic extractor, but as a manually validated gold standard resource.

3.5 Incremental construction and deduplication

NormDef-FR was built incrementally in successive batches. Each batch served to expand the corpus coverage, improve the extraction rules and correct errors observed during previous phases. The initial versions were used to stabilise the annotation schema and identify the most productive patterns. Subsequent versions increased the number of gold instances and improved duplicate checking. When merging the batches,

particular attention was paid to deduplication. Duplicates were identified using a key combining the source code, article number, article identifier and defined term. This strategy prevents the erroneous deletion of two similar definitions found in separate articles, whilst eliminating repetitions arising from successive annotation batches. Following merging, validation and the removal of duplicates, version NormDef-FR 1.0.0 contains 693 gold definitions. The final corpus covers 363 articles from 50 legal codes. This coverage demonstrates that normative definitions are not concentrated in a single area of law, but appear in a variety of codes, notably those relating to public health, the environment, taxation, consumer affairs, construction, trade, transport and energy. The detailed distribution by legal code is provided in the metadata file in the repository.

3.6 Reliability of the annotation

In order to assess the reliability of the annotation scheme, a sub-sample of `150` examples was independently annotated by a second annotator. This sub-sample was selected to include examples falling under the three main labels: `valid_definition`, `partial` and `invalid`.

Agreement was measured at two levels. The first level concerns the detailed three-class scheme: `valid_definition`, `partial` and `invalid`. The second level concerns the binary distinction used in candidate validation experiments, where `valid_definition` and `partial` are grouped into the `accepted` class. The results indicate high inter-annotator agreement, with a Cohen's Kappa coefficient of $\kappa = 0.97$ for the three-class scheme and $\kappa = 0.98$ for the binary scheme. This dual measure is important, as it allows us to assess both the difficulty of making a fine distinction between valid and partial definitions, and the robustness of the operational judgement between acceptable and invalid candidates. The disagreements observed mainly concern partial cases, legal qualifications that are close to definitions, and segments where the exact boundaries of the term or definition are ambiguous. A small number of disagreements also involved technical formula variables: these cases had not been explicitly anticipated in the annotation guidelines, and one annotator ignored them while the second annotator treated them as possible definitions because they were accompanied by explanatory text. After adjudication, we decided not to include such formula-variable explanations as normative definitions in the current release, since they primarily clarify local calculation parameters rather than define autonomous legal notions. These cases are documented as boundary cases and will be evaluated more systematically in future versions of the resource. Overall, these disagreements reflect the very nature of normative definitions, which may be explicit, implicit, contextualised or dependent on a specific scope of application.

3.7 Final corpus statistics

The NormDef-FR 1.0.0 version contains 693 gold standard normative definitions. These definitions are extracted from 363 articles covering 50 legal codes. For each instance, the corpus retains metadata enabling the definition to be linked to its source

text, including the legal code, the article number, the `LEGIARTI` identifier, the dates of force, and the source file.

Table 1 Key statistics for NormDef-FR 1.0.0

Element	Number
Gold definitions	693
Source articles	363
Legal codes covered	50
Term–definition extraction examples	693
Manually verified true negatives	300

The corpus is accompanied by several derived experimental datasets. The first contains the 693 term-definition extraction examples. The second contains 449 annotated candidates for automatic validation. The third is a manually verified article-level benchmark, consisting of 300 articles containing gold definitions and 300 true negative articles.

Table 1 presents the main aggregate statistics for the resource. Detailed statistics, including the complete list of source articles, the distribution by legal source, the distribution of definitional patterns, the distribution of scopes and the distribution of definition types, are provided in the `data/metadata.json` file in the public repository. This approach allows the article to retain a concise presentation whilst making the information necessary for auditing and reusing the resource available.

3.8 Availability, documentation and reproducibility

To facilitate the reuse, auditing and verification of results, NormDef-FR is accompanied by a structured public repository. This repository contains the gold corpus, metadata, derived experimental datasets, the main scripts used for the construction and evaluation of the resource, as well as the documentation required for its interpretation. The `README.md` file provides a concise overview of the resource and the organisation of the repository. The `DATA_CARD.md` file describes the corpus fields, key statistics, labels, distributions, derived experimental datasets, manually verified true negatives, inter-annoter agreement and known limitations. The `MODEL_CARD.md` file documents the experimental baselines, task inputs and outputs, the metrics used, and interpretation caveats. The guides `ANNOTATION_GUIDELINES_EN.md` and `ANNOTATION_GUIDELINES_FR.md` describe the annotation scheme, decision criteria, borderline cases, the labels `valid_definition`, `partial` and `invalid`, as well as the main definitional patterns. The repository distinguishes between scripts related to the construction of the resource and those related to the experiments. The `scripts/annotation/` folder contains the scripts used during the annotation stages, including batch sampling, strict extraction of term-definition candidates, manual review, construction of gold files, schema validation, batch merging, deduplication and metadata generation. The `scripts/experiments/` folder contains the scripts used to prepare the experimental datasets, create splits, construct the manually verified

article-level benchmark, train the TF-IDF and CamemBERT baselines, perform cross-validation and generate the error analysis. This structure allows the article to provide a concise overview of the method and results, whilst providing access in the repository to the files and scripts needed to inspect, reuse and verify the resource. The scripts are provided as reproducibility materials associated with the reported experiments, and the repository documentation specifies their role and the order in which they should be used. The code is distributed under the MIT licence, whilst the data, metadata, derived datasets and documentation are distributed under the Open Licence 2.0 / Etalab.

4 Experimental tasks and evaluation protocol

This section presents the experimental tasks defined based on NormDef-FR 1.0.0, the datasets used, the models evaluated and the metrics selected. The aim is not merely to measure the performance of a single model, but to characterise several levels of difficulty in the automatic extraction of normative definitions. We thus distinguish between term-definition extraction, the detection of definitional articles, and the automatic validation of term-definition candidates.

4.1 Task 1: Term–Definition Extraction

The first task involves extracting, from a text segment, the defined term and its corresponding definition. Each example is taken from a gold instance of NormDef-FR and contains a source text field, the expected term and the expected definition. The system's input is the raw text associated with the definition, whilst the expected output is a pair `term` and `definition`. This task is evaluated on the 693 term–definition extraction examples derived from the gold corpus. The data is divided into training, development and test sets, grouped by `article_key` to prevent the same article from appearing in multiple partitions. The split used contains 509 examples in the training set, 88 in the development set and 96 in the test set. There is no overlap of articles between the partitions. In this first version, we evaluate term-definition extraction using a rule-based baseline. This choice is motivated by the highly formulaic nature of many normative definitions and by the aim of establishing a transparent and interpretable benchmark. We do not train a BIO-style sequential tagging neural model in this version, as NormDef-FR 1.0.0 annotates validated term–definition pairs, but does not yet provide an exhaustive token-by-token annotation of all spans in the raw text. Automatic conversion of gold pairs into BIO labels risks introducing noise when terms are corrected, normalised, partial, or not strictly contiguous in the source text. For this task, we report the exact match of the term, the exact match of the definition, the joint exact match of the term and definition, as well as the token-level F1 score for the term and for the definition. The exact match measures the ability to reproduce the expected string exactly, whilst the F1-token allows for partially correct predictions when textual boundaries do not coincide perfectly.

4.2 Task 2: detection of definitional articles

The second task consists of predicting whether or not a legal article contains a normative definition. The system's input is the full text of an article, and the output is a binary label indicating whether the article is a definitional article or a manually verified true negative. We distinguish between two experimental frameworks. The first framework, used as a preliminary analysis, relies on weak negatives, i.e. articles not represented in the gold set. This framework allowed for an initial assessment of the separability of definitional articles, but a manual audit showed that some weak negatives actually contained definitional or quasi-definitional constructions. For this reason, the main framework selected for the final evaluation is a manually verified benchmark. This manually verified article-level benchmark contains 600 articles, distributed evenly between 300 definitional articles from the gold set and 300 manually verified true negatives. An article is considered a true negative when it contains no internal normative definition, assimilation, legal fiction, definitional qualification or terminological block. The splits are balanced and grouped by `article_key`. The training set contains 420 articles, with 210 positives and 210 negatives. The development set contains 90 articles, with 45 positives and 45 negatives. The test set also contains 90 articles, with 45 positives and 45 negatives. There is no overlap between the article partitions. We evaluate several classification models for this task. The standard baselines rely on TF-IDF representations combined with logistic regression or a linear SVM. We also evaluate a TF-IDF representation based on character n-grams, in order to capture the morphological and phrasing features frequently found in legal texts. Finally, we evaluate CamemBERT (Martin et al. 2020) as a pre-trained neural model for French. For this task, macro-F1 is used as the primary metric, alongside the F1 score for the `definition_article` class, the F1 score for the `verified_true_negative` class, and the confusion matrix. Macro-F1 assigns equal weight to both classes and ensures that the models do not favour only the positive or negative class.

4.3 Task 3: Automatic validation of candidates

The third task involves predicting whether an automatically generated term–definition candidate should be accepted or rejected. This task corresponds to the practical scenario for using the pipeline: an extractor generates candidates from legal articles, and a model must then help filter out usable candidates before they are incorporated into the corpus or reviewed by a human. The candidate validation dataset contains 449 annotated examples. The initial manual labels are `valid_definition`, `partial` and `invalid`. For the binary classification task, the labels `valid_definition` and `partial` are grouped into the `accepted` class, whilst rejected candidates are placed in the `invalid` class. This decision reflects the logic behind the corpus construction: a partial candidate may be useful after correction, whereas an invalid candidate must be excluded. The train/dev/test split contains 322 examples in the training set, 79 in the development set and 48 in the test set. The overall distribution is 306 accepted candidates and 143 invalid candidates. As with the other tasks, the partitions are grouped by `article_key` to prevent information leakage between training and evaluation. We evaluate the same families of models as for article detection: TF-IDF with logistic

regression, TF-IDF with linear SVM, TF-IDF on character n-grams with linear SVM, and Camembert (Martin et al. 2020). This task is evaluated both on the train/dev/test split and via five-fold cross-validation. Cross-validation is used because the test set is relatively small, particularly for the `invalid` class. It provides a more robust estimate of the average performance and its variability. For this task, we report the macro-F1, the F1 for the `accepted` class, the F1 for the `invalid` class, and the confusion matrix. We pay particular attention to false positives of the `invalid` to `accepted` type, as these errors would introduce noise into an automatic corpus enrichment pipeline.

4.4 Preventing cross-partition leakage

For all tasks, splits are constructed by grouping examples by `article_key`. This constraint prevents the same article, or multiple candidates from the same article, from appearing simultaneously in both the training set and the test set. It is particularly important for NormDef-FR, as a single article may contain multiple definitions or multiple candidates generated from similar textual structures. This strategy makes the evaluation more realistic. A model must not only recognise a fragment already seen in another example from the same article, but also generalise to distinct articles. We therefore explicitly verify that there is no overlap of articles between the training, development and test sets.

4.5 Summary of the experimental protocol

Table 2 summarises the experimental tasks, dataset sizes, splits, models evaluated and key metrics. It provides an overview of the experimental protocol prior to the presentation of the results.

Table 2 Summary of the experimental protocol

Task	Examples	Split	Models	Metrics
Term-definition extraction	693	509/88/96	Rules	EM, F1-token
Article detection	600	420/90/90	TF-IDF, CamemBERT	Macro-F1, F1/class
Candidate validation	449	322/79/48	TF-IDF, CamemBERT	Macro-F1, F1/class
5-fold candidate validation	449	5 folds	TF-IDF	Average Macro-F1, standard deviation

The experimental protocol is designed to distinguish between three levels of difficulty. The first level concerns the extraction of definitions from explicit patterns, where rules can be highly effective but where the exact boundaries of terms remain difficult to define. The second level concerns the detection of definitional articles, evaluated within a manually verified framework to avoid contamination from weak negatives. The third level concerns the automatic validation of candidates, which is the task closest to the actual corpus enrichment pipeline and the most sensitive to noise.

5 Experimental results

This section presents the results obtained for the three tasks defined above: term-definition extraction, the detection of definitional articles, and the automatic validation of candidates. The results are presented progressively in order to distinguish tasks strongly guided by explicit patterns from those requiring a more refined validation of the quality of the extracted candidates.

5.1 Rule-based term-definition extraction

The first experiment evaluates the rule-based baseline on the test set for the term-definition extraction task. This baseline utilises the most explicit definitional patterns, notably colon-based definitions, and constructions such as *s'entend de*, *on entend par*, *défini comme* and *est considéré comme*. Table 3 presents the overall results obtained on the 96 examples in the test set.

Table 3 Overall results of the rule-based baseline for term-definition extraction

Metric	Score
Number of examples	96
Exact match of the term	0.635
Exact match of the definition	0.865
Exact match of term + definition	0.604
Term F1-token	0.902
Definition F1-token	0.961
Non-prediction rate	0.010

The results show that definition extraction performs significantly better than exact term extraction. The baseline achieves an F1-token of 0.961 for definitions and an exact match of 0.865, confirming that definitional segments can often be identified by explicit patterns. In contrast, the exact match for the term achieves only 0.635, despite a high F1-token of 0.902. This discrepancy indicates that the rules generally identify the correct text area, but that the exact delimitation of the defined term remains more difficult. Table 4 details the results by major pattern type. The `qualification` pattern has only one example in the test set and is therefore not included in this table.

Table 4 Results of rule extraction based on the definitional pattern

Pattern	n	EM term	EM def.	F1 term	F1 def.
`term_colon_definition`	49	0.959	0.980	0.992	0.988
`s'entend de`	29	0.310	0.966	0.870	0.986
`on entend par`	10	0.500	0.200	0.714	0.826
`défini comme`	7	0.000	0.714	0.805	0.995

Definitions using colons are handled almost perfectly, with an exact match for the term of 0.959 and an exact match for the definition of 0.980. Constructions using *s'entend de* also yield very good results for the definition, but remain more difficult for the term. The patterns *on entend par* and *défini comme* present greater difficulties, particularly because the boundaries of the term can be more complex or because the introductory text can be longer. These results confirm that French normative definitions exhibit a high degree of linguistic regularity, but that the exact segmentation of the defined term constitutes a difficulty distinct from the simple detection of the definitional segment.

5.2 Detection of definitional articles on a verified benchmark

The second experiment focuses on the detection of definitional articles using the manually verified benchmark.

Table 5 presents the results obtained by the TF-IDF and CamemBERT models on the test set. In this table, *FP* denotes a true negative incorrectly predicted as a definitional article, whilst *FN* denotes a definitional article incorrectly predicted as a true negative.

Table 5 Results of definitional article detection on the verified benchmark

Model	Macro-F1	Def.-F1	TrueNeg.-F1	FP	FN
TF-IDF word + LR	0.955	0.953	0.957	0	4
TF-IDF word + SVM	0.956	0.955	0.957	1	3
TF-IDF char + SVM	0.989	0.989	0.989	1	0
CamemBERT	**1.000**	**1.000**	**1.000**	0	0

The results show that the detection of definitional articles is very well resolved when the negatives are manually verified. The best TF-IDF model, based on character n-grams and a linear SVM, achieves a macro-F1 of 0.989. It produces no false negatives in the definitional article class. CamemBERT achieves a macro-F1 score of 1.000 on this verified test set of 90 articles, with no false positives or false negatives. However, this result should be interpreted with caution. It reflects the strong lexical and phrasal distinctiveness of definitional articles in this balanced benchmark, rather than a general resolution of the problem under uncontrolled conditions. It also demonstrates that the audit and reconstruction of negatives were necessary: evaluation on weak negatives would have been less interpretable, as some articles assumed to be negative actually contained definitional or quasi-definitional constructions. CamemBERT's very high performance on this task does not mean that the entire extraction problem is trivial. Rather, it indicates that detection at the clause level is a relatively easy step when the negative examples are reliable. The tasks of exact extraction and candidate validation remain more difficult, as the following experiments demonstrate.

5.3 Automatic validation of candidates

The third experiment assesses the models' ability to predict whether a candidate term–definition pair should be accepted or rejected. This task more closely mirrors the actual operation of the pipeline, as it takes place after the automatic generation of candidates and aims to filter out usable extractions. The test set contains 48 candidates, comprising 36 accepted candidates and 12 invalid candidates. Given this small sample size, the results obtained on this test set should be interpreted in conjunction with the cross-validation presented in the following section. Table 6 presents the results obtained on this test set. The confusion matrix follows the order `[[TN, FP], [FN, TP]]`, where the negative class corresponds to the `invalid` candidates and the positive class to the `accepted` candidates.

Table 6 Results of the automatic validation of candidates on the test set

Model	Macro-F1	Accepted-F1	Invalid-F1	Confusion matrix
TF-IDF word + Logistic Regression	**0.747**	**0.895**	**0.600**	[[6, 6], [2, 34]]
TF-IDF word + Linear SVM	0.686	0.849	0.522	[[6, 6], [5, 31]]
TF-IDF char + Linear SVM	0.625	0.875	0.375	[[3, 9], [1, 35]]
CamemBERT	0.705	0.865	0.545	[[6, 6], [4, 32]]

Unlike article detection, candidate validation is significantly more difficult. The best model on the test set is logistic regression with TF-IDF word, which achieves a macro-F1 of 0.747. The F1 score for the `accepted` class is high, at 0.895, but the F1 score for the `invalid` class remains limited to 0.600. The confusion matrices show that the most problematic errors are false positives of the `invalid` to `accepted` type. For the best model, six invalid candidates are predicted as accepted. These errors are particularly critical in an automatic enrichment pipeline, as they would directly introduce noise into the corpus. Conversely, false negatives reduce recall but do not directly degrade the quality of the selected candidates. CamemBERT does not outperform the best TF-IDF baseline on this task. Despite good results on the development set, its performance on the test set reaches only 0.705 macro-F1. This result suggests that neural fine-tuning is less stable in this context of specialised, moderately sized annotated data. It also confirms that candidate validation does not depend solely on the presence of lexical markers, but on the exact quality of the extracted term–definition pair.

5.4 Cross-validation for candidate validation

Given the small size of the candidate validation test set, we supplemented the evaluation with five-fold cross-validation. This evaluation covers the 449 annotated candidates and respects the grouping by article to avoid cross-fold leakage. Table 7 presents the mean results and standard deviations for the TF-IDF models.

Table 7 Five-fold cross-validation for candidate validation

Model	Macro-F1	Accepted-F1	Invalid-F1
TF-IDF word + Logistic Regression	0.705 ± 0.033	0.849 ± 0.014	0.561 ± 0.054
TF-IDF word + Linear SVM	**0.710 ± 0.019**	0.845 ± 0.015	**0.575 ± 0.030**
TF-IDF char + Linear SVM	0.678 ± 0.038	0.835 ± 0.022	0.522 ± 0.056

Cross-validation confirms that the stabilised performance of the classical models lies around 0.70–0.72 macro-F1. The best average model is the linear SVM with TF-IDF word, which achieves 0.710 ± 0.019 macro-F1. The `accepted` class is learned relatively stably, with an average F1 score of 0.845, whilst the `invalid` class remains more challenging, with an average F1 score of 0.575. These results confirm that candidate validation is the most challenging task in the benchmark. They also show that classical models provide robust baselines in a context of limited annotated data, whilst improving the detection of invalid candidates remains a key challenge for future work.

5.5 Summary of results

The results highlight a clear hierarchy: rule-based extraction is effective for definitions (token F1 score 0.961) but imprecise regarding the boundaries of the term (EM 0.635); article detection performs very well on benchmarks with human validation (macro-F1 1.000 for CamemBERT, 0.989 for TF-IDF char+SVM); candidate validation remains the most difficult task, stabilising at around 0.71 macro-F1 in cross-validation. This hierarchy confirms that the main difficulty does not lie in the rough identification of articles likely to contain definitions, but in the fine-grained evaluation of automatically generated term–definition candidates. NormDef-FR thus enables the documentation of several levels of difficulty in the automatic processing of normative definitions: identifying articles, extracting definitional segments, and then validating the exact quality of the extracted candidates.

6 Discussion

The experimental results highlight a central property of normative definitions in French legal codes : they are both highly formulaic and legally difficult to validate. This tension explains the discrepancy observed between the very high performance achieved in detecting definitional clauses and the more modest performance achieved in the automatic validation of term-definition candidates. NormDef-FR thus allows us to distinguish several levels of difficulty that are often conflated in work on definition extraction: identifying a definitional context, extracting a definitional segment, precisely identifying the defined term, and then determining whether the extracted pair constitutes a usable normative definition.

6.1 Linguistic regularity and the limits of segmentation

A first important result concerns the high linguistic regularity of normative definitions. The rule-based baseline achieves a token F1 score of 0.961 for definition extraction,

which shows that definitional segments are often introduced by explicit and relatively stable patterns.

Forms such as *s'entend de*, *on entend par*, *défini comme* or definitions using a colon constitute highly productive cues. This result confirms that normative definitions, unlike certain encyclopaedic or discursive free-text definitions, are frequently formulated in a controlled manner in legal texts . This regularity enables a robust initial rule-based extraction, which is important for the construction of large-scale resources. However, the results also show that this regularity is not sufficient to fully solve the task. The exact match for the term is only 0.635, whilst the F1-token score for the term is 0.902.

This discrepancy is particularly revealing: the rules often identify the correct textual region, but the exact boundaries of the defined term remain difficult to determine. The difficulty therefore lies less in detecting the definitional signal than in the precise segmentation of the defined legal unit. In law, the defined term may include long expansions, restrictive complements, complex noun phrases or elements dependent on the scope of the article. A boundary error can alter the interpretation of the definition: defining *installation agrivoltaïque*, *taux de couverture d'une installation agrivoltaïque* or a longer expression does not refer to the same legal object. This observation justifies the decision to retain precise term–definition annotations rather than limiting ourselves to the detection of definitional phrases.

6.2 Article detection and construction of negatives

The second major result concerns the detection of definitional articles. On the manually verified benchmark, the task is almost perfectly solved: the best TF-IDF model achieves a macro-F1 of 0.989 and CamemBERT achieves a macro-F1 of 1.000 on the verified test set of 90 articles. These results indicate that articles containing normative definitions exhibit highly distinctive lexical and structural features when compared to true negatives verified manually. The models do not require deep legal reasoning to recognise this definitional presence: surface markers, recurring phrasing and certain typographical structures are more than sufficient in this experimental setting.

These high performance figures must nevertheless be interpreted with caution: they reflect the strong lexical and formulational distinctiveness of definitional articles in this balanced benchmark , and not a general resolution of the detection problem under uncontrolled conditions. This conclusion must be read in the light of the audit of weak negatives. The first version of the article-level task relied on articles not represented in the gold dataset. The audit showed that this strategy was insufficient, as several articles assumed to be negative actually contained definitional, assimilative or quasi-definitional constructions. This point is methodologically significant beyond NormDef-FR. In legal benchmarks, constructing a negative class based solely on absence in a gold dataset can introduce substantial contamination. An unannotated article is not necessarily a negative article: it may simply contain an instance not yet identified or not yet validated. The distinction between *weak negative* and true manually verified negative must therefore be explicitly documented, particularly in legal information extraction tasks where the exhaustiveness of the gold dataset is difficult to guarantee. The reconstruction of a manually verified benchmark thus constitutes a methodological

contribution of this work. It demonstrates that a reliable evaluation of article detection requires verified negatives, i.e. articles for which the absence of an internal normative definition has been verified. The fact that performance remains very high after this correction reinforces the conclusion that article detection is a well-managed step, whilst showing that this conclusion depends on the quality of the negative set construction.

6.3 Validation of candidates and limitations of automation

The third result, and probably the most important for the practical use of NormDef-FR, concerns the automatic validation of candidates. Unlike article detection, this task remains difficult. The best result on the test set reaches 0.747 macro-F1, and cross-validation stabilises performance around 0.71 macro-F1. The gap with article-level results is considerable. It shows that recognising an article containing a definition is not equivalent to validating a term–definition pair. A candidate may contain an explicit definitional marker and still be invalid because the term is poorly segmented, because the definition is truncated, because the segment introduces a condition rather than a definition, or because the extraction captures a legal qualification that is not actionable. An illustrative example is provided by procedural annexes, which account for 83 % of the false positives observed during the audit. A candidate such as *"si vous êtes salarié d'une jeune entreprise innovante : tout document fiscal établissant la qualité de jeune entreprise innovante conformément à l'article 44 sexies 0A du code général des impôts"* contains a colon and an apparently definitional structure, but actually constitutes a list of documents to be provided. This type of error is characteristic of articles that adopt a form-like structure without, however, introducing a normative definition in the strict sense. It illustrates why the validation of candidates cannot be reduced to the detection of a surface marker. Error analysis confirms this difficulty. False positives of the `invalid` to `accepted` type are particularly concerning, as they correspond to invalid candidates that the model would allow into the corpus. From a resource-building perspective, these errors are more serious than false negatives: a false negative reduces coverage, but a false positive directly degrades the quality of the gold. The best TF-IDF model produces six false positives among the twelve invalid candidates in the test set. The `invalid` class achieves an F1 score significantly lower than the `accepted` class, both in the test set and in cross-validation. This asymmetry is consistent with the nature of the pipeline: candidates are produced by extractors already oriented towards definitional patterns, which means that invalid candidates are often superficially similar to true positives. The `partial` label reflects an important practical reality.

In an assisted annotation pipeline, the boundary between a valid definition and a partially extracted definition is not binary. The results show that this grey area is precisely where models fail most, suggesting that the validation task requires not only a linguistic representation but also an understanding of the legal completeness of the extracted pair. By grouping `valid_definition` and `partial` into the class `accepted`, we adopt a resource-oriented perspective: a partial candidate may be retained for correction, whereas an invalid candidate must be excluded. This choice is pragmatic and reflects the logic of assisted annotation upon which NormDef-FR was built.

6.4 TF-IDF versus CamemBERT: limited supervision and specialised domain

The comparative behaviour of TF-IDF and CamemBERT is particularly instructive. CamemBERT achieves perfect performance on article detection, but does not outperform the best TF-IDF baseline on candidate validation. This result is not merely a failure of the neural model: it indicates that the difficulty of fine-grained validation does not lie solely in the general linguistic representation. The task requires recognising segmentation errors, incomplete fragments, implicit scopes, non-definitional qualifiers and subtle legal distinctions. In a context of limited supervision, with only 449 annotated candidates, traditional TF-IDF-based models remain competitive, and are even more robust. This result aligns with a common observation in specialised domains : pre-trained neural models do not automatically replace robust lexical baselines when the task relies on surface-level features and the annotated dataset remains moderate (Mosbach et al. 2021). CamemBERT achieves good performance on the development set but shows a significant drop on the test set, suggesting partial overfitting in this context of limited data. The stability of TF-IDF models, confirmed by cross-validation with moderate standard deviations, reinforces this interpretation.

This comparison has direct practical implications for future systems for extracting normative definitions. A fully neural pipeline is not necessarily the best strategy in an initial phase. The results instead support a hybrid approach: using patterns and rules to generate candidates with good recall, applying classical or neural models to prioritise or filter the candidates, then retaining human validation for ambiguous cases.

In this context, rules are not merely a weak baseline: they constitute a useful, interpretable and auditable component of the system, particularly well-suited to explicit legal formulations.

6.5 Implications for French-language Legal NLP

NormDef-FR provides a benchmark for French-language Legal NLP . Existing resources in legal French focus on article retrieval, question-answering or document classification. Whilst BSARD (Louis and Spanakis 2022) links questions to entire articles and LLeQA (Louis et al. 2024) links questions to passages and long legal answers, NormDef-FR targets a finer-grained unit.

7 Limitations and prospects

In terms of quantity and structure, version 1.0.0 of NormDef-FR comprises 693 reference definitions, a volume that remains modest for training complex neural networks. Tests conducted with CamemBERT also reveal instability in fine-tuning within this context of limited supervision. Furthermore, the current distribution favours explicit patterns, such as colons or the phrase *s'entend de*, at the expense of rarer forms such as legal qualifications, internal definitions, enumerative structures or constructions with complex scope. Finally, the absence of comprehensive BIO annotation at the token level prevents the direct evaluation of sequential labelling models. As automatic conversion risks introducing noise due to normalised boundaries, corrected terms or

partial cases, the production of a token-level aligned corpus is a priority. Methodologically, the creation of reliable negatives in automatic legal language processing presents a specific challenge. Whilst the integration of a manually verified control environment has made it possible to correct the biases of weak negatives, it confirms that the absence of a definition in the reference corpus is not sufficient to classify an article as a negative. Future work would therefore benefit from expanding this volume of control articles, improving the detection of invalid candidates, and characterising non-definitional texts more precisely, whether they are procedural, organisational, punitive or purely transitional in nature. The resource's legal and temporal scope also sets the groundwork for the future. Centred on the codes currently in force on Légifrance, the corpus excludes other major sources such as uncodified texts, case law, circulars, collective agreements or European law. Expanding this framework would allow the transferability of patterns to be tested beyond codified positive law. Furthermore, the synchronic nature of NormDef-FR captures the law at a given moment without reflecting its evolution. Developing a diachronic version of the resource would make it possible to track terminological changes and automatically update legal knowledge bases as legislative reforms take place. These limitations do not diminish the value of NormDef-FR; rather, they outline a stimulating research programme. Whilst the detection of definitional articles now appears to be well-established in the presence of reliable negatives, the fine-grained validation of candidates, the precise segmentation of terms, token-level annotation, the integration of additional legal sources, and the use of the resource in information retrieval or retrieval-augmented generation systems remain important avenues for future work.

8 Conclusion

This article has presented *NormDef-FR*, an open-source annotated corpus for the automatic study of normative definitions in French legal codes. To the best of our knowledge, it constitutes the first reference resource dedicated to this task for French law . Version 1.0.0 contains 693 gold-standard normative definitions extracted from 363 articles covering 50 distinct legal codes.

Each instance associates a defined term, a definition, a type of definition, legal metadata, information on the extraction pattern and, where the text permits, information on the scope of application. The corpus was constructed via a hybrid pipeline combining linguistic pattern detection, automatic extraction , systematic manual annotation, correction and deduplication. The reliability of the annotation is confirmed by an inter-annotator agreement of $\kappa = 0.98$ on the binary scheme and of $\kappa = 0.97$ on the three-class scheme. The dataset is distributed with comprehensive documentation and reproducible scripts under open licences. Beyond the resource itself, this work makes a methodological contribution applicable to the construction of benchmarks in Legal NLP. The demonstration that negatives constituted by mere absence in the gold set introduce substantial contamination, and the manual verification procedure enabling the production of verified true negatives, constitute a lesson applicable to any legal information extraction task where the exhaustiveness of the annotation

is difficult to guarantee. The illustrative baselines provided with the resource characterise three levels of difficulty associated with NormDef-FR. They show that French normative definitions exhibit a high degree of formulational regularity exploitable by explicit patterns, that article detection is well resolved against verified negatives, and that the fine-grained validation of term-definition candidates constitutes

the main bottleneck, with performance stabilising around 0.71 macro-F1 in cross-validation. These benchmarks are made available to enable future work to assess its performance across each of these dimensions. NormDef-FR thus provides a foundation for developing methods for legal knowledge extraction, information retrieval and retrieval-augmented generation from French normative texts.

Priority extensions include token-level annotation, increasing the gold dataset, integrating additional legal sources such as uncodified texts, case law or European law, and developing a diachronic version to track the evolution of definitions over time.

Declarations

Authors' contributions. All authors contributed to the conception and design of the study. Aboudourazakou Tetereou led the corpus construction, data collection, annotation protocol, experiments, and first draft of the manuscript. Tarik Boudaa and Dadi El Wardani contributed to the methodological design, supervision, analysis of results, and critical revision of the manuscript. All authors read and approved the final manuscript.

Funding. The authors received no funding for this study.

Data availability. The NormDef-FR corpus, derived experimental datasets, annotation guidelines, documentation, and reproducibility scripts are publicly available at: https://github.com/testayantune/normdef. The data are distributed under Licence Ouverte 2.0 / Etalab, and the code is distributed under the MIT License.

Competing interests. The authors declare no competing interests.

References

Chalkidis I, Fergadiotis M, Androutsopoulos I (2021) MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In: Moens MF, Huang X, Specia L, et al (eds) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 6974–6996, https://doi.org/10.18653/v1/2021.emnlp-main.559, URL https://aclanthology.org/2021.emnlp-main.559/

Chalkidis I, Jana A, Hartung D, et al (2022) LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 4310–4330, https://doi.org/10.18653/v1/2022.acl-long.297, URL https://aclanthology.org/2022.acl-long.297/

Cohen J (1960) A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20(1):37–46. https://doi.org/10.1177/001316446002000104, URL https://journals.sagepub.com/doi/10.1177/001316446002000104

DILA - Direction de l'information légale et administrative (2025) LEGI - French Codes, Laws and Consolidated Regulations. URL https://www.data.gouv.fr/fr/datasets/legi-codes-lois-et-reglements-consolides/, prepared and published by @VinceGx33 on Hugging Face

Douka S, Abdine H, Vazirgiannis M, et al (2021) JuriBERT: A Masked-Language Model Adaptation for French Legal Text. In: Aletras N, Androutsopoulos I, Barrett L, et al (eds) Proceedings of the Natural Legal Language Processing Workshop 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp 95–101, https://doi.org/10.18653/v1/2021.nllp-1.9, URL https://aclanthology.org/2021.nllp-1.9/

Guha N, Nyarko J, Ho DE, et al (2023) LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. https://doi.org/10.48550/arXiv.2308.11462, URL http://arxiv.org/abs/2308.11462, arXiv:2308.11462 [cs.CL]

Hearst MA (1992) Automatic Acquisition of Hyponyms from Large Text Corpora. In: COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics, URL https://aclanthology.org/C92-2082/

Koreeda Y, Manning C (2021) ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. In: Findings of ACL: EMNLP 2021. Association for Computational Linguistics, pp 1907–1919, https://doi.org/10.18653/v1/2021.findings-emnlp.164

Lewis P, Perez E, Piktus A, et al (2021) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. https://doi.org/10.48550/arXiv.2005.11401, URL http://arxiv.org/abs/2005.11401, arXiv:2005.11401 [cs.CL]. Accepted at NeurIPS 2020

Lippi M, Palka P, Contissa G, et al (2019) CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law 27:117–139. https://doi.org/10.1007/s10506-019-09243-2

Louis A, Spanakis G (2022) A Statutory Article Retrieval Dataset in French. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 6789–6803, https://doi.org/10.18653/v1/2022.acl-long.468, URL https://aclanthology.org/2022.acl-long.468/

Louis A, Van Dijck G, Spanakis G (2024) Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. In: Proceedings

of the 38th AAAI Conference on Artificial Intelligence. AAAI Press, Vancouver, Canada, URL https://arxiv.org/abs/2309.17050

Martin L, Muller B, Ortiz Suárez PJ, et al (2020) CamemBERT: a Tasty French Language Model. In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 7203–7219, https://doi.org/10.18653/v1/2020.acl-main.645, URL https://aclanthology.org/2020.acl-main.645/

Mosbach M, Andriushchenko M, Klakow D (2021) On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. https://doi.org/10.48550/arXiv.2006.04884, URL http://arxiv.org/abs/2006.04884, arXiv:2006.04884 [cs.LG]. ICLR 2021

Navigli R, Velardi P (2010) Learning Word-Class Lattices for Definition and Hypernym Extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp 1318–1327

Spala S, Miller N, Dernoncourt F, et al (2020) SemEval-2020 Task 6: Definition Extraction from Free Text with the DEFT Corpus. In: Herbelot A, Zhu X, Palmer A, et al (eds) Proceedings of the Fourteenth Workshop on Semantic Evaluation. International Committee for Computational Linguistics, Barcelona (online), pp 336–345, https://doi.org/10.18653/v1/2020.semeval-1.41, URL https://aclanthology.org/2020.semeval-1.41/