Creating Terminological Resources in the Digital Age for Less-resourced Languages

Mercè Vàzquez
Universitat Oberta de Catalunya
Barcelona, Spain
mvazquezga@uoc.edu

Abstract

Multilingual terminological resources contain the most representative knowledge of specialized domains and allow professionals to create and translate specialized content in order to spread knowledge. Today, representative and useful multilingual terminological resources are available for the most resourced languages. This reduces or limits the development of knowledge in less-resourced languages across different specialized domains, mainly those that are constantly evolving and creating or adapting new concepts as needed. In this paper we present our methodology for carrying out terminological projects in Catalan, based entirely on open access linguistic resources and using natural language processing tools. The main objective of this research is to maximize the Catalan terminology currently available in open access, using a combination of natural language processing tools. The results are supervised by linguists and terminologist experts before being publicly available to the public. The findings of our research provide a new approach to terminology work, making it possible to design high-volume multilingual terminological projects that are manually revised by linguists and terminologists in the context of less-resourced languages.

Keywords: computational terminology, linguistic resources, terminology, less-resourced languages

1. Introduction

Satisfying the need for terminology in all fields of knowledge is one of the greatest challenges for less-resourced languages, as the reduced number of terminological resources makes it more difficult to access existing specialized knowledge, to create new specialized concepts (especially in the most innovative fields) and to translate specialized knowledge into these languages. Less-resourced languages suffer from a lack of available linguistic resources, such as linguistic corpora, terminology dictionaries and technologies that help to maximize the resources available. Because of this limitation, there is a dearth of reliable regulatory documents in specialized fields, which would be necessary to establish the relevant terminology. It also means having to rely on documents published in other languages and translating them, or using bridge languages to create equivalent terms in the less-resourced languages.

As shown by the results of a survey on lexicographic practices and lexicographers' needs across Europe (Tiberius et al., 2022), more natural language processing resources are needed for less-resourced languages, and their availability needs to be firmly established. According to the authors, there is also a need to integrate lexicographic data into natural language processing applications, and terminological resources into dictionaries published on open access websites.

Against this backdrop, the design of terminology resources in the digital age should take into account the range of open access linguistic resources that can be used to overcome the limitations suffered by less-resourced languages and to meet users' needs for access to terminological information. As observed in the course of our research, the considerable number of open access linguistic resources scattered across the internet have become crucial in accessing available multilingual terminological resources, including those for less-resourced languages such as Catalan. Moreover, the use of open access linguistic resources makes it possible to create terminological resource products according to specific requirements, ensuring that the content can be linked to external information and made available on open access platforms (Heid, 2014).

The main objective of this research is to maximize the Catalan terminology currently available using natural language processing tools and open access linguistic resources, and to publish the results revised by linguists and terminologists in an open access format. To this end, our research focuses on the European Union's interinstitutional terminology database –Interactive Terminology for Europe (IATE) – in order to compile the Catalan equivalents of terms in IATE's majority languages (English, French and Spanish). As the largest multilingual terminology database, IATE provides a model for compiling terminology equivalents in Catalan. Thus, our approach is to first compile the largest possible number of terms in Catalan using open access linguistic resources and to match them with their English, French and Spanish equivalents, as well as the IATE code corresponding to each entry, in order to obtain terminological information from the 24 official languages of the European Union.

The methodology used confirms that it is possible to increase the availability of terminology resources in Catalan in the three chosen specialized domains (law, economics and medicine) and to ensure that the results of such projects are made publicly available. The use of open access linguistic resources in terminology projects entails major changes in terms of project design, the selection of source documents, data processing, linguistic reviews to ensure data quality control and, finally, the delivery of publicly available terminology resources.

The paper is structured as follows: Section 2 presents our methodology, Section 3 describes the results and Section 4 provides some concluding remarks and ideas for future research.

2. Methodology

In order to overcome the initial limitations in terms of specialized knowledge and terminology available in Catalan, we chose the IATE terminology database to identify the corresponding equivalents in Catalan. Since our main objective was to compile high-quality terminology entries, we selected the IATE entries with the highest reliability. Our method for compiling high-volume terminological resources involves the use of open access knowledge and natural language processing tools, combined with a manual review of the compiled data, as described below.

First, we selected three specialized domains (law, economics and medicine) from the IATE terminology database in order to retrieve their open access terminology in three languages: English, French and Spanish. Having classified this terminology by domain, we compiled the open access terminology available in Catalan related to these domains. We focused our research on Terminologia Oberta (TO),[1] published by the TERMCAT Terminology Centre, which contains many terms that are available to the public and can be downloaded by specific domains. The terminology in this resource allowed us to find translation equivalents in Catalan, along with additional information about the terms, such as their grammatical category, definition, notes and reference source.

We also used Wikipedia,[2] the free online encyclopaedia, as a resource from which reliable terminology can be extracted (Oliver et al., 2017). From Wikipedia, we extracted the terminology contained in the *Categories* section located at the bottom of encyclopaedia entries, which is complementary to the information provided in the main body of text. This terminological content was extracted within the DBpedia project (Lehmann et al., 2015), which provides access to the encyclopaedia's content in a database structure that is easier to process computationally than direct dumps from Wikipedia.

We then selected specialized knowledge related to the fields of law, economics and medicine to build the corresponding linguistic corpus and increase the Catalan terminology compilation. We used the Official Journal of the Government of Catalonia (DOGC) and created the corresponding corpus (Oliver, 2017) in Catalan and Spanish. The creation of a multilingual linguistic corpus based on information published in specialized domains from authoritative information sources enables the identification of a large number of terminological units and their context of use, and also ensures the quality of the data collection.

Having created the linguistic corpus, we used a terminology extraction tool to automatically identify the most representative terms from this corpus. To perform this task, we used the automatic terminology extraction tool TBXTools (Oliver and Vàzquez, 2015), which is an open source tool that implements linguistic and statistical methods for multiword term extraction, and which is gradually being developed with new filtering methods to improve the results of terminology extraction from specialized corpora (Vàzquez and Oliver, 2018).

The automatic identification of terminology from the linguistic corpus provided a set of candidate terms that had to be manually reviewed by experts before being selected as terms of a specific domain. The candidate terms automatically extracted from the corpus using TBXTools were manually reviewed by linguists and terminologists to ensure the quality of the data compiled for each domain.

Since the main objective of our research was to identify the Catalan equivalents of English, French and Spanish terms downloaded from IATE, we also created a parallel corpus. The identification of translation equivalents in large corpora, such as the DOGC corpus, is a very costly process in terms of computational processing, so an efficient indexing algorithm for identifying translation equivalents based on statistical machine translation models computed with Moses (Koehn, 2007) was run in TBXTools. This feature makes it possible to search for translation equivalents from any available parallel corpus, especially in the case of large corpus compilations.

Once the Spanish and English equivalents of the Catalan terms were found in the parallel corpus, before publishing the results by domain, we created a preliminary database containing the terminological information from IATE (English, Spanish and French terms) together with the Catalan equivalents and their grammatical category, definition (if available) and the source of information. In order to ensure that the meaning of the IATE terms matched the Catalan equivalents, we also added the IATE ID to obtain all the information about the IATE entries that was available at the time and that would be useful for future updates. During the manual review, we were able to verify that the IATE terms and specialized domains were correctly matched with their Catalan equivalents by means of the IATE ID, with a group of linguists and terminologists ensuring the fit of each concept (terms and equivalents) in the particular domain. At the end of the manual review process, we accepted the Catalan equivalents when the meaning was consistent with the IATE reference terms (English, Spanish and French) and domains (Table 1) and rejected those that were not reliable (Table 2).

[1] https://www.termcat.cat/en/terminologia-oberta

[2] https://en.wikipedia.org/wiki/Outline_of_academic_disciplines

Language	IATE ID 34774	IATE ID 1091562
English term	rights in rem	succession property
Spanish term	derechos reales	bienes sucesorios
French term	droits réels	biens successoraux
Catalan term	drets reals	actiu hereditari
Source	DOGC	TO

Table 1: Review results of correct terms (law)

Language	IATE ID 3584038	IATE ID 113058
English term	exit from the territory	contract for services
Spanish term	salida del territorio	arrendamiento de servicios
French term	*not available*	*not available*
Catalan term	sortida del territori	contracte de servei
Source	DOGC	TO

Table 2: Review results of incorrect terms (law)

The Catalan equivalents were compiled from authoritative information sources so that this terminology could be made freely available to end users (translators, specialists, etc.). To this end, the manual review of the equivalents retrieved using natural language processing tools was carried out by linguists and terminologists, following a standard process to ensure the quality of the results.

After the manual review of the preliminary results, the validated Catalan equivalents were ready to be published in an open access dictionary related to IATE.

The use of openly available online linguistic resources for terminology projects makes it possible to overcome data dispersion, downloading and access problems; to concentrate terminological material in a single terminological output (Tarp, 2012), and to provide less-resourced languages such as Catalan with a large set of terminological resources.

3. Results

The most significant result of the research carried out using this methodology was the creation of an open access digital terminology dictionary[3] with 16,231 entries in Catalan, Spanish, English and French, which complements the content of the European Union's IATE terminology database. The entries in this dictionary also include the corresponding IATE code to allow consultation in the other languages of the European Union.

Broken down by domain, the dictionary contains 3,292 entries in law, 478 entries in economics and 3,829 entries in medicine. It also includes the compiled results in the European Union domain, with 632 entries, and those in various other domains, with 8,000 entries from the IATE database. The precision achieved in the compilation of these results is shown in Table 3.

Domain	TO precision	Wikipedia precision	DOGC precision
Law	82.08%	61.71%	95.37%
Economics	74.76%	52.78%	
Medicine	87.12%	69.54%	
European Union	77.60%		
Other	63.16%		

Table 3. Results in terms of precision

This is an innovative research result, as the methodology had never before been used to produce a high-volume terminology dictionary with a multidisciplinary team of terminologists, linguists, translators and engineers. The large amount of data collected was intended to test the ability of natural language processing tools to provide new linguistic resources for less-resourced languages. It is also innovative in that the natural language processing tools we have developed are freely available on the GitHub platform, which hosts open source tools.

Furthermore, the manual review led to some changes in the terminology work process. Linguists and terminologists need to understand the value of the candidate terms extracted from the automatic terminology extraction tool in order to select those in each domain. They also need to determine which terms compiled from different linguistic resources, such as Terminologia Oberta or Wikipedia, are reliable for each domain, and to find the most reliable Catalan equivalents between different proposals automatically extracted from parallel corpora. Finally, they need to be able to manage automatic data extracted from natural language tools in order to ensure the quality of the data compiled in the resulting online dictionary.

4. Conclusions and Future Research

Open access linguistic resources can be used by less-resourced languages to create a wide variety of terminological resources that can be made available to end users.

The design of new terminology projects based on open data also enriches terminology outputs with additional information, such as the reference source of each equivalent and links to external sources,

[3] https://www.termcat.cat/en/diccionaris-en-linia/264/presentacio/en

which can be adapted to the needs of the project's target users.

This research introduced a new way of working with terminology, in which linguists and terminologists manually oversee a large amount of automated data extracted from natural language processing tools to build an online dictionary.

The approach presented is not specific to Catalan, so it can be used for other languages to increase linguistic resources in an open access format. A final advantage of this new approach to open access projects is the ability to integrate data from different information sources and to unify resources published in different formats, making it easier for users to consult the resource.

As for future research, we are considering exploring new types of linguistic resources that can be integrated into the design of terminological resources, and developing new tools to automatically update the content of terminological projects using the methodology presented in this paper.

5. Bibliographical References

Heid, Ulrich (2014). Natural Language Processing Techniques for Improved User-friendliness of Electronic Dictionaries. In *Proceedings of the XVI Euralex International Congress: The User in Focus*, pages 47–62. Bolzano/Bozen, Italy.

Koehn, Philipp; Hoang, Hieu; Birch, Alexandra; Callison-Burch, Chris; Federico, Marcello; Bertoldi, Nicola; Cowan, Brooke; Shen, Wade; Moran, Christine; Zens, Richard; Dyer, Chris; Bojar, Ondrej; Constantin, Alexandra; Herbst, Evan (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Prague, Czech Republic.

Lehmann, Jens, Isele, Robert; Jakob, Max; Jentzsch, Anja; Kontokostas, Dimitris; Mendes, Pablo N.; Hellmann, Sebastian; Morsey, Mohamed; van Kleef, Patrick; Auer, Sören; Bizer, Christian (2015). DBpedia–a Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web,* 6(2):167–195.

Oliver, Antoni; Vàzquez. Mercè (2015). TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction. In *International Conference on Recent Advances in Natural Language Processing* (RANLP 2015), pages 473–479. Hissar, Bulgaria.

Oliver, Antoni; Vàzquez, Mercè; Ubide, Georgina. (2017). Estudi de la fiabilitat de la Viquipèdia com a recurs terminològic. *Revista Tradumàtica. Tecnologies de la Traducció*, 15, 10–20.

Oliver, Antoni (2017). El corpus paral· lel del Diari Oficial de la Generalitat de Catalunya: compilació, anàlisi i exemples d'ús. *Zeitschrift für Katalanistik*, 30, 269–291.

Tarp, Sven (2012). Online Dictionaries: Today and Tomorrow. *Lexicographica: International Annual for Lexicography*, 28(1):253–268.

TERMCAT, Centre de Terminologia (2019). Terminologia oberta. Barcelona: TERMCAT, Centre de Terminologia. https://www.termcat.cat/en/terminologia-oberta

Tiberius, Carole; Kallas, Jelena; Koeva, Svetla; Langemets, Margit; Kosem, Iztok (2022). An insight into lexicographic practices in Europe. *Dictionaries and Society*. Proceedings of eLex 2022, pages 509–51.

Vàzquez, Mercè; Oliver, Antoni. (2018). Improving term candidates selection using terminological tokens. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *24*(1):122–147.