Assessment of Term Extraction Methodology Applied to Terminology Work. A Case Study of Catalan Legal Domain*

Mercè Vàzquez[1,*,†] and Patricia Morales-Hurtado[1,†]

[1] Universitat Oberta de Catalunya, Rambla del Poblenou 154-156 08018 Barcelona, Spain

Abstract

Computational terminology is one of the most productive lines of research in applied linguistics. One of its subspecialties is the automatic management of domain-specific vocabulary through automatic term extraction tools. These tools also introduce significant methodological changes to terminology work, enabling the analysis of larger corpora and the extraction of more multilingual terms. In this paper, we assess the terminology work process within the Catalan legal domain, using automatic term extraction methodology and a subsequent manual review of the extracted term candidates. The aim of the present research is to enhance the existing compilation of Catalan legal terminology by applying computational terminology methodology and publishing the results in an open access dictionary. The results show that term candidates can be extracted efficiently, and that the applied methodology significantly improves the manual revision process, thereby saving time and facilitating effective term candidate selection.

Keywords

Computational terminology, corpus linguistics, automatic term extraction, linguistic resources

1. Introduction

Computational terminology is one of the most productive lines of research in applied linguistics to automatic collection, management and analysis of terminology. This advancement is applied in broad domains in Natural Language Processing (NLP) such as information retrieval, question answering systems, ontology building, machine translation, summarising among others. And also the automatic management and analysis of terminological data is used for working on terminological metamodel work, database structure, data exchange or data fairification [1]. One of the computational terminology subspecialties involves the automatic management of domain-specific vocabulary using automatic term extraction (ATE) tools. Efficiently collecting terminology from corpora using ATE tools can increase and disseminate specialized knowledge across domains and language [2]. Likewise, ATE tools introduce significant methodological changes in terminology work—a field focused on the systematic collection, description, processing, and presentation of concepts and their designations [3]. These tools increase the number of corpora analyzed and expand the extraction of multilingual terms [4, 5, 6]. Moreover, this new approach promotes translations from international languages into minority languages, especially when using machine translation systems. For efficient translation of documents across domains and languages, these systems require a robust foundation of knowledge corpora and terminology [7, 8, 9]. Furthermore, ATE tools efficiently maximize terminology from less-resourced languages and facilitate the publication of results in open-access terminological dictionaries [10]. However, despite the significant advances introduced by computational terminology in recent years, the integration of this new methodology into terminology work and translation systems requires testing ATE tools across various domains and languages, as well as evaluating the results obtained. Moreover,

* 5th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2026, June 25-26, 2026, Zadar, Croatia.

* Corresponding author.

† These authors contributed equally.

mvazquezga@uoc.edu (M. Vàzquez); pmoraleshu@uoc.edu (P. Morales-Hurtado)

0000-0002-7983-4029 (M. Vàzquez); 0009-0009-2527-516X (P. Morales-Hurtado)

training terminologists, linguists, engineers, and translators during the process is essential for adapting standard procedures into a partially digitalized methodology for compiling corpora and processing candidate terms extracted by ATE tools. In this paper, we assess the terminology work process applied to the Catalan legal domain, using TBXTools with Token Slot Recognition (TSR) [11, 12] and a subsequent manual review of extracted term candidates. The primary aim of this research is to maximize the current Catalan legal terminology, with a special focus on European Union and international law regulations, by applying the computational terminology methodology. It also aims to publish the results in the open-access *Terminology of IATE in Catalan* e-dictionary [13], thereby completing the entries from the IATE (Interactive Terminology for Europe) database available in Catalan. The study explores how computational terminology contributes to improving terminology selection from understudied domains and languages that require maximized terminology compilation for specialized knowledge translation.

For its main objective, this paper describes the compilation of new terms from the European Union and international law domain in Catalan, English, and Spanish, using the aforementioned tool and filter. Additionally, it assesses the tool's efficiency in terminology work.

This paper is structured as follows: Section 2 describes the methodology of terminology extraction applied. Section 3 presents the results and discusses them in detail. The paper concludes with final remarks and ideas for future research.

2. Methodology

With the aim of maximizing Catalan legal terminology using computational terminology methodology, we built corpora of international and European Union law in Catalan, English, and Spanish to identify Catalan equivalents from IATE's database entries [14]. We publish the terms in the *Terminology of IATE in Catalan* e-dictionary.

To do so, we compiled specialized information publicly available in the Legal Portal of Catalonia [15] published by the Catalan Government, as it contains official translations of legal documents. It contains European Union and international law regulations, addressing issues related to human rights (civil, political, social, and cultural), founding treaties, and other European Union legislation. We also included information about international treaties from the Government of Andorra, as there are few linguistic differences between the Catalan language from Catalonia and that of Andorra [16]. Catalan and Andorran Language Policy Departments have previously collaborated to create official terminological resources, such as the TERMCAT resource [17, 18, 19]. Additionally, these subdomains were chosen because the terminology applies to multiple countries, meaning the same concepts exist in various languages. Following the structure of the Legal Portal of Catalonia, five corpora were created for each language: founding treaties, EU-derived legislation, universal human rights, European human rights, and other human rights. The 'other human rights' corpus included documents that were agreements and covenants between states, but not necessarily within an institution like the United Nations or the European Union. To summarize them for this study, we classified them by domain: European Union law (EU-derived legislation and European human rights), international law (universal human rights and other human rights), and international treaties (founding treaties). The number of tokens compiled in the corpora is shown in Table 1.

Table 1
Size of corpora sorted by domain and language

Domain of corpora	Tokens in Catalan	Tokens in Spanish	Tokens in English
European Union law	1,262,896	1,709,591	1,419,274

International law	291,079	345,394	302,473
International treaties	162,999	236,602	226,773
Total	1,716,974	2,291,587	1,948,520

After corpora compilation, we continued with the terminology extraction process to compile Catalan terminology and its corresponding terms in Spanish and English from the specific domain of international and European Union law. To do so, we used TBXTools, an open access terminology extraction tool, which employs linguistic and statistical methods for multiword term extraction [11] based on chi-square, t-score, log likelihood or PMI, among other measures. Additionally, it uses the Token Slot Recognition (TSR) method, a filtering approach based on a reference term list and term frequency within a corpus, to rank extracted term candidates from domain-specific corpora [12].

We automatically extracted candidate bigrams, trigrams, and quadrigrams (as most terminological units in the legal field are of this length) from Catalan, Spanish, and English corpora. We then filtered the results using a reference term list comprising terminology not present in the compiled corpora. The terms selected for the reference list came from the TERMCAT dictionary [13], which can be downloaded from the webpage [20] in different formats (namely, HTML, XML, and PDF). To extract high-quality candidates, we ensured our sources were up-to-date and endorsed by official institutions (TERMCAT and the EU, in this case). Using the TSR method, term candidates were ranked by their similarity to domain-specific corpora and by frequency. Similarity here refers to the degree of coincidence between terms in the candidate list and those in the reference term list. The method compares the two lists, considering the coincidence of words and their position within phrases. Term candidates with low similarity to the legal domain were given a low ranking and excluded from the manual revision process. Thus, the TSR method provides a more accurate and precise term candidate selection, which can improve the manual revision process conducted by specialists.

Once term candidates were extracted from the corpora, a manual revision process was undertaken to validate the terminology selection. The authors were responsible for this manual revision. We first considered asking domain experts to assist; nevertheless, we decided to perform the review ourselves for the following reasons. Spanish and English term candidates were selected only if they appeared in a downloadable version of IATE [14]. We refined the selection of entries by choosing International Relations (08), European Union (10), Law (12), and CJEU Law (14); all classifications IATE uses based on the EuroVoc thesaurus [21]. Catalan candidates were first checked against TERMCAT dictionaries [22] in the Law, Administration and Civil Protection, and Social Sciences sections. For candidates not found in these sources, our Catalan corpora consist of official translations from the Government of Catalonia. Therefore, terms extracted from them are considered the most used, reliable, and relevant in the legal field. Given the high number of candidates, we manually supervised 250 term candidates for each bigram, trigram, and quadrigram extraction. Thus, a total of 750 term candidates were manually supervised for each language and corpus to select terminology for the domain analysis.

3. Results and discussion

This section describes the results of using TBXTools for term extraction. To better exemplify the analysis, we focus on one corpus from our experiment (in this case, the Government of Andorra's treaties). While the TSR method was applied in all instances, we opted for this particular corpus because it was the most recently evaluated and incorporated the latest changes and updates from the term reference list. The statistics presented are exclusively from the mentioned documents, but the examples are drawn from all the corpora used.

The proposed filtered terms can be divided into three categories: full coincidence with the reference list, partial coincidence with the reference list, and no coincidence with the reference list. The partial coincidence and no coincidence categories can each be further separated into two subgroups: good and bad candidates. This classification is illustrated in Figure 1.

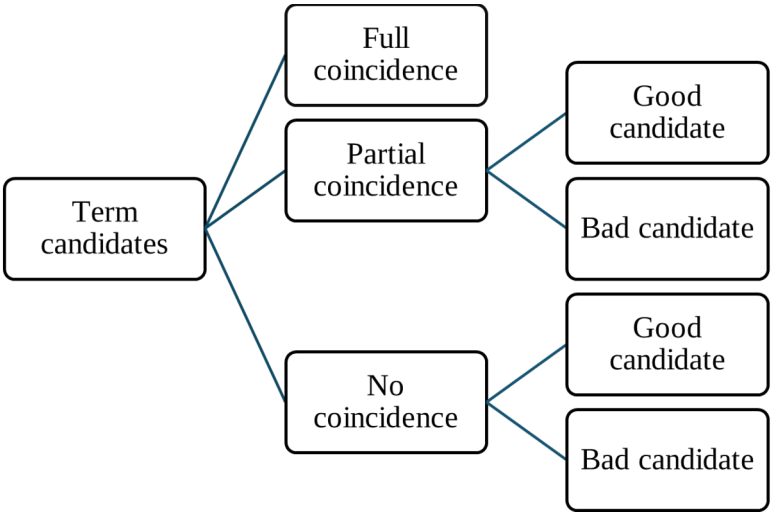

Figure 1: Classification of term candidates (source: authors of this article).

Firstly, terms with full coincidence are positioned at the top of the candidate list, as the reference list designates them as top priority (for example, *Nacions Unides* [United Nations], *persona física* [natural person], or *procediment legislatiu* [legislative procedure]). This prioritization allows for faster evaluation and elimination, enabling a focus on potential new terminology.

Secondly, partial coincidence provides relevant data for the research. While proving highly beneficial for identifying good candidates, (87 of the 141 good candidates found, or 61.7%, were attributable to the reference list), this category also elevates a significant number of non-viable candidates (303 of 496 bad candidates, or 61.09% of the total). For instance, *autoritat europea* [European authority] and *acte delegat* [delegated act] are considered new terms because the literal phrases are not in the reference list, but *autoritat* [authority] and *acte* [act] are present. Conversely, *informació sol·licitada* [requested information] or *mesures apropiades* [appropriate measures] are deemed bad candidates (as they do not define a specific concept within the legal field), even though *informació* [information] and *mesures* [measures] appear on the list.

Lastly, no coincidence terms are identified based on their frequency in the corpus. Two examples illustrate this category: *cap de missió* [head of mission] and *necessitats bàsiques internes corresponents* [corresponding basic internal necessities]. *Cap de missió* [head of mission] does not appear in the reference list, but after consulting the corresponding TERMCAT dictionary [19] it was confirmed to be a valid term. In contrast, *necessitats bàsiques internes corresponents* [corresponding basic internal necessities] does not delineate a specialized concept, despite its frequent use in the legal field.

After analysing the results, we confirmed that applying TBXTools with the TSR method expedites term identification compared to traditional approaches such as manual extraction. This

offers a benefit in workflow management, as users can simultaneously run the program and carry out other project tasks. Moreover, the reference term list improves the quality and reliability of term candidates by prioritizing those with full or partial coincidence, making them more noticeable to the reviewer. Users can also update the reference term list with new terms and fine-tune subsequent extractions with greater precision. In contrast, we need to consider the hardware being used because older or less powerful computers may experience slower processing analysis, and the filtering method requires more time to complete the task.

As a result of the manual review, we identified 341 terms in Catalan, 255 terms in Spanish, and 262 terms in English. Table 2 illustrates these statistics sorted by n-grams and language. These terms will expand the content of the *Terminology of IATE in Catalan* e-dictionary. In addition, the new linguistic resources created within the framework of this research, including machine translation and automatic terminology extraction tools, will serve as a basis for evaluating how European and international regulations can be promoted in Catalan.

Table 2
Identified terms sorted by n-grams and language

N-grams	Catalan	Spanish	English
Bigrams	170	122	138
Trigrams	121	100	79
Quadrigrams	50	33	45
Total	341	255	262

4. Conclusion and future work

Computational terminology methodology introduces relevant advancements in terminology work by processing corpora using automatic term extraction tools. The amount of terminological data analyzed with this methodology facilitates the compilation of term units in different languages and domains, which is particularly relevant for less-resourced languages.

The present research confirms that efficient terminology collection from corpora, using automatic term extraction tools, enables the creation of specialized terminological databases and dictionaries across different domains. Indeed, these linguistic resources are invaluable for translators, who often dedicate considerable time to finding the most precise terms for specialized domains.

The term extraction method applied to terminology work done within the Catalan legal domain confirms that term candidates can be extracted efficiently. In our case, we used TBXTools with a TSR filter to statistically extract term candidates from Catalan, Spanish, and English corpora in the fields of European Union and international law. This method introduces improvements during manual revisions of filtered term candidates, saving time and reducing the effort involved in term candidate selection. Furthermore, it is a reliable method for less-resourced languages.

As future research, we plan to expand the selection of terminological units in other IATE domains and also enhance term extraction methods to improve the selection of term candidates

from specialized corpora. We also encourage researchers and scholars to study less-resourced languages in natural language processing to represent as many languages as possible. Computational linguistics can benefit immensely from language diversity, and it democratizes knowledge within the scientific community.

Acknowledgements

This study was supported by the Industrial Doctorates Plan from the Department of Research and Universities of the Government of Catalonia (reference number: 2025 DI 00112); the project Computational terminology applied to the Catalan Legal Domain (N50000000RA505), funded by UOC University in the Research Accelerator call 2024, and the project TamTAS (PCI2025-167063-2), funded by MICIU/AEI/10.13039/501100011033 and European Union in the CHIST-ERA call 2025 Science in your own language.

Declaration on Generative AI

During the preparation of this work, the authors used LanguageTool in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] E. Lefever, A. R. Terryn, Computational Terminology, in: Y. Peng, H. Huang, D. Li (Eds.), New Advances in Translation Technology. New Frontiers in Translation Studies, volume Part F3024, Springer, Singapur, Singapur, 2024, pp. 141–159. doi:10.1007/978-981-97-2958-6_8.

[2] K. Ahmad, M. Rogers, 8.4.1 Corpus linguistics and terminology extraction, in: S. E. Wright, G. Budin (Eds.), Handbook of Terminology Management. Volume 2: Application-Oriented Terminology Management, John Benjamins Publishing Company, 2001, pp. 725–760. doi:10.1075/z.htm2.28ahm.

[3] International Organization for Standardization, ISO 1087:2019. Terminology work and terminology science — Vocabulary, International Organization for Standardization, 2019. URL: https://www.iso.org/en/contents/data/standard/06/23/62330.html.

[4] K.-D. Schmitz, Developments in computational terminology management and its influence on terminology science and terminology work, in: Papers of the 5th conference: "Hellenic language and terminology," volume 6 of Hellenic language and terminology 2005, University of Cyprus, Nicosia, Cyprus, 2005. URL: http://www.eleto.gr/download/Conferences/5th%20Conference/5th_24-11-Schmitz%20Klaus-Dirk_Paper.pdf.

[5] A. Nuopponen, Dimensions of Terminology work, Terminologija 25 (2018) 6–22. URL: http://lki.lt/wp-content/uploads/2018/12/Terrminologija_25_maketas.pdf.

[6] T. Wissik, Impact of automatic term extraction on terminology work: A qualitative interview study in institutional settings, Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 31 (2025) 110–135. doi:10.1075/term.00085.wis.

[7] G. Dinu, P. Mathur, M. Federico, Y. Al-Onaizan, Training neural machine translation to apply terminology constraints, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3063–3068. doi:10.18653/v1/P19-1294.

[8] M. Exel, B. Buschbeck, L. Brandt, S. Doneva, Terminology-constrained neural machine translation at SAP, in: A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, M. L. Forcada (Eds.), Proceedings of the 22nd Annual Conference of the European Association for

Machine Translation, European Association for Machine Translation, Lisboa, Portugal, 2020, pp. 271–280. URL: https://aclanthology.org/2020.eamt-1.29/.
[9] E. Michon, J. Crego, J. Senellart, Integrating domain terminology into neural machine translation, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 3925–3937. doi:10.18653/v1/2020.coling-main.348.
[10] M. Vàzquez, A. Oliver, E. Casademont, Using open data to create the Catalan IATE e-dictionary, Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 25 (2019) 175–197. doi:10.1075/term.00035.vaz.
[11] A. Oliver, M. Vàzquez, TBXTools: A Free, fast and flexible tool for Automatic Terminology Extraction, in: R. Mitkov, G. Angelova, K. Bontcheva (Eds.), Proceedings of the International Conference Recent Advances in Natural Language Processing, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 2015, pp. 473–479. URL: https://aclanthology.org/R15-1062/.
[12] M. Vàzquez, A. Oliver, Improving term candidates selection using terminological tokens, Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 24 (2018) 122–147. doi:10.1075/term.00016.vaz.
[13] TERMCAT, Centre de Terminologia, Universitat Oberta de Catalunya (UOC), Terminologia de IATE en català, 2019. URL: https://www.termcat.cat/ca/diccionaris-en-linia/264/.
[14] Translation Centre for the Bodies of the European Union, IATE, 2018. URL: https://iate.europa.eu/home.
[15] Generalitat de Catalunya, Portal Jurídic de Catalunya, 2025. URL: http://portaljuridic.gencat.cat/ca/inici/.
[16] E. Valls i Alecha, Canvi morfològic vs. canvi fonològic en català nord-occidental, Treballs de sociolingüística catalana 23 (2013) 57–79. doi:10.2436/20.2504.01.51.
[17] Agència Catalana de Turisme, TERMCAT, Centre de Terminologia, Diccionari de turisme, 2023. URL: https://www.termcat.cat/ca/diccionaris-en-linia/312/.
[18] TERMCAT, Centre de Terminologia, Diccionari de cooperació al desenvolupament, 2019. URL: https://www.termcat.cat/ca/diccionaris-en-linia/190/.
[19] TERMCAT, Centre de Terminologia, Diccionari de relacions internacionals, 2018. URL: https://www.termcat.cat/ca/diccionaris-en-linia/246/.
[20] TERMCAT, Centre de Terminologia, Terminologia Oberta, 2026. URL: https://www.termcat.cat/ca/terminologia-oberta.
[21] Translation Centre for the Bodies of the European Union, IATE-FAQ, 2026. URL: https://iate.europa.eu/faq.
[22] TERMCAT, Centre de Terminologia, Diccionaris en Línia, 2026. URL: https://www.termcat.cat/ca/diccionaris-en-linia.