Legal Spanish Terminology Extraction: Gold-standard Generation and LLMs Evaluation

Lucía Palacios[1], Beatriz Guerrero García[2], Patricia Martín-Chozas[1], Elena Montiel-Ponsoda[1],
[1]Technical University of Madrid, Spain
[2]University of Salamanca, Spain
{lucia.palacios, patricia.martin, elena.montiel}@upm.es
bguer98@usal.es

Abstract

This study aims to develop a gold-standard for terminological extraction in Castilian Spanish within the domain of labour law. To achieve this, a methodology was developed based on established linguistic theories and reviewed by a team of expert terminologists. Departing from previous extraction studies and reference theoretical frameworks, candidate terms were identified by their morphosyntactic patterns, enriched by assessing their degree of specialisation in reference resources. The candidate terms were then subjected to manual validation. To evaluate its applicability, we assessed the performance of the LLaMA3-8B and Mistral-7B language models in extracting labour law terms from the latest version of the *Real Decreto Legislativo 2/2015 Ley del Estatuto de los Trabajadores*. YAKE was also included as a statistical baseline for comparison between traditional methods and generative approaches. All models were evaluated against the validated gold-standard.

1 Introduction: ATE, Legal Spanish Terminology & Gold-Standard Generation

Terminological units provide linguistic access to specialised concepts within specific contexts (Cabré, 2009), making them fundamental to languages for specific purposes (LSP). Their relevance has attracted interdisciplinary attention—from linguistics and cognitive sciences to specialised fields and computer science (Faber and L'Homme, 2022)—where term extraction is both the starting point and the cornerstone, as it is the (semi)automatic process of identifying candidate terms from a given speciality domain based on linguistic and/or statistical studies of textual corpora (Peña and Peña, 2015). The development of natural language processing systems has made it possible to process larger amounts of text and obtain refined extractions that contributed to the creation of linguistic, specialised lexicographic and ontological resources that feed machine translation engines and chatbots, as well as facilitating data mining and knowledge discovery tasks (Tran et al., 2023).

To this matter, field-specific terminology is a necessity, despite certain challenges such as accessibility issues and linguistic particularities (Bagot, 1999). The legal domain is particularly challenging due to the characteristics of legal jargon, i.e. the variety of language used in texts related to the application and practice of law, including judicial, administrative, and notarial documents (Española and del Poder Judicial, 2017). Within these texts, labour terminology is focused specifically on terms used in legal contexts related to the labour field that allow specialists to communicate with precision and without ambiguity, in order to achieve the terminological ideal of univocity (Española and del Poder Judicial, 2017). It is characterised by a significant presence of derived forms—nouns, adjectives, and adverbs—along with abbreviations and acronyms referring to organisations and legal provisions. It also features prepositional phraseological units and frequent nominalisations of verbs, which sometimes also function as specialised nouns. These nominalisations are not only integral to specialised phraseology, but also play a key role in the transmission of specialised knowledge. Furthermore, the coexistence of specialised terms with general discursive units gives rise to overlaps between general and domain-specific meanings (Hourani-Martín, 2023; Vitalaru, 2019; Española and del Poder Judicial, 2017; Hormigo, 2011; Cabré et al., 1996).

Previous studies have highlighted the complexity involved in the extraction of legal terminology in Castilian Spanish (Martín-Chozas et al., 2022; Rico et al., 2019; Martín-Chozas and Calleja, 2018), which partially derives from the linguistic

characteristics of terms. Their length and internal morphosyntactic complexity generate noise in automatic 'raw' extractions such as phraseology—units of specialised meaning that do not represent specialised concepts (Bagot, 1999; Cabré et al., 1996)—, and incomplete terms or concatenations of two separate terms that do not form a single terminological unit, among other problems. This shows that while labour law terminology fits into the generalised conventions of legal language, it may have certain distinctive features that need to be identified in order to nuance Automatic Term Extraction (ATE) task. These are particularly complex due to the lack of standardised, normalised and authoritative resources that could support the development of training and evaluation datasets (Breton et al., 2025). This gap is particularly noticeable when compared with other domains with numerous monolingual and multilingual resources, as outlined in the survey of Tran et al. (2023)—like biomedicine, where reference data bases such as SNOMED CT exist (Gaudet-Blavignac et al., 2021)—while very few exist for the legal field in Spanish.

For this reason, the TeresIA project aims to develop tools that allow the integration and linking of existing resources to create a meeting point for terminologies in Spanish. Likewise, it aims to enrich existing resources by means of methodologies based on AI techniques. Therefore, in this line of work, the aim of the present paper was to develop a gold-standard resource focused on Castilian Spanish within the specific subdomain of labour law, as manually curated gold-standards remain the most common and consistent evaluation method for this specific task (Tran et al., 2023). For its construction, the proposed methodology for the development of reliable evaluation datasets for the ATE task is based on well-established linguistic theories and aligned with best practices in terminology work.

To demonstrate the practical value of the proposed resource, it was used to evaluate the term extraction performance of two state-of-the-art open-source large language models, LLaMA3-8B and Mistral-7B, on the *Real Decreto Legislativo 2/2015, Ley del Estatuto de los Trabajadores* (Statute of Workers Rights in the Spanish Jurisdiction), the fundamental act that legally regulates and protects labour relations in Spain, and provides the basis for the rest of the labour legislation. Using prompting techniques, zero- and one-shot, adapted for term identification and compared the output of each model against our gold-standard. Additionally, we included YAKE as a statistical baseline for comparison. We compared the performance of the LLMs with each other and also with a classical statistical extraction method.

Building on this work, our paper makes the following key contributions:

1. Domain-adaptable methodology for ATE gold-standard construction
2. Reusable gold-standard dataset
3. Assesment of YAKE, LLaMA3-8B, Mistral-7B language models in extracting Spanish labour law terminologies

2 Previous Work and Theoretical Background

2.1 Terminology Extraction: Theoretical and NLP Techniques

ATE tasks are an example of the interdisciplinary appeal of terminology. From a linguistic perspective, they facilitate the theoretical linguistic work carried out by terminologists (Bagot, 1999) and are the basis for creating specialised-linguistic resources, such as dictionaries or databases that make validated terminologies available to general language users, students, or language professionals in institutions or public services, such as translators and interpreters (Rodríguez-Tapia, 2024; Seghiri, 2017; Sierra, 2010). However, the development of these linguistic materials requires a computational approach.

Terminology extraction from natural language texts as an NLP task has been the focus of automation efforts since the earliest approaches (Maria Teresa et al., 2006), which were based on linguistic (Bourigault, 1992; Evans and Zhai, 1996), statistical (Frantzi et al., 2000; Nazar, 2011), and hybrid perspectives (Fkih and Omri, 2012; Lossio-Ventura et al., 2016). All of these methods rely heavily on external resources and are typically implemented as supervised systems, whose performance largely depends on the quality of the underlying corpora (Tran et al., 2023). More recent techniques incorporate machine learning (Ljubešić et al., 2019; Hossari et al., 2019), deep learning (Rokas et al., 2020; Sugimoto et al., 2021), graph-based methods (Ala and Sharma, 2020; Kimura

et al., 2021), and hybrid strategies (Zheng et al., 2023). While these approaches have led to more productive outcomes, the need for annotated data remains.

Latest developments in LLMs have opened promising avenues for generating high-quality terminology extraction systems. These models are pre-trained on huge and diverse amounts of data, which leads to one of the key advantages of LLMs, that they do not require task-specific annotated data to perform accurately, as have reported several studies across various domains (Goel et al., 2023; Brkić Bakarić and Lalli Paćelat, 2024; Breton et al., 2025).

2.2 Evaluation for the ATE Task

Evaluation methods for terminology extraction tasks are generally classified, according to the survey by Tran et al. (2023), into manual and automatic approaches. Manual evaluation requires domain experts to determine how effectively the system extracted the candidate terms (Justeson and Katz, 1995). This approach is highly costly in terms of human effort. In contrast, automatic evaluation methods compare the extraction results with predefined term sets, such as terminological dictionaries (Macken, Lieve and Lefever, Els and Hoste, Veronique, 2013; Kayadelen et al., 2020) or gold-standard datasets (Tran et al., 2023).

The gold standard-based method uses terminological lists manually compiled by domain experts as a validation reference (Loginova et al., 2012). These datasets often include a wider variety of term variants while still preserving the terminological nature of conceptual units within the specialised domain. By reflecting the complexity of domain-specific terminology, the gold-standard approach is regarded as a more robust and suitable option for evaluating terminological quality, as it combines automatic tools with expert domain knowledge.

Despite significant progress in terminology extraction technologies, the lack of unified and effective validation standards remains a major challenge. Current evaluation methods are often based on specific datasets or metrics tailored to particular languages, most commonly English, or to domains traditionally studied within the ATE task (e.g. the widely used resources such as the ACL RD-TEC 2.0 dataset (QasemiZadeh and Schumann, 2016) and TermEval (Rigouts Terryn et al., 2020)).

For these reasons, a methodology specifically designed to address the needs of the labour domain in Spanish is proposed. The approach aims to address linguistic and conceptual aspects of legal language while being flexible enough to be applied across different subdomains.

3 Methodology for Gold-standard Generation and Validation

The methodology used to create the gold-standard for labour law terminology extraction in Spanish is based on morphosyntactic patterns of the terms. Previous studies show the importance of considering the morphosyntactic structure of terms to refine tATE tasks and that within a thematic domain it is possible to track terminology following said patterns, as they allow to determine which lexematic combinations correspond to linguistic concept representations, that is, terms, and which do not (Cabré, 2009). Therefore, the automatic extraction was directed towards candidates that met this set of conditions, i.e. that matched the morphosyntactic patterns and that their items were specialised where required, in order to obtain candidates specific to the labour domain and avoid noise or non-terminological candidates.

3.1 Selection of Morphosyntactic Patterns

The initial list of morphosyntactic patterns was compiled building on previous studies that identified recurrent morphosyntactic patterns of terms across various domains—including the legal field in Spanish (Martín-Chozas et al., 2022; Rico et al., 2019; Martín-Chozas and Calleja, 2018; Bagot, 1999). This first stage focused on noun-based structures, as they were the most frequent, and a first candidate patterns list was drafted. Subsequently, a filtering process was applied to identify which patterns were representative of Spanish labour law terminology. Using the *Estatuto de los Trabajadores* and CQL queries in Sketch Engine, the selected patterns were tested to determine if they showed term candidates of the labour domain that followed the linguistic definition of terminology as language units with specialised meaning in specific pragmatic contexts (Cabré, 2009). After manually reviewing the contexts of the candidates, the patterns that did not yield terminological units were excluded. Once manually filtered, the same experiment was replicated with ChatGPT to determine the reliability of the manually-verified patterns. The results of the automatic test were the same as the

Morphosyntactic Patterns	
1	N (specialised) + Adj (specialised)
2	N (general / not specialised) + Adj (specialised)
3	N (specialised) + Adj (specialised) + Adj (specialised)
4	N (general / not specialised) + Adj (specialised) + Adj (specialised)
5	N (specialised) + Adj (specialised) + Prep + Art (opcional) + N (specialised)
6	N (specialised) + Adj (specialised) + Prep + Art (opcional) + N (specialised) + Adj (specialised)
7	N (specialised) + Prep + Art (opcional) + N (specialised)
8	N (specialised) + Prep + Art (opcional) + N (specialised) + Adj (specialised)
9	N (specialised) + Prep + Art (opcional) + N (general / not specialised) + Adj (specialised)
10	N (specialised) + Prep + Art (opcional) + N (specialised) + Prep + Art (opcional) + N (specialised)
12	N (specialised) + Prep + Art (opcional) + N (specialised) + Prep + Art (opcional) + N (specialised) + Adj (specialised)

Table 1: Morphosyntactic Patterns for Terminological Extraction.

manual one, which confirmed the reliability of the patterns and led to the final selection used for TE, as shown in the table 1 Morphosyntactic Patterns for Terminological Extraction.

In order to detect the candidate terms, two filters were established to ensure the reliability of the results. Firstly, a frequency filter that identified language strings corresponding to the patterns with a minimum frequency of 10 occurrences in the corpus. Secondly, a filter that would check the specificity of the pattern-components where required—as seen in table 1 Morphosyntactic Patterns for Terminological Extraction—to ensure that the candidates were actual terms from the labour domain and not a lexical unit with no specific meaning. For candidates corresponding to the pattern "N" (noun), linked resources were used to identify those units present in the *Estatuto* that had an exact match with entries in the *Vocabulari de Dret* (Law Glossary) by Termcat, the reference lexicographical law-resource in Catalan, elaborated by Law experts, providing terminological equivalents in Spanish. For more complex morphosyntactic patterns, its detection was combined with the presence of its components in the same linked resource, both when there was an exact and a partial match. Particularly in the case of adjectives, which had to be specialised for the pattern to be fulfilled, an extraction of adjectives with an independent entry in the *Vocabulari* was carried out, as well as those part of a nominal entry, in order to obtain a list of specific adjectives. This step was necessary to ensure the specialised nature of the units.

As a result of the extraction, a list of candidate terms was obtained for each pattern, together with their respective contexts within the *Estatuto* and tagged indicating which part of the pattern was found in the linked resources.

3.2 ATE Validation

The extraction results were manually validated to determine if they met the term definition provided earlier (Cabré et al., 1996). Firstly, the terms with complete coincidence with the *Vocabulari de Dret* were validated. In the cases where no match nor a partial match was found, it was assessed whether the candidates followed the morphosyntactic pattern regarding the specialised nature of their components. Initially, a total of 3,803 candidate terms that followed the morphosyntactic patterns were identified. For the rest of the candidates that had not passed the first filter, it was checked whether they appeared in their entirety in the reference legislation in Spain, *BOE*, or in the *Diccionario panhispánico del español jurídico* (Pan-Hispanic Dictionary of Legal Spanish), and only considering operational terms those for which there was agreement from validators in the manual validation.

Morphosyntactic Patterns	
1	N (specialised) + Adj (specialised)
2	N (general / not specialised) + Adj (specialised)
3	N (specialised) + Adj (specialised) + Adj (specialised)
4	N (general / not specialised) + Adj (specialised) + Adj (specialised)
5	N (specialised) + Adj (specialised) + Prep + Art (opcional) + N (specialised)
6	N (specialised) + Adj (specialised) + Prep + Art (opcional) + N (specialised) + Adj (specialised)
7	N (specialised) + Prep + Art (opcional) + N (specialised)

Table 2: Final Morphosyntactic Pattern List.

3.3 Final Pattern List

As a result of manual extraction, the need to nuance the patterns was verified, to take into account specific adjectives and to avoid the detection of phraseological units as if they were terminological. The latter is derived from the excessive nominalisation in the legal field abovementioned, where much of the phraseology is formed with deverbal nouns which can add clutter to the search for patterns starting from the detection of nouns—e.g. for the pattern "N + prep + N" we found the term *estatuto de los trabajadores* (statute of workers rights), but also the phraseological unit *representación de los trabajadores* (representation of employees). At this point, the specificity of the components of the patterns becomes particularly important, as it is the key element in determining whether the candidates were actual terms of the labour domain—e.g. for the pattern "N + Adj + prep + N" the term *trabajador demandante de empleo* (job-seeker) was found, but also the discursive unit *trabajador considerado en su totalidad* (worker taken as a whole) with no specialised meaning. For this reason, entire patterns had no valid term candidates, but only phraseological or discursive units. These patterns were excluded, which resulted in the final list of operative patterns.

Inter-annotator agreement metrics for manual validation resulted in a Cohen's Kappa of 0.7230 and an overall agreement rate of 84.57%, which indicates substantial agreement among validators. The result of this process was the identification of 752 terms, corresponding to the different patterns.

4 Evaluation of LLMs with Gold-Standard Dataset

For assessing the evaluation capabilities of our gold-standard for terminology extraction within the Spanish labour law domain, we conducted two experiments employing LLMs for extracting terms. These experiments are based on prompt engineering techniques, zero- and one-shot. Through these experiments, we aim to analyze the performance of the LLMs in identifying relevant legal terminology, measuring how well their outputs align with the manually curated gold-standard.

In addition to these LLM-based approaches, we also employed the statistical keyword extraction model YAKE (Campos et al., 2020). We applied YAKE directly with the PyPI library [1], to the *Estatuto de los Trabajadores*, using the resulting terms to evaluate their overlap with the gold standard. The objective of including YAKE was to provide a baseline for comparison for the performance of LLMs. This allowed us to compute changes in performance ($\Delta P/\Delta R$) between the LLM-based methods and the statistical baseline, showing the relative improvements in precision and recall from the LLM-based extraction.

All the results and code can be found on the following repository:
https://github.com/luxxiferr/gold_standard_llms_terminolgoy

To carry out experiments, we selected two state-of-the-art LLMs: LLaMA 3 with 8 billion parameters and Mistral with 7 billion. Since the advent of LLMs, proprietary models have become the industry benchmark for a wide range of NLP tasks, including information extraction (Breton et al., 2025). Despite their strong performance, these models often present limitations such as access restrictions, high computational resource requirements, and cost barriers.

Additionally, we selected LLaMA 3 and Mistral due to their relatively small model sizes compared to some of the largest LLM variants. Their sizes allow us to run inferences efficiently with-

[1] https://pypi.org/project/yake/

out requiring powerful GPUs or extensive cloud infrastructure. These advantages were key in our decision to rely on LLaMA 3 and Mistral for these experiments.

4.1 Prompting Techniques

In our experiments, we crafted textual prompts that reflect the criteria used to create our gold-standard dataset, with the goal of guiding the selected models in identifying candidate legal terms. We apply two main prompting strategies: zero-shot and one-shot.

Our zero-shot prompt explicitly defines the task and the morphosyntactic patterns that candidate terms must follow. The prompt also specifies the domain of interest (labour law terminology in Spanish) and requests a clean list of candidate terms extracted from a given legal text fragment. This approach tests the model's ability to generalize the task from the instructions alone, reflecting the linguistic patterns we applied during manual term identification. The zero-shot prompt can be seen in Table 3 (it has been translated from Spanish to English). The one-shot prompt is exactly the same as the zero-shot, but we included a positive extraction example after the instructions. The example can be seen in Table 4.

By basing both prompts on the patterns and criteria used in manual term identification, we ensure that the models have an accurate context for extracting candidate terms. This facilitates a meaningful comparison between the model outputs and human-validated terminology, enabling a robust evaluation of the LLMs' ATE performance.

4.2 Evaluation Methodology

After performing inference with our selected LLMs and applying post-processing to their outputs (e.g., deleting repeated prompt fragments, removing explanatory or justificatory text, and isolating only candidate term expressions), we proceeded to evaluate the quality of the extracted candidate terms by comparing them against our manually curated gold-standard dataset. Our evaluation framework is based on the one proposed by Breton et al. (2025) for ATE using LLMs in the legal domain.

We began by employing classic information extraction metrics: Precision, Recall, and F1 Score. In addition, we consider the F2 Score, which weights Recall more heavily, recognizing that, in terminology extraction, missing relevant terms could be worse than including some false positives. The absence of relevant terms can limit the coverage and overall usefulness of the generated resources, whereas false positives can often be filtered out or corrected in later stages. However, these metrics don't take into account slight mismatches between expert annotations and model predictions. For example, an expert might annotate the term as a single unit, *labour legislation*, while a LLM might extract *the current labour legislation*. Under conventional strict matching, the model's partial extraction would be counted as a False Positive, despite correctly identifying the core concept. To overcome this, we followed the mentioned framework by Breton et al. (2025), that implements the division of the True Positive category into two subcategories. These subcategories are: Perfect Match (the term exactly matches the gold-standard annotation) and Partial Match (it is a valid subset of a term). To quantify partial matches, the normalised Levenshtein distance is introduced, which is a metric that measures the minimal number of character edits needed to transform one string into another. Then, it is normalised to allow comparison across terms of different lengths.

4.3 Results and Discussion of Perfomance

Each set of extracted terms was evaluated for all 40 articles of the *Estatuto de los Trabajadores*. The models' outputs were compared against our gold-standard annotations using the evaluation framework described. Table 5 presents the terminology extraction results for each model.

Overall, one-shot prompting increased the number of exact true positives for both models, improving precision, while partial true positives remained stable or slightly decreased. Mistral produced fewer false positives than LLaMA 3 under one-shot conditions, reflecting higher precision, and both models reduced false negatives, thereby improving recall, especially LLaMA 3.

A key part of the evaluation was the normalised Levenshtein distance, measuring similarity between extracted terms and the gold-standard validated terms. The distances range from 0.647 to 0.708. This indicates that many partial matches are close to the gold-standard terms but still differ notably. Despite the improvements seen with one-shot prompting, the results indicate that neither model consistently extracts legal terminology accurately enough. The high number of false positives and false negatives, combined with moderate

You are an expert in Spanish labour law terminology.
Task: Extract only the candidate terms that appear in the following excerpt from a legal article of the Workers' Statute.
The terms must be:

- Nouns or complex nominal structures that represent concepts specific to labour law.
- Terminological units within the legal-labour domain.

They must follow one of the following morphosyntactic patterns:
Return a clean list of candidate terms:

- One term per line
- No numbering, no quotation marks, no symbols

Excerpt text:

```
"""{texto_articulo}"""
```

Candidate terms:

Table 3: Zero-shot prompt used before extracting candidate terms in the legal-labour domain.

precision and recall values, reflects the limitations in these models' capabilities for this task.

We calculated the differences in precision (ΔP) and recall (ΔR) to quantify how much better or worse each LLM performs compared to the YAKE baseline. In the zero-shot setting, Mistral achieved a precision of 0.677, much higher than YAKE (with $\Delta P = +0.2099$) but at a significant cost in recall ($\Delta R = -0.5119$). LLaMA3-8B, with zero-shot prompting, showed a precision gain of $\Delta P = +0.2180$ and a recall loss ($\Delta R = -0.5488$), with a comparable $\Delta P/\Delta R$ of -0.3972. These results show that LLMs are more selective but fail to retrieve a substantial number of relevant terms when extracting terms with zero-shot prompts.

In the one-shot setting, where models received an example for guidance, both LLMs improved in recall but at the cost of precision. Mistral's recall is 0.487 (narrowing the gap to $\Delta R = -0.4399$), while precision dropped to 0.450 ($\Delta P = +0.1918$), resulting in a $\Delta P/\Delta R$ of -0.4360. LLaMA3-8B increased recall to 0.545 ($\Delta R = -0.3825$) but only achieved a modest precision gain ($\Delta P = +0.0694$), producing the lowest ratio in the group ($\Delta P/\Delta R = -0.1815$). Considering overall performance, Mistral zero-shot achieved the highest F_1 score (0.440), while YAKE obtained the highest F_2 score (0.610), reflecting its strong recall. Additionally, YAKE exhibited the lowest average normalised Levenshtein distance (0.5506), outperforming all LLM configurations in partial match quality because the terms are all extracted as they are from the text. Overall, YAKE offers broader extraction because it relies on statistical features such as term frequency and co-occurrence, allowing it to capture a wide range of candidate terms. This results in high recall, but with very low precision because of the large number of false positives. LLMs are more selective in their outputs. Zero-shot prompts make LLMs more conservative and one-shot prompts produce more generalised results, increasing recall.

5 Conclusions & future work

This study presents several key contributions: a domain-adaptable methodology for constructing gold-standard datasets for ATE, the creation of a reusable gold-standard dataset specifically for the Spanish labour law domain, and an assessment of LLaMA3-8B and Mistral-7B language models in extracting domain-specific terminologies, complemented by YAKE, which served as a statistical baseline to contextualize LLM performance. Thanks to its methodological and theoretical basis, the present gold-standard can not only be used in linguistic or computational tasks within this domain, but can also serve as a reference for the development of future resources in other areas, such as mercantile, administrative, or criminal law. To do so, it would be necessary to test the morphosyntactic patterns in order to identify the patterns specific to those domains. Moreover, after the experiments carried out, the reliability of the present gold-standard has been tested against LLMs. Therefore, at this stage, we consider the following courses of action for further research. First, to validate the *Estatuto* in its integrity in order to have a wider body of terminology, applying the same methodology fol-

Excerpt text: Article 2. Special employment relationships.
1. The following shall be considered special employment relationships:
a) That of senior management personnel not included in Article 1.3.c). b) That of household service.c) That of convicted persons in penitentiary institutions. d) That of professional athletes. e) That of artists engaged in performing, audiovisual, and musical arts, as well as persons performing technical or auxiliary activities necessary for such work. f) That of persons involved in commercial operations on behalf of one or more employers without bearing the risks and rewards thereof. g) That of workers with disabilities providing services in special employment centers. h) (Repealed) i) That of minors subject to internment measures for the fulfillment of their criminal responsibility. j) That of medical residents in specialist training programs. k) That of lawyers providing services in law firms, whether individually or collectively. l) Any other work expressly declared by law to be a special employment relationship.
2. In all the cases mentioned above, the regulation of these employment relationships shall respect the fundamental rights recognised by the Constitution.
Candidate terms:

- Noun: activity, article, account, compliance, development, execution, employment, internment, law, regulation, residency, risk, service, work
- Noun + adjective: technical activity, professional athlete, fundamental right, penitentiary institution, commercial operation, criminal responsibility, employment relationship
- Noun + adjective + adjective: [none identified]
- Noun + adjective + preposition + (optional article) + noun (+ optional adjective): special employment center
- Noun + preposition + (optional article) + noun: [none identified]

Table 4: Example used in the one-shot prompt. Spanish text tranlasted into English.

Model	TPe	TPp	FP	FN	Prec	Rec	F1	F2	Avg Lev
LLaMA 3 (0-shot)	389	350	814	1215	0.476	0.378	0.421	0.384	0.708
LLaMA 3 (1-shot)	740	409	2362	961	0.327	0.545	0.409	0.481	0.676
Mistral (0-shot)	475	351	940	1164	0.677	0.415	0.440	0.425	0.695
Mistral (1-shot)	730	310	1273	1095	0.450	0.487	0.468	0.479	0.647
YAKE	1484	561	5886	161	0.258	0.927	0.403	0.610	0.551

Table 5: Terminology extraction and evaluation results. TPe: Exact true positives, TPp: Partial true positives, FP: False positives, FN: False negatives, Prec: Precision, Rec: Recall, F1: F1-score, F2: F2-score, Avg Lev: Average normalised Levenshtein distance.

lowed in the present paper. Secondly, to study other morphosyntactic patterns whose central elements are not nouns in order to enrich existing terminological resources, which currently show limitations in accounting for certain elements—e.g. specific adjectives. Finally, to analyse phraseology with the excluded patterns that only showed discursive or phraseological units, which could be used to create specialised-phraseology detection filters.

Future work involving LLMs will focus on fine-tuning LLaMA3-8B, Mistral-7B, and other models on the gold-standard resource. These models will be able to recognize terminology more accurately, generate more precise results, and gain a deeper understanding of the legal domain terminology by including the validated terms into the training data. Moreover, we plan to implement explainability strategies and carry out error studies on model outputs. Thus, we can analyze the behavior of LLMs and find frequent mistakes or ambiguities in terminology extraction. This approach could ensure a feedback loop that continuously improves the accuracy and reliability of terminology recognition in future experiments.

Nevertheless, we hope that this gold-standard will contribute to the improvement and enrichment of existing terminological resources in Spanish.

Acknowledgments

This research has been conducted within the framework of the TeresIA project, funded by the European Union's Next GenerationEU / PRTR funds through the Spanish Ministry of Economy and Digital Transformation (now the Ministry for Digital Transformation and Public Service).

Author Beatriz Guerrero García holds a USAL 2023 pre-doctoral contract, co-financed by Banco Santander.

References

Hema Ala and Dipti Sharma. 2020. Graph based automatic domain term extraction. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): TermTraction 2020 Shared Task*, pages 1–4, Patna, India. NLP Association of India (NLPAI).

Rosa Estopà Bagot. 1999. Extracció de terminologia: elements per a la construcció d'un seacuse (sistema d'extracció automàtica de candisats a unitats de significació especialitzada).

Didier Bourigault. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *COLING 1992 Volume 3: The 15th International Conference on Computational Linguistics*.

Julien Breton, Mokhtar Mokhtar Billami, Max Chevalier, Ha Thanh Nguyen, Ken Satoh, Cassia Trojahn, and May Myo Zin. 2025. Leveraging llms for legal terms extraction with limited annotated data. *Artificial Intelligence and Law*, pages 1–27.

Marija Brkić Bakarić and Ivana Lalli Paćelat. 2024. Quick and easy term extraction: Making use of llm-based chatbots. *Corpora in Language Learning, Translation and Research: Book of Abstracts*, pages 31–33.

M. Teresa Cabré, Rosa Estopà, and Mercè Lorente. 1996. Terminología y fraseología. In *V Simposio de Terminología Iberoamericana*.

Maria Teresa Cabré. 2009. La teoría comunicativa de la terminología, una aproximación lingüística a los términos. *Revue Francaise de Linguistique Appliquee*, 14:9–15.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Real Academia Española and Consejo General del Poder Judicial. 2017. *Libro de estilo de la Justicia*.

David A. Evans and Chengxiang Zhai. 1996. Noun phrase analysis in large unrestricted text for information retrieval. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 17–24, Santa Cruz, California, USA. Association for Computational Linguistics.

Pamela Faber and Marie-Claude L'Homme, editors. 2022. *Theoretical Perspectives on Terminology*, volume 23. John Benjamins Publishing Company.

Fethi Fkih and Mohamed Nazih Omri. 2012. Complex terminology extraction model from unstructured web text based linguistic and statistical knowledge. *International Journal of Information Retrieval Research (IJIRR)*, 2(3):1–18.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: The c-value/ nc-value method. volume 3, pages 115–130.

Christophe Gaudet-Blavignac, Vasiliki Foufi, Mina Bjelogrlic, and Christian Lovis. 2021. Use of the systematized nomenclature of medicine clinical terms (snomed ct) for processing free text in health care: Systematic scoping review. *J Med Internet Res*, 23(1):e24594.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023. Llms accelerate annotation for medical information extraction.

María Tadea Díaz Hormigo. 2011. Sobre los denominados sustantivos deverbales de acción. *Lorenzo Hervás*.

Murhaf Hossari, Soumyabrata Dev, and John D. Kelleher. 2019. Test: A terminology extraction system for technology related terms. In *Proceedings of the 2019 11th International Conference on Computer and Automation Engineering*, ICCAE 2019, page 78–81, New York, NY, USA. Association for Computing Machinery.

Dunia Hourani-Martín. 2023. Unidades fraseológicas suboracionales en un corpus de textos jurídicos: el esquema [prep. + sust. + prep.]. *Revista de Filologia de la Universidad de La Laguna*, pages 187–206.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Tolga Kayadelen, Adnan Ozturel, and Bernd Bohnet. 2020. A gold standard dependency treebank for Turkish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5156–5163, Marseille, France. European Language Resources Association.

Yusuke Kimura, Kazuma Kusu, Kenji Hatano, and Tokiya Baba. 2021. Automatic terminology extraction using a dependency-graph in nlp. In *Innovations in Bio-Inspired Computing and Applications: Proceedings of the 11th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2020) held during December 16-18, 2020 11*, pages 411–421. Springer.

Nikola Ljubešić, Darja Fiser, and Tomaž Erjavec. 2019. Kas-term: Extracting slovene terms from doctoral theses via supervised machine learning.

Elizaveta Loginova, Anita Ramm, Helena Blancafort, Marie Guégan, Tatiana Gornostay, and Ulrich Heid. 2012. Reference lists for the evaluation of term extraction tools.

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. 2016. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, 19:59–99.

Macken, Lieve and Lefever, Els and Hoste, Veronique. 2013. TExSIS: bilingual terminology extraction from parallel corpora using chunk-based alignment. *TERMINOLOGY*, 19(1):1–30.

Pazienza Maria Teresa, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2006. *Terminology Extraction: An Analysis of Linguistic and Statistical Approaches*, volume 185, pages 255–279.

Patricia Martín-Chozas and Pablo Calleja. 2018. Challenges of terminology extraction from legal spanish corpora. In *2nd Workshop on Technologies for Regulatory Compliance*, pages 73–83.

Patricia Martín-Chozas, Karen Vázquez-Flores, Pablo Calleja, Elena Montiel-Ponsoda, and Víctor Rodríguez-Doncel. 2022. Termitup: Generation and enrichment of linked terminologies. *Semantic Web*, 13:967–986.

Rogelio Nazar. 2011. A statistical approach to term extraction. *International Journal of English Studies (IJES)*, 11.

Gilberto Anguiano Peña and Catalina Naumis Peña. 2015. Extracción de candidatos a términos de un corpus de la lengua general. *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*, 29:19–45.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868, Portorož, Slovenia. European Language Resources Association (ELRA).

Mariano Rico, Pablo Calleja, Patricia Martín Chozas, Patricia Martín, and Elena Montiel. 2019. Extracting terminologies in the legal domain: a syntactic pattern-based approach for spanish. In *JURIX Iberlegal Workshop*.

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 85–94, Marseille, France. European Language Resources Association.

Sergio Rodríguez-Tapia. 2024. *Gestión terminológica, corpus especializados y extracción automática de terminología en español*. Comares.

Aivaras Rokas, Sigita Rackevičienė, and Andrius Utka. 2020. Automatic extraction of lithuanian cybersecurity terms using deep learning approaches.

Miriam Seghiri. 2017. Corpus e interpretación biosanitaria: extracción terminológica basada en bitextos del campo de la neurología para la fase documental del intérprete. *Panace@ Revista de Medicina, Lenguaje y Traducción*, 18:123–132.

Chelo Vargas Sierra. 2010. *Combinatoria terminológica y diccionarios especializados para traductores*, pages 17–46. Comares.

Kento Sugimoto, Toshihiro Takeda, Jong-Hoon Oh, Shoya Wada, Shozo Konishi, Asuka Yamahata, Shiro Manabe, Noriyuki Tomiyama, Takashi Matsunaga, Katsuyuki Nakanishi, et al. 2021. Extracting clinical terms from radiology reports with deep learning. *Journal of Biomedical Informatics*, 116:103729.

Hanh Thi Hong Tran, Matej Martinc, Jaya Caporusso, Antoine Doucet, and Senja Pollak. 2023. The recent advances in automatic term extraction: A survey.

Bianca Vitalaru. 2019. Organización temática de terminología jurídica para traductores: proceso de elaboración de una ontología del proceso penal en español-rumano-inglés. *Hermēneus. Revista de traducción e interpretación*, pages 463–514.

Cheng Zheng, Na Deng, Ruiyi Cui, and Hanhui Lin. 2023. Terminology extraction of new energy vehicle patent texts based on bert-bilstm-crf. In *International Conference on Emerging Internetworking, Data & Web Technologies*, pages 190–202. Springer.