A Multi-Stage Heuristic Filtering Pipeline for Refining a Spanish Legal Corpus for Natural Language Processing[1]

Nikolai TIURIN
https://orcid.org/0009-0002-9664-3547; nikolai.tiurin@autonoma.cat
Universitat Autònoma de Barcelona (SPAIN)

Xavier BLANCO ESCODA
https://orcid.org/0000-0001-8210-3668; xavier.blanco@uab.cat
Universitat Autònoma de Barcelona (SPAIN)

Suggested citation: TIURIN, N., BLANCO, X. (2025), A multi-stage heuristic filtering pipeline for refining a Spanish legal corpus for natural language processing, *Langue(s) & Parole,* **10**, 37-56, https://doi.org/10.5565/rev/languesparole.153

Abstract

This research presents a multi-stage heuristic pipeline to refine the Spanish *Boletín Oficial del Estado* (BOE) corpus for Natural Language Processing tasks. Raw legal corpora are often filled with noise, including OCR errors, lists, tables, and non-textual placeholders, making them unsuitable for training language models. Our methodology first normalizes the text by correcting character-level errors and repairing hyphenation. Subsequently, it applies a series of filters based on quantifiable metrics, such as newline character ratios, non-alphabetic character counts, and misspelled word percentages, to detect and discard structurally and semantically unsuitable segments. A key contribution is the novel Combined Borderline Score (CBS), which identifies and removes marginal segments that are close to multiple failure thresholds. The result is a significantly cleaner corpus of legal texts, providing a high-quality foundation for training models for tasks like automatic text simplification and offering a reusable methodology for cleaning other large and diverse legal texts.

Keywords: natural language processing (NLP), corpus cleaning, legal text processing, heuristic filtering, data pre-processing

Résumé

Cette recherche présente une séquence heuristique en plusieurs étapes pour nettoyer le corpus du *Bulletin Officiel de l'État* (BOE) espagnol à des fins de traitement

[1] This research is funded by the European Union in the frame of the iRead4Skills project (Intelligent Reading Improvement System for Fundamental and Transversal Skills Development) (Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07).

automatique du langage naturel. Les corpus juridiques bruts contiennent souvent beaucoup de « bruit », comme des erreurs d'OCR, des listes et des tableaux, qui les rendent inadaptés à l'entraînement de modèles linguistiques. Notre méthodologie commence par normaliser le texte en corrigeant les erreurs dans les caractères et en ajustant les coupures de mots par des traits d'union. Ensuite, elle applique une série de filtres basés sur des métriques quantifiables telles que le ratio de sauts de ligne, le pourcentage de caractères non alphabétiques et de mots mal orthographiés, afin d'écarter les segments structurellement ou sémantiquement inadaptés. Une contribution clé est le Score Combiné de Seuil (SCS), une technique novatrice qui identifie et élimine les segments marginaux proches de nombreux seuils d'exclusion. Le résultat est un corpus de textes juridiques nettement plus propre, fournissant une base de haute qualité pour l'entraînement de modèles destinés à des tâches telles que la simplification automatique de textes, et offrant une méthodologie réutilisable pour le nettoyage d'autres grands corpus juridiques hétérogènes.

Mots clé : traitement automatique du langage naturel (TALN), nettoyage de corpus, traitement de textes juridiques, filtrage heuristique, prétraitement de données

Resumen

Esta investigación presenta una secuencia heurística de múltiples etapas para limpiar el corpus del *Boletín Oficial del Estado* (BOE) español para tareas de Procesamiento del Lenguaje Natural. Los corpus jurídicos en bruto suelen contener una gran cantidad de "ruido", como errores de OCR, listas y tablas, lo que los hace inadecuados para el entrenamiento de modelos de lenguaje. Nuestra metodología en primer lugar normaliza el texto corrigiendo errores en los caracteres y ajustando la separación de palabras con guiones. Posteriormente, aplica una serie de filtros basados en métricas cuantificables como la proporción de saltos de línea, el porcentaje de caracteres no alfabéticos y el de palabras mal escritas para rechazar segmentos estructural o semánticamente inapropiados. Una contribución clave es la novedosa Puntuación Combinada de Umbral (PCU), que identifica y elimina segmentos marginales cercanos a múltiples umbrales de exclusión. El resultado es un corpus de textos jurídicos significativamente más limpio, que proporciona una base de alta calidad para entrenar modelos para tareas como la simplificación automática de textos y ofrece una metodología reutilizable para la limpieza de otros grandes corpus jurídicos heterogéneos.

Palabras clave*:* procesamiento del lenguaje natural (PLN), limpieza de corpus, procesamiento de textos jurídicos, filtrado heurístico, preprocesamiento de datos

Resum

Aquesta investigació presenta una seqüència heurística de múltiples etapes per a netejar el corpus del *Butlletí Oficial de l'Estat* (BOE) espanyol per a tasques de Processament del Llenguatge Natural. Els corpus jurídics en brut solen contenir una gran quantitat de "soroll", com ara errors d'OCR, llistes i taules, que els fan inadequats per a l'entrenament de models de llenguatge. La nostra metodologia en primer lloc normalitza el text corregint errors de caràcters i ajustant la separació de paraules amb guions. Posteriorment, aplica una sèrie de filtres basats en mètriques quantificables com la proporció de salts de línia, el percentatge de caràcters no alfabètics i el de paraules mal escrites per a rebutjar segments estructuralment o semànticament inapropiats. Una contribució clau és la nova Puntuació Combinada de Llindar (PCL), que identifica i elimina segments marginals propers a múltiples llindars d'exclusió. El resultat és un corpus de textos jurídics significativament més net, que proporciona una base d'alta qualitat per a entrenar models per a tasques com la simplificació automàtica de textos i ofereix una metodologia reutilitzable per a la neteja d'altres grans corpus jurídics heterogenis.

Paraules clau: processament del llenguatge natural (PLN), neteja de corpus, processament de textos jurídics, filtratge heurístic, preprocessament de dades

1. Introduction

Texts pertaining to legal domain represent a significant field of study in Natural Language Processing (NLP). Legal texts, such as laws, regulations, contracts and agreements, determine many aspects of our everyday lives, but their inherent complexity, specialized terminology and confusing syntactic structures with intertwining and nested clauses make these documents largely inaccessible to the general public. This problem of poor readability of legal texts is widely recognized and Spanish government has launched several initiatives aimed at reduction of complexity of legal documents (Vallejo, 2021). The recent advancements in development of Large Language Models (LLMs) present an unprecedented opportunity to alleviate this problem (Anh *et al.,* 2023), (Quevedo *et al.,* 2024) and provide people that do not have education or professional background in law and legal procedures with effective tools that can facilitate their understanding of matters described in complex legal documents. Enormous capabilities of LLMs in processing of textual data make them very suitable for automatic text simplification, summarization and legal information retrieval (Cemri *et al.,* 2022), (Arfat *et al.,* 2024) and leveraging these capabilities can prove very efficient in reduction of complexity of legal texts and improvement of their readability.

However, the success of any sophisticated NLP application is fundamentally dependent on the quality and suitability of the underlying training data. While vast repositories of legal documents are publicly available, they are rarely in a “ready-to-use” format for training language models and this is particularly true for official gazettes like the Spanish *Boletín Oficial del Estado* (BOE), a comprehensive digital archive of Spanish law. The BOE corpus, as collected and made available by Gutiérrez-Fandiño *et al.* (2019), is an invaluable resource, offering thousands of documents and millions of words that represent regulations governing numerous aspects of human activities, such as commerce, education, medicine, etc. However, in its raw form, it is a heterogeneous collection plagued by a wide array of noise and artifacts that pose significant challenges for downstream NLP tasks (Hua *et al.,* 2022).

Initial analysis reveals that the raw BOE corpus is far from a clean collection of legal texts, it is riddled with systemic issues stemming from its multi-decade history and varied digitization processes. These issues include:

- OCR and digitization errors: Scanned documents introduce optical character recognition (OCR) artifacts, such as character substitutions (e.g., a cedilla “¸” instead of a comma), incorrect sentence segmentation, especially in cases with end-of-line hyphenation when words get split in two parts that appear in two different lines.
- Structural and formatting artifacts: Apart from standard legal text in form of articles, paragraphs and subparagraphs, the BOE corpus also contains a vast amount of structured and semi-structured data, including tables, itemized lists, annexes with technical descriptions and specifications of goods, and purely numerical data. These segments lack the syntactic and semantic structure required for tasks like simplification, they do not define rights, obligations or procedures which is typical for laws, regulations and contracts, but rather provide detailed information regarding the objects of regulation and frequently contain highly specialized technical data which is not actually suitable for simplification as it involves risks of omitting very important details and hindering correct and complete understanding of the document.
- Textual placeholders: A significant portion of the documents (over 13% in our initial analysis) are not actual legal texts but rather

placeholders, such as the string *"Texto no disponible. Consulte el documento PDF de esta disposición."*, which corrupt the dataset with irrelevant entries.

- Multilingual content and specialized jargon: International treaties and technical regulations often include text in multiple languages or consist of highly specialized, non-legal terminology (e.g., lists of chemical compounds or biological species) that is not common in general legal language.
- Heavily referential text: Many articles in legal documents primarily consist of citations and references to other legal provisions (e.g., *"de acuerdo con los artículos 22 y 86 al 89 del Estatuto"*). While legally sound, these segments are not self-contained and cannot be meaningfully simplified without complex, external context retrieval.

Feeding such noisy data directly into a fine-tuning process for a task like text simplification would be counterproductive, the machine learning model could learn to replicate OCR errors, generate lists instead of complete and grammatically correct sentences, and struggle to parse the underlying legal concepts, which would ultimately lead to its poor performance and unreliable outputs. Therefore, a rigorous and systematic pre-processing and filtering methodology is not just a preliminary step but a critical research challenge in itself (Shaheen *et al.,* 2020).

To address this challenge, this paper introduces a robust, multi-stage heuristic filtering pipeline designed to semi-automatically identify and remove unsuitable segments from the raw BOE corpus (Sánchez *et al.,* 2025). Our methodology moves beyond simple text cleaning to perform a deeper, content-aware analysis, we first normalize textual content by correcting common OCR errors and resolving hyphenation and then we employ a series of quantifiable, heuristic-based filters to discard segments that are structurally or semantically inappropriate. These filters leverage metrics such as the ratio of newline characters, the percentage of non-alphabetical characters, and the rate of misspelled words (as a proxy for non-Spanish text or strings with dense technical terminology). Finally, we introduce a novel Combined Borderline Score (CBS), which aggregates these heuristics to identify and rank segments that are marginal on multiple criteria, allowing for a more nuanced and effective final filtering stage.

The primary contributions of this work are:

- A systematic characterization of the diverse noise and artifacts present in a large, real-world Spanish legal corpus.
- The design and implementation of a modular, multi-stage filtering pipeline that uses a combination of normalization and quantifiable heuristics to refine the corpus.
- The introduction of a Combined Borderline Score (CBS), a novel technique for identifying and removing text segments that are unsuitable for NLP tasks based on multiple weak signals.
- The creation and release of a significantly refined, analysis-ready version of the BOE corpus, providing a valuable resource for future research in Spanish legal NLP.

The following sections detail the methodology of this pipeline, present the results of its application to the BOE corpus, and discuss the properties of the resulting dataset and its implications for training language models for legal text simplification.

2. Data and Methods

To address the challenges of noise and diversity of content outlined in the introduction, we developed a comprehensive, multi-stage pipeline to transform the raw BOE corpus into a refined dataset suitable for downstream NLP tasks (Lin, Cheng, 2024). This process consists of a sequence of stages, each detailed in the following subsections. It starts with an initial characterization and structural processing of the source data, followed by a text normalization and repair stage targeting low-level artifacts. Subsequently, we apply a pipeline of heuristic-based filters to remove semantically unsuitable content and the final stage is the application of our novel Combined Borderline Score (CBS) for a more nuanced filtering pass.

2.1 Corpus Description and Initial Processing

The foundation of this research is the largest part of the corpus collected by Gutiérrez-Fandiño *et al.* (2019), which is derived from the *Boletín Oficial del Estado* (BOE). For the remainder of this paper, we will refer to this as the BOE corpus. In its raw form, this dataset consists of a 3,6 GB text file structured around the "TEXTO ORIGINAL" marker, which serves as a separator for individual documents. The total amount of documents in the BOE corpus summed up to 216 484, with an average length of 17 097 characters per document.

However, a closer look at the corpus contents revealed significant data quality issues that necessitated an initial filtering stage. It was found that 29 097 documents marked with a "TEXTO ORIGINAL" separator are indeed only 67 characters long and their text is *"Texto no disponible. Consulte el documento PDF de esta disposición"*. These entries, likely artifacts from an automated web-scraping process encountering broken links, represent over 13% of the total document count. In fact, the BOE corpus contains 588 more documents with length below 150 characters. To purge these non-documents and avoid noise that could be caused by their presence, we established a 150-character threshold as a minimum for any text segment to be included in our dataset and removal of these short non-documents from the BOE corpus produced a refined set of 186 799 items with an increased average length of 19 804 characters.

Following this initial filtering, the next step in obtaining short segments suitable for feeding to an LLM was to split the large, valid documents into smaller, semantically meaningful units. Legal texts, by their nature, provide a convenient way to achieve this since they are usually well structured and divided into sections, chapters, articles, paragraphs, etc. The splitting was accomplished through a two-tiered segmentation process: first, we split the document texts into major sections using explicit high-level markers such as "Artículo", "Capítulo", "Anexo" and others and then, to obtain shorter text segments that would express a single idea or concept, we implemented another splitting algorithm.

The algorithm consists in scanning the texts for markers that could serve as indicators for splitting, such as numbers followed by a period at the beginning of a line (e.g., 1.), single letters followed by a period or a closing parenthesis (e.g., a., a)), and common bullet points (*, •, -). The algorithm checks every segment obtained in the previous step and returns a list of strings which represent the original string split at these markers. If no matching pattern is found, the algorithm returns the original text segment it received as input. This initial structural processing yielded a large collection of text segments of varying length and quality, which then served as the input for the deeper normalization and heuristic filtering steps described in the following sections.

2.2 Text Normalization and Repair

After the initial structural segmentation, the resulting text segments required significant normalization to correct low-level textual artifacts. These artifacts, primarily originating from the Optical Character Recognition (OCR) process of scanned documents, can introduce "noise" that would distort the process of machine learning (Adamczyk, Hula, 2024), (Silveira *et al.,* 2023). Our normalization pipeline consisted of two main stages: character-level correction and hyphenation repair.

First, we addressed the presence of non-standard characters incorrectly recognized by OCR programs. For example, in some cases these texts contain a character called a cedilla (¸) where a comma should appear, or a single comma quotation mark (‚) that, while visually similar, is processed differently by computer algorithms. Given the diversity of such errors, we decided to resort to filtering out all characters which are not included into a compiled allowlist. The list of allowed characters, aside from letters and numbers, is:

!"#$%&'()*+,-./:;<=>?@[]^_{|}~¡£¥§°±¿×—•…‰€−≠≤≥.

As for the characters that were filtered out, many of them represented different kinds of blank spaces, such as non-breaking space, thin space, hair space, zero-width space, etc. We have replaced some of the non-standard characters with their standard alternatives, for example, the less-frequent variants of the "less-than" sign (‹, <) were replaced with the standard one (<). We also replaced all non-standard spaces, including tabulation signs and double spaces, with regular single spaces, and replaced the common Spanish short form *nº* (for *número*) with a hash sign (#).

Second, we developed a procedure to fix issues with hyphenated words split across lines. Since many of the documents in our corpus were digitized by scanning hard copies, they contained numerous instances of hyphenation inside words. To detect these instances, we searched for strings containing at least two letters followed by a hyphen and a newline symbol. We then concatenated the two word parts with second part retrieved from the beginning of the next line and checked if this candidate word is present in a vocabulary of the Spanish language. Crucially, we also verified that at least one of the two original parts (the one preceding the hyphen or the one on the new line) is *not* present in the vocabulary as a standalone word. This check is intended to avoid the erroneous merging of two distinct words that are legitimately separated by a hyphen.

For the vocabulary check, we evaluated both `spacy` and `hunspell` Python libraries. While `hunspell` is more demanding in terms of diacritic accent presence, it correctly handled ambiguous cases where spacy failed (e.g., correctly joining *adminis-* and *tración* where `spacy` recognized both as valid separate entries) and since it includes algorithms that correctly recognize morphological variations of Spanish words, it does not yield as much false negative results as `spacy` does. After comparison, we settled on using the `hunspell` vocabulary as it helped to implement more hyphenation corrections. In total, the total amount of lines that end with a hyphen detected by our algorithm was 41 495, and our approach allowed to correct 24 340 of these instances. These normalization and repair steps produced a cleaner set of text segments, preparing them for the higher-level heuristic filtering described next.

2.3 Heuristic-Based Filtering Pipeline

Following normalization, we applied a multi-stage filtering pipeline to identify and remove segments unsuitable for text simplification based on their structural and content-based properties and this pipeline consists of a sequence of heuristic-based filters, each designed to target a specific type of noise present in the corpus.

2.3.1 Duplicate Segment Removal

The first filtering step addressed data redundancy which is common in text corpora (Garcia *et al.,* 2024). Here, we have used the built-in functionality of `pandas` Python library `DataFrame` class to drop duplicate segments across the entire dataset. This method only identifies exact duplicates and, consequently, it has not affected any segments that differ by a single letter, white space or punctuation sign. This procedure detected and removed 340 796 duplicate segments out of 6 155 257 total segments, which makes up 4,6% of the overall segment count. The vast majority of duplicate segments were quite short, which is expected in legal texts that contain many standardized phrases and sentences.

2.3.2 Newline Character Ratio for List Detection

A significant portion of legal documents consists of not very cohesive content like lists and tables. We can assume that such lists are not suitable for simplification, as omitting an item can have critical consequences for the reader. To identify list-like structures, our algorithm calculated the density of newline characters ("\n" in Unicode) within each text segment. We observed that segments with a high ratio of these characters consistently corresponded to text formatted with frequent visual line breaks, which is very common in annexes or tables.

Manual analysis of several dozen segments led us to the conclusion that a value of 1,9% can be a good threshold to separate semantically cohesive sentences from list-like structures. Application of this 1,9% threshold resulted in the rejection of 386 915 segments or 5,2% of the BOE corpus. While this threshold is quite soft, it effectively removes many segments that look like lists or contain a mixture of regular sentences followed by short, disjointed lines.

2.3.3 Non-Alphabetical Character Ratio for Content-Type Filtering

The next step in our pre-processing was the calculation of non-alphabetical characters in all segments. The rationale is that if a segment contains no text, or if text makes up a minor part of a segment, this can serve as an indicator that this segment does not fit our purpose of training a model for automatic simplification. We observed that both extremes of the range of values for this metric indicate certain issues with text segment contents. An excessively low percentage of non-alphabetical characters often means that the text segment has missing white spaces and words are agglutinated together. Conversely, an excessively high percentage indicates that the text segment may contain no text at all, only numbers, mathematical or typographic symbols, and white spaces.

Based on this, we established a two-sided filter:

- A lower threshold of 10%: All text segments that have less than 10% non-alphabetical characters were rejected. This small filter removed 905 segments where text was corrupted or agglutinated.
- An upper threshold of 29%: All text segments that contain 29% or more of non-alphabetical characters are considered unsuitable. This value may seem low and too strict since it includes white spaces and punctuation signs, but analysis of text segments that are just above this threshold shows that even though many of them do contain full sentences with complex syntactic structures, they also include a lot

of references to other laws or regulations and it does not look very probable that they can be simplified without retrieving the corresponding text from these referenced documents and converting the original text segment into a completely new text. Apart from that, simplifying texts with resolution of external references is an enormously complex task on its own and it is outside the scope of this research. Application of this 29% low-pass filter resulted in the rejection of 545 860 segments (7,3% of the corpus).

2.3.4 Misspelled Word Ratio as a Proxy for Language and Domain

Another specific aspect of the texts in the BOE corpus is the presence of text strings written in languages other than Spanish, which can introduce "noise" into our dataset. Rather than using a standard language identification library, which can be unreliable on short or mixed-language segments, we opted for a different method. We used the `hunspell` spellchecking Python library to check every text segment and if the spellchecker detects too many words that are not familiar to it, we can conclude that the segment contains text written in a different language or is otherwise unsuitable.

This method serves a dual purpose. Many segments where the content is rather "technical" than "legal" contain plenty of words that might not be present in a general-purpose spellchecking dictionary, such as lists of chemical substances or species of plants. The spellchecker allows us to detect these rare terms, and a significant amount of unfamiliar words in a segment indicates that it does not contain the kind of cohesive legal text segments we would like to keep. After analysis, we settled on 25% of misspelled words as the threshold value. All segments where misspelled words make up more than 25% of the total word count are considered unsuitable for automatic simplification. Application of this threshold resulted in the rejection of 233 459 text segments.

2.4 Combined Borderline Score for Nuanced Filtering

The hard threshold values applied in the previous steps effectively detect text segments that are clear outliers. However, these thresholds are not very strict, as we wanted to avoid excessive rejection of potentially useful data. In fact, many of the remaining segments have their parameter values quite close to the hard thresholds. Our next step, therefore, was to establish a combined score that would allow us to detect text segments that are borderline on multiple criteria simultaneously. This Combined Borderline Score (CBS), in theory, should be able to point out the segments where two or three parameters are very close to their respective threshold values, allowing us to perform a more refined filtering of segments where the content may be more "technical" than "legal".

The composition of this combined borderline score is quite simple: we divide the measured value of each parameter by the corresponding threshold and sum up the results. If a measured value is very close to the threshold value, the division result is very close to 1.

A small modification of this procedure is only needed for calculation of CBS in terms of non-alphabetical characters. Since this parameter has both minimum (10 %) and maximum (29 %) threshold values, as described above, division by the minimum threshold yields higher values for those segments where proportion of non-alphabetical characters is further away from the threshold value. At the same time, the segments that are closer to the threshold will produce lower values closer to 1 as in case with calculations for the other two parameters. In order to compensate for this effect, we just subtract the division result from 2 and lower values will again be closer to one and follow the descending pattern where text segments that are more distant from the threshold produce values closer to zero. We select the higher of the two division results as it is more indicative of closeness to a threshold and its contribution to the final CBS values will be more significant.

The parameters and thresholds used in this calculation are formally defined in Table 1.

Symbol	**Description**	**Threshold Value**
Pnl	Measured percentage of newline characters	Tnl = 1,9%

Pna	Measured percentage of non-alphabetical characters	Tna_low = 10,0%
		Tna_hi = 29,0%

Pms	Measured percentage of misspelled words	Tms = 25,0%

Table 1. Parameters and hard thresholds used in the calculation of the CBS.

The formula for the Combined Borderline Score (CBS) is as follows:

$$\text{CBS} = \text{Pnl} / \text{Tnl} + \text{Pms} / \text{Tms} + \max(\text{Pna} / \text{Tna\_hi}, 2 - \text{Pna} / \text{Tna\_low})$$

In cases where all three measured parameters are very close to their respective thresholds, the value of the CBS will be close to 3.

The final step was to determine a reasonable threshold value for this combined score. As with the hard thresholds, the determination of the CBS threshold is based on the manual analysis of several dozen segments. This process consists of sorting the text segments by their CBS in descending order and iteratively selecting and evaluating a threshold. After analysis, we have decided to settle on a Combined Borderline Score threshold of 1,6. Although this value is just above the midline of the possible range [0, 3], in practice, it turns out that more than 95% of the segments with a CBS at or above 1,6 are not suitable for automatic simplification. Manual analysis of these segments reveals that while they may be close to what we could call "legal text", their content, syntactic structure or semantics make them unsuitable for simplification even in a manual mode. This final filtering step allows for a more nuanced removal of "noisy" data than the individual hard thresholds alone. Application of the 1,6 threshold for combined borderline score has resulted in rejection of 162 098 text segments (2,1% of the refined BOE corpus).

3. Results

Filtering Step	Segments Remaining	% of Total	Characters Remaining	% of Total
Initial Split (Artículo, etc.)	7 431 216	100%	3,4 billion	100%
Length Filter (>150 char)	6 155 257	82,8%	3,3 billion	97,5%
Duplicate Removal	5 814 461	78,2%	3,0 billion	88,1%
Hard Thresholds (Combined)	5 022 424	67,6%	2,7 billion	79,4%
CBS Filter (>1.6)	4 860 326	65,4%	2,6 billion	76,5%

Table 2. Data cascade table.

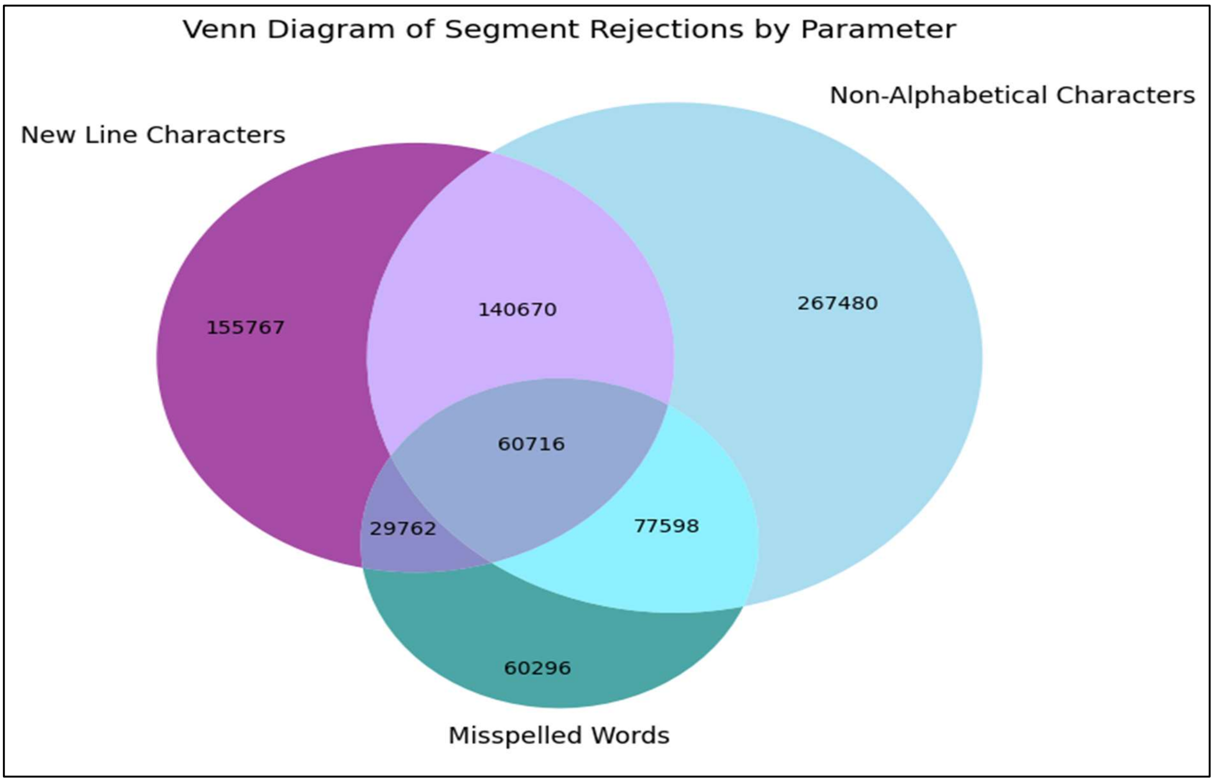

Figure 1. Overlaps of filtering thresholds.

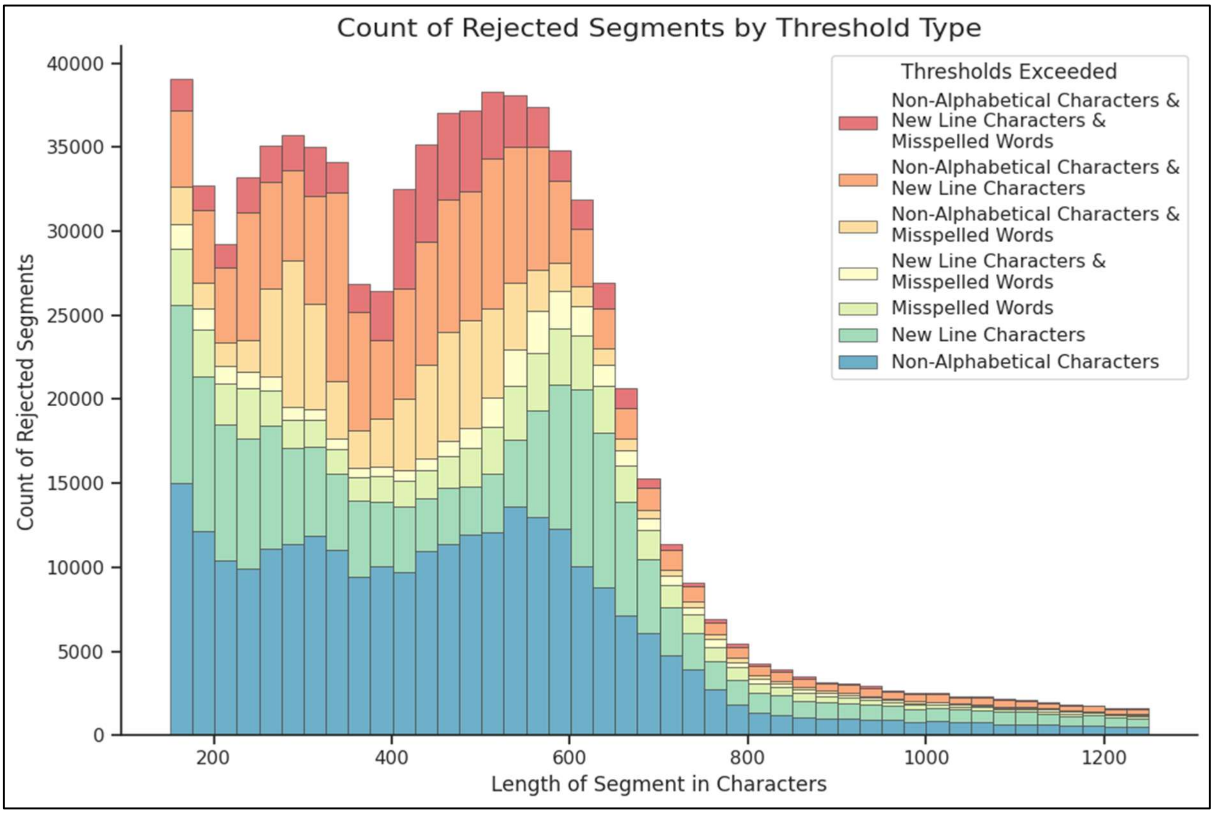

Figure 2. Distribution of rejection thresholds across segment length in characters

Table 1 shows how sequential application of our methods affected the size of the BOE corpus, reducing the amount of "noisy" segments that could contaminate the training dataset and consequently affect the performance of the machine learning model. As the table confirms, application of all filters has discarded about one fourth of the initial corpus volume measured in characters and there is still plenty of data to work with.

Figure 1 provides visual representation of overlaps for the three hard thresholds in terms of segments where the thresholds were exceeded and amount of segments that have not passed one or several hard filters.

Figure 2 shows the shows the distribution of all three hard thresholds and their combinations that were exceeded and resulted in rejection of the corresponding segments. As the histogram shows, there are no sharp changes in proportions of the exceeded thresholds, the rejections are distributed quite evenly across various character counts in text segments. This is a good indicator that count of characters in a segment does not affect triggering of any threshold or combination of thresholds.

4. Discussion

The application of our multi-stage filtering pipeline resulted in the rejection of a substantial portion of the original BOE corpus, over 10% of segments by count and nearly 9% by character volume were removed by the hard thresholds alone. This outcome underscores the critical need for deep, content-aware cleaning of large, web-sourced legal corpora before their use in sensitive NLP tasks.

4.1 Effectiveness of the Heuristic Pipeline

The pipeline's effectiveness originates from its layered approach: the initial removal of exact duplicates was a necessary first step, and the finding that the vast majority of duplicate segments are quite short was expected, as legal texts generally have a lot of standardized phrases. More significantly, the heuristic-based filters proved highly capable of identifying distinct categories of noise. The Venn diagram presented in the results (Figure 1) illustrates that many rejected segments triggered multiple filters. This "double triggering" provides us with more confidence that the segments that were rejected are in fact not suitable for our purposes and reinforces the validity of the chosen heuristics.

By systematically removing the most obvious outliers first, such as lists, numerical tables, and segments with high proportion of misspelled words, we were able to more effectively identify subtle issues in the remaining data. This process can be compared to looking for a lost key in tall grass: if we trim the grass first, it becomes easier to find the key. Similarly, removing the most prominent "noise" made the remaining borderline cases stand out more which allowed us to use the Combined Borderline Score for more precise filtering of corpus texts.

4.2 Analysis of Filtered Content and Methodological Justifications

A key aspect of our methodology lies in the design of heuristics that act as proxies for semantic unsuitability. The newline character ratio, for instance, proved to be a simple yet powerful indicator of non-cohesive text. We can assume that such lists are not suitable for simplification; if we delete any item from a list and present the reduced list to a reader, it can have critical consequences. For example, omitting a protected species of fish, herbs or mushrooms from a list in a conservation law could lead a reader to cause ecological damage and face serious legal responsibility.

Similarly, using a spellchecker as a filter for misspelled words was not primarily about correcting text but rather aimed at identifying content that does not belong to our target legal domain. This method allowed us to detect segments written in languages other than Spanish, but also served a dual purpose: it flagged segments where the content is rather "technical" than "legal", such as lists of chemical substances or species of plants, which contain many words not present in a general-purpose spellchecking dictionary.

4.3 Limitations and Inevitable Trade-offs

While effective, our automated approach involves necessary trade-offs and has limitations. One of the most significant challenges is handling heavily referential text. For instance, consider the following segment, which was filtered out due to its high percentage of non-alphabetical characters (29%):

> *(3) El artículo 17, apartado 1, letra c), del Reglamento de Ejecución (UE) nº 391/2013 dispone que la Comisión evalúe las tarifas unitarias de 2014 para las zonas de tarificación presentadas por los Estados miembros a la Comisión con fecha límite de 1 de junio de 2013, de acuerdo con los requisitos del artículo 9, apartados 1 y 2, de dicho Reglamento. Esta evaluación deberá comprobar la conformidad de*

> *las tarifas unitarias de 2014 con los Reglamentos de Ejecución (UE) n o 390/2013 y (UE) n o 391/2013.*

This is indeed a valid legal text, but increasing its clarity and readability would require retrieving information from multiple referenced documents. This research does not intend to cover all possible methods for simplification of legal texts, and we have deliberately chosen to exclude such segments, as their simplification would require a complex reference-resolution system beyond the scope of this work.

Furthermore, the heuristic nature of the pipeline means that false negatives are inevitable. In our hyphenation correction algorithm, for example, a word like *contenido* split as *con-* and *tenido* would not be joined, as both parts are legitimate separate words. It does not seem possible to cover all probable cases, and we only have to accept inevitable errors and hope they do not deteriorate our final results too much. Trade-offs like this are inevitable in the automatic processing of vast amounts of text, and we suppose that cases where our approach correctly identifies two independent words and does not join them would outnumber the cases where it fails.

4.4 Implications for the Final Corpus and Future Work

The rigorous filtering process has produced a corpus that is substantially cleaner and more focused on cohesive legal text. However, this has an important implication: the final training data has its own distinct characteristics, which means that at the stage of testing our future machine learning model, we might need to apply similar pre-processing procedures to input texts in order to make them more similar to the training dataset, thereby reducing domain shift and improving performance. The refined corpus now serves as a high-quality foundation for the next stage of this research: using it to create a parallel dataset for fine-tuning a language model for automatic simplification.

5. Conclusion

The successful application of advanced NLP techniques to texts from the legal domain depends on the availability of high-quality, specialized data. This paper has described the development of a robust, multi-stage pipeline designed to transform a large raw legal repository, the Spanish BOE corpus, into a refined dataset suitable for downstream tasks like automatic text simplification. Our work was aimed at resolving the common issues present in large corpora that contain a lot of noise, including OCR artifacts, structural inconsistencies and semantically unsuitable content.

Our methodology combined low-level text normalization with a sequence of content-aware heuristic filters. The hard threshold values applied to the *BOE* corpus allowed us to detect text segments that are clear outliers, while the introduction of a Combined Borderline Score (CBS) provided a more nuanced tool for identifying segments that are still unsuitable for our purpose even without exceeding any hard threshold. The entire procedure was designed to minimize the "noise" that could distort the process of machine learning and to improve the quality of the training dataset, which will hopefully lead to a better outcome for this entire work.

This process, however, is not without its limitations, we acknowledge that our automated approach involves necessary compromises, it does not seem possible to cover all probable cases, and we must accept inevitable errors. Trade-offs like these are inevitable in the automatic processing of vast amounts of text, for instance, our decision to exclude texts with a lot of references to other documents was a conscious choice to narrow the scope of the simplification task, as resolving such references is a complex research problem in its own right.

The primary contribution of this work is the resulting refined corpus, which serves as a critical resource for future research. A key implication of this process is that at the stage of testing our future machine learning model, we might need to apply similar pre-processing procedures to input texts in order to make them more similar to the training dataset. As a result, the refined *BOE* corpus now can be used as a solid foundation for our central research objective: creating a high-quality parallel corpus to fine-tune a language model for the automatic simplification of legal texts in Spanish. The filtering pipeline itself also can be used as a general methodology that can be adapted for cleaning other large, heterogeneous corpora of legal texts and, possibly, with certain modifications in other domains.

References

ADAMCZYK, D., HULA, J., Efficient use of large language models for analysis of text corpora, *Proceedings of Recent Advances in NLP Applications*, 2024, 695-705. https://doi.org/10.5220/0012349800003654

ANH, D. H., DO, D.-T., TRAN, V., MINH, N. L., The impact of large language modeling on natural language processing in legal texts: A comprehensive survey, *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, 2023, 1-7. https://doi.org/10.1109/KSE59128.2023.10299488

ARFAT, Y., COLELLA, M., MARELLO, E., Legal text analysis using large language models, *Recent Advances in NLP and AI Applications* 2024, 258-268. https://doi.org/10.1007/978-3-031-70242-6_25

CEMRI, M., ÇUKUR, T., KOÇ, A., Unsupervised simplification of legal texts, *ArXiv, abs/2209.00557*, 2022. https://doi.org/10.48550/arXiv.2209.00557

GARCIA, E., SILVA, N., SIQUEIRA, F., GOMES, J., ALBUQUERQUE, H. O., SOUZA, E., LIMA, E., DE CARVALHO, A., RoBERTaLexPT: A Legal RoBERTa Model pretrained with deduplication for Portuguese, *Proceedings of the 16th International Conference on Computational Processing of Portuguese, 1*, 374-383, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics, 2024.

GUTIÉRREZ-FANDIÑO, A., ARMENGOL-ESTAPÉ, J., GONZÁLEZ-AGIRRE, A. VILLEGAS, M., Spanish legalese language model and corpora. *arXiv preprint arXiv:2110.12201*, 2021. https://doi.org/10.48550/arXiv.2110.12201

HUA, W., ZHANG, Y., CHEN, Z., LI, J., WEBER, M., LegalRelectra: Mixed-domain language modeling for long-range legal text comprehension. *ArXiv, abs/2212.08204*, 2022. https://doi.org/10.48550/arXiv.2212.08204

QUEVEDO, E., CERNÝ, T., RODRÍGUEZ, A., RIVAS, P., YERO, J., SOOKSATRA, K., ZHAKUBAYEV, A., TAIBI, D., Legal natural language processing from 2015 to 2022: A comprehensive systematic mapping study of advances and applications. *IEEE ACCESS,* 2024, **12**, 145286-145317. https://doi.org/10.1109/ACCESS.2023.3333946

SÁNCHEZ, D. B., ALDAMA GARCÍA, N., BARBERO JIMÉNEZ, Á., GUERRERO NIETO, M., MORALES, P. M., SERRANO SALAS, N., GARCÍA HERNÁN, C., HAYA COLL, P., MONTIEL PONSODA, E., CALLEJA IBÁÑEZ, P., MEL: Legal Spanish language model. *ArXiv, abs/2501.16011,* 2025. https://doi.org/10.48550/arXiv.2501.16011

SHAHEEN, Z., WOHLGENANNT, G., FILTZ, E., Large scale legal text classification using transformer models. *ArXiv, abs/2010.12871*, 2020. https://doi.org/10.48550/arXiv.2010.12871

SILVEIRA, R., PONTE, C., ALMEIDA, V., PINHEIRO, V., FURTADO, V., LegalBert-pt: A pretrained language model for the Brazilian Portuguese legal domain, *Advances in Legal Language Modeling*, 2023, 268282. https://doi.org/10.1007/978-3-031-45392-2_18

VALLEJO, R. G., Sobre la modernización del lenguaje jurídico: una mirada en España e Italia, *ELUA: ESTUDIOS DE LINGÜÍSTICA. UNIVERSIDAD DE ALICANTE*, 2021, **35**, 109-123. https://doi.org/10.14198/ELUA2021.35.6

Nikolai TIURIN is a PhD candidate in Natural Language Processing at the Autonomous University of Barcelona. His doctoral research focuses on automatic simplification of Spanish legal texts. His research interests include text simplification, statistical methods in NLP, and corpus linguistics.

Xavier BLANCO is a professor of French Philology at the Autonomous University of Barcelona, where he has taught Lexicology, Semantics, and History of Language. He is the author of numerous publications on lexicography and automatic lexical processing.