Preprint of the published version in: Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, Volume 25, Issue 2, Nov 2019, p. 175 – 197

DOI: https://doi.org/10.1075/term.00035.vaz

Using open data to create the Catalan IATE e-dictionary

Mercè Vàzquez, Antoni Oliver
Universitat Oberta de Catalunya |

Elisabeth Casademont
TERMCAT

Abstract

Linguistic resources currently available to the public in the form of open data are an important repository for user consultations and an essential source of information for creating e-dictionaries. However, access to these linguistic resources is still limited because the information is dispersed over different sources and in different formats and is not available in all languages, thereby hindering consultation and automatic recovery. This paper presents a method for maximising use of open access linguistic resources and integrating them into specialised e-dictionaries. The method combines automatic compilation of terminology data with the creation of specialised linguistic corpora to produce a Catalan version of the IATE (InterActive Terminology for Europe) database. The paper presents a new methodological advances applied here to the production of terminological e-dictionaries, using open access linguistic resources. We observe that, for the first time, this new methodology enables economics, law and health dictionaries corresponding to the Catalan versions of the IATE to be created. In conclusion, the new methodology presented here permits the creation of new models of specialised e-dictionaries, facilitates the compilation and consultation of terminology in any language and unifies the access format for terminology data. Future studies will complete the definition and integration of open access linguistic resources that can be included in our methodology.

Keywords:
e-dictionaries, lexicography, open data, terminological dictionaries, terminology, natural language processing

1. Introduction

Dictionaries are ‘a set of basic linguistic units set out in a particular order and providing certain information’ (TERMCAT, 2010) and also they can vary widely in their characteristics, function and size. Overall, dictionaries can be classified into three main groups: general language dictionaries, encyclopaedic dictionaries and terminological dictionaries. General language dictionaries contain words considered to be of general use in a language, encyclopaedic dictionaries contain current knowledge on one or more subjects, based on a structure of key words; and terminological dictionaries combine the characteristics of the other two types, as they are both onomasiological (like encyclopaedic dictionaries) and characterise concepts (like general language dictionaries), as explained in TERMCAT, 2010. Each type of dictionary has certain specific characteristics, as shown in Table 1.

Table 1. Types of dictionary

Types of dictionary	**General language dictionaries**		**Encyclopaedic dictionaries**	**Terminological dictionaries**	
Specific characteristics of each type of dictionary	Alphabetic dictionary	Ideological dictionary	Alphabetic dictionary	Alphabetic dictionary	Ideological dictionary
	Monolingual dictionary	Multilingual dictionary	Monolingual dictionary	Multilingual dictionary	
	Prescriptive dictionary	Descriptive dictionary	Prescriptive dictionary	Prescriptive dictionary	
	Synchronic dictionary	Diachronic dictionary: historical or etymological dictionary	Diachronic dictionary	Synchronic dictionary	
	Textual dictionary	Visual dictionary		Textual dictionary	Visual dictionary

Today, a large and diverse number of linguistic resources are available in open access that facilitate the production of dictionaries, and terminological dictionaries in particular. These linguistic resources provide detailed information on words and terms in a text for use in producing terminological dictionaries, which can thus offer prescriptive content offering basic knowledge of a subject to the general public. As Tarp (2012) explains, the interaction between e-dictionaries and external sources of information, such as the Internet and corpora, permit information to be imported for inclusion in dictionary entries. However, open access linguistic resources are awkward to consult, as they are spread over different databases, their data are difficult to download and use, they are provided in different formats and are not usually available in all languages.

This paper presents a method for maximising use of open access linguistic resources and integrating them into specialised e-dictionaries. The method combines automatic compilation of terminology available from different sources of information, such as terminology dictionaries, manuals and encyclopaedias, with the creation of specialised linguistic corpora to thereby produce a Catalan version of the IATE (InterActive Terminology for Europe) database~~.~~

In the paper we present a method for creating a Catalan version of the IATE for the subject domains of economics, law and health. This was done using different open access linguistic resources. First of all, we used terminology dictionaries available in Catalan, Spanish or English corresponding to the IATE subject domains of economics, law and health. These dictionaries are part of the Terminologia Oberta (Open Terminology) service in TERMCAT, the centre for terminology in Catalan, and Wikipedia, the free encyclopaedia. Secondly, we worked with parallel corpora in Catalan and Spanish for the subject domains of law and administration, which were created mainly from the *Diari Oficial de la Generalitat de Catalunya* (Official Journal of the Government of Catalonia). Given the large volume of parallel corpora, we used translation models based on statistical machine translation systems to optimise the term search process.

The main objective of our study is to build a new methodology in producing terminological e-dictionaries based on the use of open access linguistic resources and which integrate the terminology available from different sources of information and in different formats. This main objective is based on two hypotheses. The first is that identifying terminology from different open access linguistic resources facilitates the presence of any given language in specialised e-dictionaries. The second hypothesis is that using linguistic resources makes it possible to construct dictionaries more efficiently and rapidly and also include information that is currently spread over different reference sources.

To achieve the main objective and test the two working hypotheses, this paper describes the application of a method to create terminological dictionaries in Catalan, using different types of open access linguistic resources. This method is applied to the creation of specialised dictionaries in Catalan in the subject domains of economics, law and health, using terminology in the IATE database along with terminology dictionaries and parallel corpora.

The paper is structured into the following sections: the second section describes the state of the art with regard to the design and production of specialised e-dictionaries; the third section presents the open access linguistic resources used to produce the terminological e-dictionaries; the fourth section explains the working method for constructing the Catalan version of IATE for several subjects, the fifth and sixth sections presents the results obtained in the production of each dictionary and discusses the results; the seventh section describes the publication of terminological e-dictionaries; finally, the last section presents the conclusions and future lines of work.

2. Background

Currently, e-lexicography is undergoing a process of transformation, thanks to possibilities in information management provided by online resources ranging from corpora to lexicographical data, together with new programming tools and strategies. This ground-breaking change can be seen in the diversity of recently published studies on e-lexicography (Pastor and Alcina, 2010; Trap-Jensen, 2010; Rundell and Kilgarriff, 2011; Fuertes-Olivera and Bergenholtz, 2011; Tarp, 2012; Heid et al., 2012; Granger and Paquot, 2012; Abel and Meyer, 2013; L'Homme and Cormier, 2014; Declerck et al., 2015; Ball, 2016).

This transformation in lexicography has arisen because the printed dictionary, defined as 'a book in which a systematic representation is given of one or more aspects of (parts of) the vocabulary of one or more languages, using two dimensions: the macro- and microstructure' (Martin and Van der Vliet, 2003), has evolved into a dictionary published in electronic format, which 'formalises information through digital resources and differs from dictionaries published on paper in their use, presentation of data, search possibilities and technical aspects' (Gelpí, 2003). Furthermore, in recent years, the perception of the traditional dictionary has radically changed. It is no longer considered a 'set of structured texts from which a user may be able to extract textual data that allow him, by means of an interactive interpretation, to derive information. Another really distinctive feature of dictionaries and other lexicographical tools is that they provide quick and easy access to the specific types of data from which a specific type of user can retrieve information that may cover their specific types of needs', as stated by Tarp (2008). Granger (2010), in turn, points out that 'recent research has illustrated the clear necessity to adapt dictionaries to users' needs and technological advances have simultaneously put this development within the reach of dictionary producers'. Similarly, Tarp (2009) explains that 'lexicographic needs are not abstract needs, but are always related to specific types of users who find themselves in a specific type of social situation'. Multiple studies on users' needs have been published in recent years as a way to improve dictionary consultation and better understand the relationship between types of dictionary, types of dictionary user and types of dictionary use (Nessi, 2013). Varantola (2002), for instance, claims that dictionary users can be classified into three broad categories: language learners, non-professional users, and professional users—who 'normally use a dictionary to perform a task that they get paid for—, but other user variables can also affect their behaviour as age, mother tongue, second or foreign language, language proficiency level, educational level, level of skill in dictionary use, role (as a teacher, learner, translator, traveller, player of word games etc.), location (geographically, and within the home, place of work or educational institution)'. This highly significant change in the concept of lexicographical work, thanks to a change in technology, has led the discipline to explore new ways of organising dictionary entries and setting out information, incorporating external information in lexicographical works and adapting dictionaries to different users' needs and profiles.

The evolution from printed dictionary to e-dictionary can be seen in how lexical data are accessed (Lew, 2013). As De Schryver (2012) states, users have switched from 'looking up' to 'searching'. Indeed, e-dictionaries provide various search strategies (Engelberg and Lemnitzer, 2009), such as incremental, wildcard, Boolean, filter and external text searches, among others. Bearing in mind the variety of strategies, Pastor and Alcina (2010) establish a classification of search techniques used by e-dictionaries to improve and facilitate the work of lexicographers and terminologists when creating standardised dictionaries.

Furthermore, when creating e-dictionaries, lexicographers use strategies that differ from those applied to printed dictionaries when distributing the information in entries. Thus, the dynamic nature of e-dictionaries allows lexicographers to present dictionary data more flexibly. Furthermore, entries can be structured in a variety of ways and recovered from a single lemma, contained in a single database (Gouws, 2014).

Along the same lines, Tarp (2008) considers that the process of producing an e-dictionary has developed from more specific to more general, providing a set of lexical data that can be used for both e-dictionaries and printed dictionaries. Thus, in his opinion, natural language processing tools and techniques enable the lexical data to be used over and over again. By doing this it facilitates the reuse of data for the various potential uses that can be given to dictionaries. Furthermore, natural language processing improves the usability of dictionaries, improving guidance within the text and access to data (Heid, 2014). Considering these statements, the use of open access lexical data combined with natural language processing should be the basis on which to build e-dictionaries, as our research shows through its applied methodology.

When producing terminological dictionaries, one must bear in mind how they differ from general language dictionaries, encyclopaedias or handbooks, and thus apply the requirements for such works. As specified in TERMCAT (2010), terminological dictionaries have certain characteristics in common with general language dictionaries, as they characterise concepts with the traits necessary to distinguish them from other concepts, and encyclopaedic dictionaries, by using an onomasiological approach. However, unlike these two types of works, terminological dictionaries do not have general language entries, as they contain only specialised language from a single field. Furthermore, terminological dictionaries are generally prescriptive, i.e. they fix the use of the terms and concepts contained in them and usually include equivalents in other languages. They aim to offer plenty of help for users in understanding terms in a text, by providing detailed and summarised information. In addition, they display the information alphabetically, which facilitates consultation and guarantees that the definitions of a word are valid for a broad range of concepts. In short, terminological dictionaries provide an approach to the basic knowledge of a given subject through an initial field map (concept map), subject indices, terms related to other definitions and notes.

With regard to the development of terminological dictionaries, TERMCAT published the first e-dictionary in 2004 and now all the centre's terminological compilations are published online. According to Serra (2014), this change is due to the importance of disseminating terminological work. TERMCAT's policy of terminological dissemination focuses on making full use of technological tools and online options. Technology facilitates wide dissemination, while also being free and easily proliferated. There are three advantages of e-dictionaries: low cost of publishing, immediacy in updating content and free, universal access. Furthermore, the TERMCAT online dictionaries can be downloaded from the Terminologia Oberta section in different formats with Creative Commons licences. The changes in the production of terminological dictionaries by this terminology centre are similar to those in all centres, institutions and organisations and that produce specialised lexicographic works.

Another example of the creation of terminological e-dictionaries using a freely accessible online corpus, similarly to Catalan IATE e-dictionary, is the TERMIS project (an applied research project titled 'Terminology databanks as the bodies of knowledge: The model for the systematisation of terminologies'), whose aim is to make Slovenian terminology accessible in digital format As noted in Logar and Kosem (2013), 'TERMIS is the development of a freely accessible online dictionary-like terminology database for the discipline of public relations. The development of an online dictionary editing system that is easy to use so that an expert in the field, i.e. a terminologist, can start using it without any prior knowledge.'

With regard to open access information used in terminology work, there are some relevant projects, such as the European Union terminology offered by TermCoord through a public website and free tools (Maslias, 2014); the TTC project (Terminology Extraction, Translation Tools and Comparable Corpora), which focuses on the development of tools that automatically generate bilingual terminologies based on monolingual text data collected from the web (Blancafort, 2013); and the IUPAC project, an open access terminological wiki on chemometrics (Hibbert, 2008). This research projects show that open access lexical resources are required to build e-dictionaries in terms of reusing information and facilitating dictionaries creation in terms of time consuming and human resources.

Thus, according to L'Homme (2014), technology is within reach of lexicographers for a wide variety of tasks. Tools that permit compilation of data from corpora offer a number of functions, such as extracting collocations, identifying patterns and locating semantic relations. The use of corpora together with improvements to processing tools represent a substantial change in lexicography. This method provides different visions of textual data and offers lexicographers linguistic evidence when making decisions on what data to include in dictionaries. Furthermore, lexical databases in structured formats, such as XML, permit the information they contain to be automatically recovered and efficiently relocated or linked to new terminological projects.

3. Materials

With the aim of compiling terminology in Catalan we experimentally implemented a method that permits the identification and extraction of terminology in different specialist domains, available in various languages from a variety of sources. The linguistic resources used to create the terminological dictionaries described in this paper and discussed below are open access reference resources for terminologists and translators. These resources fall into two groups: 1) Terminological resources, as Terminologia Oberta from TERMCAT. We have also experimented with the use of Wikipedia as a terminological resource. 2) Parallel corpora, we have used the DOGC corpus, a Catalan Spanish parallel corpus created from laws of the Catalan government.

3.1. IATE database

IATE is the European Union interinstitutional terminology database. It is one of the most important terminology resources, as it provides public access to a large amount of terminology data in the 24 official languages of the European Union, along with Latin. However, it does not include Catalan. The most frequent means of consulting the resource is via a web form with a variety of search criteria (language, subject domain, etc.). It can also be downloaded as a large TBX (TermBase eXchange) file and a small program can also be downloaded to create subsets of the TBX file and then select specific languages and specialities, thus providing a smaller file which works with most computer-assisted translation tools, such as OmegaT, SLD-Trados, Memsource and the vast majority of tools, as TBX is a standard language. Table 2 shows the number of terms in English and Spanish in the subject domains of law, economics and health and the number of terms for all domains included in the IATE. IATE terms that are full form and have reliability rates equal to or above 3, i.e. reliable or very reliable, are also shown.

Table 2. Terms in English and Spanish in the IATE's law, economics and health subject domains (v. 30/01/2018)

Subject domain	English terms	Spanish terms
Law (12)	8,336	8,309
Economics (16)	9,698	9,647
Health (2841)	55,831	55,305
Total	73,865	73,261

3.2. Terminologia Oberta

TERMCAT is the terminology centre for Catalan. It was established in 1985 to ensure the development of Catalan terminology. TERMCAT's mission includes the task of standardising neologisms from the speciality lexicon and creating terminological resources in the domains of science, technology and the humanities.

Most TERMCAT data are also available in Terminologia Oberta[1], a web service that offers downloadable general interest terminology data in several formats, subject to Creative Commons licenses. Today, Terminologia Oberta offers over 100 downloadable terminological resources, which users can select by title, field, update date and download format.

At present, various formats are available for download:
- Tagged XML files provide all the content for terminology entries.
- HTML files provide terminology records related to terms and thus exclude definitions.
- PDF files provide personalised downloads for registered users.

Terminology entries contain the preferred term in Catalan and any synonyms, as well as equivalents in Spanish, French and English. XML files include a definition in Catalan and explanatory notes, when necessary.

Terminology data obtained from Terminologia Oberta in XML or HTML may be integrated into language engineering tools and used for other purposes, according to the Creative Commons licence to which they are subject. XML files are subject to an Attribution-NoDerivs 3.0 Unported (CC BY-ND 3.0) licence, i.e. terminology records may be shared, but no derivatives or remixes can be based on them. HTML files are subject to an Attribution 3.0 Unported (CC BY 3.0) licence, so terminology records can be shared and adapted. Both licences have specifications relating to attribution.

We classified the various terminological resources obtained from Terminologia Oberta by subject domains, and thus identified potential sources of terminological information for each subject. Specifically, in the first stage of this project, we chose to include the subject domains of law, economics and health, because these

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

subject domains are relevant in terms of societal needs and are the most common domains in which obtaining available open access lexical resources occurs.

3.3. Wikipedia

Wikipedia is a free encyclopaedia created collaboratively by thousands of volunteer users. It is a multilingual resource currently available in 291 languages, although the size of the resource varies considerably depending on the language. There are 15 languages with over 1,000,000 articles and 43 more languages with over 100,000 articles. Critics of Wikipedia claim that the collaborative way it is constructed, in which any user can create or edit articles, means that adequate quality cannot be guaranteed. However, the internal review system and the large number of volunteers willing to edit articles, correct mistakes and improve content means that in practice the end quality is very high. This resource can be effectively used in terminological research (Oliver, 2017), as most specialist domains are included. All the articles are marked with one or more categories, which in many cases can be considered specialist domains. These categories have no predetermined structure, so they can be created by the user; however, recommended categories are provided, as is a classification system by knowledge domain[2], which is comparable to other classification systems by domain, such as the IATE.

Wikipedia, as well as being accessible online, can also be fully downloaded, thus permitting efficient computer processing. The dBPedia project[3] (Lehmann, 2015) distributes content from various language versions of Wikipedia as information structured into importable database files, thereby permitting even more efficient computer processing.

3.4. The DOGC corpus

The *Diari Oficial de la Generalitat de Catalunya*[4] (DOGC) is an official media outlet in which the laws and regulations of the Government of Catalonia are published. Most texts in the journal are published in Catalan and Spanish. Since 2007, the journal has been published exclusively online on its website, where texts can be downloaded and processed to create a large parallel Catalan-Spanish corpus for the legislative and administrative subject domains. DOGC texts like laws and regulations are in the public domain and can be compiled and redistributed freely with no licencing issues. Oliver (2017) describes the DOGC corpus creation process with a number of examples of use of the corpus, include terminological research.

4. Methods

The method we designed to create the Catalan version of the IATE database and identify and extract terminology from the different subject domains is based on using two types of resource: terminology dictionaries and parallel corpora. The terminological e-dictionaries were created using this set of open access resources.

4.1. Terminology dictionaries

Open access terminology dictionaries, available in Catalan as well as Spanish and English, were used to create the Catalan version of the IATE database. Specifically, we used the TERMCAT Terminologia Oberta terminology sets for the subject domains of economics, law and health and two Terminologia Oberta terminology compilations that contain terms from various specialities. The terms from both IATE and TERMCAT are marked by subject domain; however, these domains are not equivalent. For example, the TERMCAT subject domain “Botanics” corresponds to the IATE domains “6006 Plant product” and “3606 Natural and applied sciences”. Other domains, such as the TERMCAT domain “Law” does in fact coincide with IATE’s “12-Law”. Therefore, we manually created a match between the subject domains of the two

[2] https://en.wikipedia.org/wiki/Outline_of_academic_disciplines
[3] https://wiki.dbpedia.org/
[4] http://dogc.gencat.cat

resources. Thus, in the law domain, we used both the general terminology compilations and nine terminological law dictionaries. However, in the economics domain, we used two general terminology compilations and six terminological economics dictionaries. And in the health domain, we used two general terminology compilations and three terminological dictionaries. Table 3 shows the number of entries in the Terminologia Oberta terminological dictionaries used.

Table 3. Entries in Terminologia Oberta terminological dictionaries

Terminological dictionaries	Entries
TO Law	6,559
TO Economics	1,881
TO Health	20,388

In addition, Wikipedia was used as a terminological resource, taking into account the article titles, to identify Catalan equivalents of IATE terms. In this case, we found a major limitation with regard to subject domains, as Wikipedia entries are not marked by domain, but include a category mark instead, which in some cases might be similar to a subject domain. Thus the Wikipedia entry categories do not always match the IATE, and therefore this information cannot be used. For this reason, a manual review of the translation equivalents extracted from this resource was essential. Table 4 shows the number of English-Catalan dictionary entries created from Wikipedia entries (specifically, using the dBPedia). Obviously, not all the entries in this dictionary are terms, as there are many Wikipedia entries on people and places and others name entities. However, as this dictionary was used only to search for known terms in English, the translation was very likely to be the equivalent term in Catalan.

Table 4. Number of entries extracted from Wikipedia

Wikipedia	Entries
ENG-CAT Wikipedia	396,072

4.2. Parallel corpora

Parallel Spanish-Catalan corpora were used to identify translation equivalents in IATE corresponding to the subject domains of law and administration (Vàzquez, 2018). To do this, we produced a parallel Catalan-Spanish corpus of texts published in the DOGC, updated towards the end of 2017. Given the large size of the parallel corpus, we used translation models based on statistical machine translation systems in order to optimise the search process for Catalan equivalents. This was carried out using Moses (Koehn, 2007) to calculate the phrase tables. Phrase tables contain n-grams in the source language with possible translations in the target language, together with a series of figures indicating the probability of each translation being the right one for the source n-gram. Below is a fragment of a phrase table calculated from the DOGC corpus, specifically the entries with the source language n-gram *fondos de inversión* (investment trust). Note that the last entry is the one with the highest probability rates and the n-gram in the target language corresponds to the Catalan translation of the term.

```
fondos de inversión ||| els fons d' inversió ||| 0.142857 0.401642 0.00900901 0.00192854 ||| 0-1 1-2 2-3 ||| 7 111 1 ||| |||
fondos de inversión ||| fons d' inversió de ||| 0.2 0.401642 0.00900901 0.0533415 ||| 0-0 1-1 2-2 ||| 5 111 1 ||| |||
fondos de inversión ||| fons d' inversió tant ||| 1 0.401642 0.00900901 2.34972e-05 ||| 0-0 1-1 2-2 ||| 1 111 1 ||| |||
fondos de inversión ||| fons d' inversió ||| 0.78125 0.401642 0.900901 0.200147 ||| 0-0 1-1 2-2 ||| 128 111 100 ||| |||
```

When using large parallel corpora, these tables may contain hundreds of millions of entries. For instance, the Spanish-Catalan phrase tables calculated from the DOGC corpus contain 189,326,800 entries. In order to streamline the search process with these tables, we implemented an indexing and storage process in an SQLite database, to facilitate fast searches. These indexing functions and the automatic translation equivalent search functions are included in the TBXTools automatic terminology extraction tool (Oliver, 2015). Table 5 shows the size of the corpora used.

Table 5. Size of parallel corpora

Corpus	Segments	Words CAT	Words SPA	Words ENG
DOGC corpus	8.074.284	188.908.522	197.991.183	-

4.4. Terminological dictionary creation process

When creating a conventional dictionary, the content consists of an entry, known simply as the article in general language dictionaries and the record in terminological dictionaries. Furthermore, a dictionary may consist of main entries, which contain the lexical category, equivalents, definition and notes, and secondary entries, which consist of the lexical category and reference to the record to which they belong. By contrast, the specialised e-dictionaries we created from open access linguistic resources do not have a fixed structure, consisting mainly of the content available in the sources and possibly including additional supplementary information. Thus, unlike conventional terminological dictionaries, the new terminological resources we compiled to create the IATE database in Catalan consist of the entry in English, the translation equivalents in Catalan and Spanish, the grammatical form, the subject domain, the level of reliability in IATE and the IATE reference. After creating the new terminological resources, the process of creating the corresponding terminological e-dictionaries began.

TERMCAT advised on the methodology throughout the project, which complemented the knowledge and expertise of both researchers at the Universitat Oberta de Catalunya.

TERMCAT pointed out a few aspects to be taken into account throughout the work process. First of all, it stressed the need to check the definition of IATE terms to ensure that the proposed Catalan equivalents were correct. The concept behind the term is essential in guaranteeing the validity of the Catalan translation.

TERMCAT also recommended providing data other than the Catalan equivalent, which is of the utmost importance from a lexicographical point of view and can be of use to users (TERMCAT, 2010). One such recommendation was the inclusion of the part of speech or lexical category of the Catalan term, indicating its syntactic function. The terminology centre further considered the need to provide users with the source of the

Catalan term, i.e. the linguistic, terminological or specialised resource in which the Catalan equivalent was found. This information could be used to judge the degree of reliability of the Catalan term (Gelpí, 2004).

Finally, as IATE terms are classified according to a concept map, available on their website, TERMCAT proposed including the domain or field associated with the term. Information on the term domain contextualises the term within a specific subject area.

TERMCAT also guided the UOC's work on the layout of the information for the online dictionary. It is important to record all data in a structured set of domains within a terminology entry. In addition to the aforementioned items (i.e. part of speech, source of the Catalan equivalents and the term domain), the online dictionary will contain IATE terms in English and Spanish and the specific IATE ID number of the terminology record from which data were taken.

Each terminology record in the IATE database has a specific identification number. This ID number will enable users to connect the terminology record in the future online dictionary to the original terminology record in IATE. Thus, they will be able to check IATE for the information provided in the set of domains for IATE terminology entries, such as terms in other official European languages, definitions and notes.

5. Results

In our study we have designed a new methodology of data generation and processing with the aim of producing terminological e-dictionaries using open access linguistic resources. We applied this new methodology to building the IATE database in Catalan, specifically in the domains of economics, law and health. Below we describe the results obtained in constructing these terminological dictionaries from terminology dictionaries and parallel corpora.

5.1. Terminology dictionaries

We used the terminology dictionaries in the TERMCAT Terminologia Oberta to construct the IATE database in Catalan in the subject domains of law, economics and health. The use of this open access resource allowed us to automatically identify equivalent terms in Catalan with precision levels of 75% to 87% with respect to the total number of terms in Spanish in the IATE in each speciality. Table 7 shows the number of terms automatically obtained in Catalan from the terminology dictionaries that were found to be correct after a manual check, together with the precision and recall values for each subject domain.

Table 7. The number of terms obtained from the Terminologia Oberta terminology dictionaries

Subject domains	Terms obtained	Correct terms	Precision	Recall
Law	2,098	1,722	82.08%	20.73%
Economics	210	157	74.76%	1.63%
Health	10,085	8,786	87.12%	15.89%

The results show that the precision for law and health was over 80%, but below 75% for economics. Despite this, the accuracy may be considered adequate and the method efficient in all three cases, as long as the automatically obtained results are revised manually. With regard to recall, in the case of economics it was very low, at only 1.63%, as there are very few economic terms in the Terminologia Oberta dictionaries (just 1,881).

Wikipedia was also used as an open access terminology resource in constructing the IATE subject domains of law, economics and health in Catalan. In this case, first we automatically identified the equivalents in Catalan present in the encyclopaedia and then carried out an automatic assessment based on the terms in Terminologia Oberta and in the DOGC parallel corpus for the subject domain of law (Table 8).

Table 8. Number of terms obtained from Wikipedia as a terminological resource

Subject domain	Terms obtained	Terms assessed	Correct terms	Precision	Recall
Law	561	397	245	61.71%	4.17%
Economics	598	36	19	52.78%	3.27%
Health	3,544	1,011	703	69.54%	4.46%

As can be seen, the results for both precision and recall obtained with Wikipedia as a terminology resource are lower than those obtained from the TERMCAT Terminologia Oberta dictionaries. This may be for two reasons. Firstly, when identifying terms, we did not take into account information on specialities, as the categories assigned to Wikipedia articles, which could be considered equivalent to specialities, do not match the IATE specialities. It should be remembered that the categories assigned to Wikipedia articles are free and chosen by their authors and editors. This might explain the low accuracy. Secondly, Wikipedia is an encyclopaedic resource, not a terminology resource. Although it may have entries that match terms, many entries are for name entities, such as people, places and things. This might explain the low coverage.

5.2. Parallel corpora

The DOGC parallel corpus is an open access linguistic resource which we used to identify the IATE terms in Catalan for the subject domain of law (Table 9).

Table 9. Number of terms obtained from the DOGC parallel corpus

Subject domain	Terms obtained	Correct terms	Precision	Recall
Law	2,678	2,554	95.37%	30.74%

As can be seen, the DOGC parallel corpus provides excellent results for both precision and recall. The use of parallel corpora to identify terminology is a very effective strategy, as long as a big enough corpus is available for the languages and speciality involved.

We provide some examples obtained from the Terminologia Oberta terminology dictionaries, Wikipedia as a terminological resource and the DOGC parallel corpus in order to show the entry structure in Catalan IATE e-dictionary. Terminologia Oberta and Wikipedia are open access lexical resources which provides terms from each subject domain, and DOGC parallel corpus is an open access lexical resource to provide terms from law subject domain. The entry structure in Catalan IATE e-dictionary includes the IATE ID number; the Catalan terms from Terminologia Oberta, Wikipedia and DOGC parallel corpus; the grammatical category, and the English and Spanish terms from IATE (Table 9).

Table 9. Terms obtained from the Terminologia Oberta, Wikipedia and DOGC parallel corpus

	Catalan IATE e-dictionary Entry structure		
IATE's subject domains	Terminologia Oberta	Wikipedia	DOGC parallel corpus
Law (12)	**IATE-34834** *ca* execució forçosa n f *ca* potestat d'execució forçosa n f *es* ejecución forzosa *en* enforcement	**IATE-905458** *ca* *dret positiu* *es* derecho positivo *en* positive law	**IATE-34663** *ca* disposicions financeres *es* disposiciones financieras *en* financial provisions
Economics (16)	**IATE-757391** *ca* broker [en] n m, f *ca* mediador n m, f *ca* mediadora n m, f *es* intermediario *en* middleman	**IATE-1696383** *ca* estabilitat econòmica *es* estabilidad económica *en* economic stability	
Health (2841)	**IATE-35093** *ca* principi actiu n m *es* principio activo *en* active ingredient *en* active substance *en* pharmacologically	**IATE-1489091** *ca* neuropatologia *es* neuropatología *en* neuropathology	

6. Discussion

The results described above show that using open access linguistic resources permits terminology to be compiled from a variety of sources available in different formats and new terminological dictionaries to be produced easily and efficiently for all languages.

Furthermore, combining linguistic resources meant the maximum number of terms for the subject domains of law, economics and health for the Catalan version of the IATE could be compiled. If new linguistic resources in Catalan become available in these domains, then the number of terms in these online terminological dictionaries can be increased. Thus, the use of open access sources means dictionaries created from this new lexicographical methodology can be steadily enlarged.

In addition, the availability of open access linguistic resources raises the possibility of creating new large general language products and applying tools and methods from computational linguistics. Without these resources, undertaking new large-scale terminology projects, such as the creation of the multilingual IATE database for new languages, would be unthinkable.

It is also worth stressing that the use of different types of lexicographical resources provides extremely useful materials, such as monolingual, parallel corpora and comparable corpora, which facilitate tasks such as automatic terminology extraction, improving automatic and assisted translation and the creation of terminology databases useful to computer-assisted translation tools, while applying natural language processing techniques.

The new methodology for creating terminological e-dictionaries described in this paper can be applied to any language. As shown in the results of the assessment presented here, the methodology for producing dictionaries based on parallel corpora is very effective and provides excellent results for both precision and recall.

In short, initiatives by institutions, companies and official organisations to create linguistic resources have facilitated the birth of e-lexicography, thereby enabling new strategies for creating general language and terminological dictionaries to be put into practice, thanks to new technologies.

7. Publication of terminological e-dictionaries

The Catalan IATE e-dictionary will be published on the TERMCAT website. The number of terms compiled using the new lexicographic method is shown in Table 10.

Table 10. Total terms in the terminological dictionaries

	Catalan terms	
IATE subject domain	Law	5,337
	Economics	808
	Health	13,629

Regarding the general structure of the dictionary, a presentation will inform users about the specific content of the dictionary and the research and production methodology adopted.

Guidelines on how to look up information will also be provided. Terms can be looked up in three ways: using the alphabetical indices (there is a different index for each language); using the topic-based access; and making simple or advanced searches in any of the dictionary languages. Multi-accessibility allows users to find a specific term through multiple routes and thus choose the one that best suits them (Gelpí, 2004). In addition, the authorship of the dictionary, the work team and the bibliography used throughout the process will be indicated.

As for the dictionary structure, by default, lexical entries are sorted by alphabetical order of the Catalan terms, but users can switch the main language at any time or look up the terms by topic, using the topic index.

Finally, the dictionary structure will provide different lexicographical data distributed in the terminological domains for the entry: terms in Catalan, equivalent IATE terms in Spanish and English, the source of the Catalan terms, grammatical category of the Catalan terms, IATE ID number for the terminological entry in the European database, and the subject domain attributed to the term in IATE.

8. Conclusions and future work

This paper has presented a new methodology applied to the production of terminological e-dictionaries, using open access linguistic resources. Implementation of this new methodology focused on the construction of the IATE database in Catalan. The results from the research show that it is possible to use and combine open data available in different repositories to easily and effectively create new resources.

In terms of creating dictionaries, using open data streamlines the production process in various ways. Firstly, all terminological projects generally start by defining the project, i.e. defining the conceptual scope of the product, the content and the target public. Part of the conceptual definition is specified by establishing a concept map and stating the approximate number of terms it will contain. In our case, using the IATE database as the starting point means that the project is based on the concept map used to organise the terms in this European database. Thus, with this new working method, this first phase, which is usually rather time-consuming, is not required, as it uses a tool from another product, in this case the IATE.

With regard to creating the product, the project definition and concept map serve as a guide throughout production and are particularly important in extraction, i.e. when selecting the terms to include. The working method described in this paper permits the terminological work to start at this point, i.e. with the concept map defined and the terms selected. Thus, with regard to the steps involved in producing a terminological dictionary, the terminological work strictly starts with the information completion phase, in this case completion of the Catalan terms, although, as commented above, certain actions that are not usually involved in producing a terminological dictionary have to be carried out.

The data are also revised in a different way. When using this new methodology in terminological resource production, it is necessary to ensure that the open data used in automatic completion are of an adequate quality. Therefore, before starting the automatic completion process, the resources to be used must be adequately assessed. Thus the revision of the form (Catalan, in this case) also differs from that carried out in the ‘traditional’ process of producing a terminology product. The revision is not a strictly linguistic assessment (assessing the adequate form); rather, it focuses on checking that the automatic completion has been carried out correctly (i.e. the correct form from the open resources used has been added). For instance, in the law subject area, during revision it is checked that the Catalan term ‘propietari, propietària’ corresponds to the ‘owner’ IATE’s term.

For users, the improvement is substantial. The unification of data in a single resource helps them find the answer to their search quickly, easily and economically, without having to consult various resources. Furthermore, data transparency (e.g. an explanation of the working process or the source of the terms) is a guarantee of quality for users, as it allows them to assess the resource used as a source of reference from different points of view.

Furthermore, given that the data are available in open format, users can reuse them with linguistic engineering tools in accordance with their needs.

With regard to data dissemination, the idea is to include the Catalan terms in the IATE database so they are available to all users in the source database. However, this idea does not fall within the strict scope of the terminological project and the new methodology presented here. Eventually, the Catalan version of the IATE, as well as being available to the public on the TERMCAT website, will also be available in the IATE internal version, thanks to the agreement signed between TERMCAT and TermCoord, the European Parliament Terminology Coordination Unit. The main role of TermCoord is to assist translators with their day-to-day tasks and facilitate terminology research and terminology management in the translation units, as well to increase the European Parliament’s contribution to IATE. Therefore, TermCoord and TERMCAT share a common interest in developing and offering terminology to their users. This is why both institutions signed an agreement to enrich the contents of the internal IATE (Nin, 2016). More specifically, TERMCAT provides TermCoord with new terminology in Catalan, in a joint effort to enrich and update the contents of IATE with relevant terminology data from TERMCAT.

With regard to future challenges, the most immediate one is updating the data. It should be borne in mind that the result of this project, an e-dictionary, is not a finished product; rather, it is one that will grow in step with the resources on which it is based. The IATE database expands on a daily basis in terms of the number of entries, and new resources that include Catalan are constantly created. Thus the challenge is to keep the growth of this corpus of IATE terms in Catalan alive.

In closing, with regard to the creation of terminological dictionaries, a key challenge is the systematic incorporation of this new methodology (or, at least, part of this new methodology in the habitual working method. The new methodology provides linguistic engineering tools for producing specialist dictionaries, thus streamlining the different work phases. This is particularly notable in the extraction phase. Undoubtedly, incorporating these tools into the production of dictionaries will permit extraction from many more sources (in this case, Terminologia Oberta, the DOGC and Wikipedia, among others).

References

Abel, Andrea, and Christian M. Meyer. 2013. «The Dinamics Outside the Paper: User Contributions to Online Dictionaries». In *Proceedings of the eLex 2013 Conference,* 17-19: 179-194. Tallinn, Estonia. http://eki.ee/elex2013/proceedings/eLex2013_13_Abel+Meyer.pdf

Ball, Liezl. 2016. *An Evaluative Study to Determine to What Extent Technology can be Used in e-Dictionaries to Provide Relevant Information on Demand*. Dissertation. University of Pretoria. https://repository.up.ac.za/handle/2263/61338

Blancafort, Helena, Francis Bouvier, Béatrice Daille, Ulrich Heid, and Anita Ramm 2013. «TTC Web Platform: from Corpus Compilation to Bilingual Terminologies for MT and CAT Tools» In *Futures in technologies for translation (TRALOGY II)*, 14-pages, Paris, France.

Declerck, Thierry. Eveline Wandl-Vogt, Simon Krek, and Carole Tiberius. 2015. «Towards Multilingual eLexicography by Means of Linked (Open) Data». In *Proceedings of the 4th Workshop on the Multilingual Semantic Web (MSW4)*, ed. by Jorge Gracia, John P. McCrae, and Gabriela Vulcu, 51–58. http://ceur-ws.org/Vol-1532/

de Schryver, Gilles-Maurice. 2012. «Trends in Twenty-five Years of Academic Lexicography». *International Journal of Lexicography* 25(4): 464–506. Oxford University Press. https://doi.org/10.1093/ijl/ecs030

Engelberg, Stefan, and Lothar Lemnitzer. 2009. *Lexikographie und Wörterbuchbenutzung*, 4th edition. Tübingen: Stauffenburg Verlag.

Fuertes-Olivera, Pedro A., and Henning Bergenholtz (eds.). 2011. *e-Lexicography: the Internet, Digital Initiatives and Lexicography*. London / New York: Continuum International Publishing Group.
Gelpí, Cristina. 2003. «El estado actual de la lexicografía: los nuevos diccionarios». In *Lexicografía española,* ed. by Antonia María Medina Guerra, 307–332. Barcelona: Ariel.

Gelpí, Cristina. 2004. «Diccionaris digitals especialitzats per la temàtica: estat actual i perspectives». *Scripta Nova: revista electrónica de geografía y ciencias sociales*, 8(170). Barcelona: Universidad de Barcelona. http://www.ub.es/geocrit/sn/sn-170-69.htm

Gouws, Rufus H. 2014. «Article Structures: Moving from Printed to e-Dictionaries». *Lexikos*, *24*(1): 155–177. https://www.ajol.info/index.php/lex/article/view/112563/102316

Granger, Sylviane, and Magali Paquot, 2010. «Customising a General EAP Dictionary to Meet Learner Needs». In *Proceedings of ELEX2009: eLexicography in the 21st century: New challenges, new applications*, ed by Sylviane Granger, and Magali Paquot, 87-96. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, Sylviane, and Magali Paquot, (eds.). 2012. *Electronic Lexicography*. United Kingdom: Oxford University Press.

Heid, Ulrich, Daan J. Prinsloo, and Teo J. D. Bothma. 2012. «Dictionary and Corpus Data in a Common Portal: State of the Art and Requirements for the Future». *Lexicographica*, *28*(1): 269–291.
https://doi.org/10.1515/lexi.2012-0014

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*: 47-62.
http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202014/euralex_2014_002_p_47.pdf

Hibbert, David Brynn, Pentti Minkkinen, Nicolas M. Faber, and Barry M. Wise. 2009. «IUPAC project: A glossary of concepts and terms in chemometrics». *Analytica chimica acta*, *642*(1-2): 3-5.

Koehn, Philipp; Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 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,*177–180. Association for Computational Linguistics. Prague, Czech Republic.
https://bit.ly/2wqZIs6

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

Lew, Robert. 2013. «Online Dictionary Skills». In *Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013 Conference*, ed. by I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets and M. Tuulik, 16–31. Tallinn, Estonia. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.
https://repozytorium.amu.edu.pl/bitstream/10593/8138/1/Lew_2013%20Online%20dictionary%20skills.pdf

L'Homme, Marie-Claude, and Monique C. Cormier. 2014. «Dictionaries and the Digital Revolution: A Focus on Users and Lexical Databases». *International Journal of Lexicography*, *27*(4): 331–340. Oxford University Press.
https://doi.org/10.1093/ijl/ecu023

Logar, Nataša, and Iztok Kosem. 2013. «TERMIS: a Corpus-driven Approach to Compiling an e-Dictionary of Terminology». In *Electronic Lexicography in the 21st Century: Thinking Outside the Paper: Proceedings of the eLex 2013 Conference,* 164-178. Tallinn, Estonia.
http://eki.ee/elex2013/proceedings/eLex2013_12_Logar+Kosem.pdf

Martin, Willy, and Hennie van der Vliet. 2003. «Design and Production of Terminological Dictionaries». In *A Practical Guide to Lexicography*, ed. by P. van Sterkenburg, 6: 333-365. John Benjamins Publishing.

Maslias, Rodolfo. 2014. «Combine EU Terminology with Communication and Ontology Research». In *Terminology and Knowledge Engineering,* 9 pages. Berlin, Germany.

Nesi, Hilary 2000. «Electronic Dictionaries in Second Language Vocabulary Comprehension and Acquisition: the State of the Art». In *Proceedings of the Ninth Euralex International Congress*, ed. by Ulrich Heid, Stefan Evert, Egbert Lehmann, and Christian Rohrer, 839–847. Stuttgart, Germany.
Nesi, Hilary. 2013. «Researching Users and Uses of Dictionaries». *The Bloomsbury Companion to Lexicography*: 62-74.
Nin, Anna. 2016. *Catalan Terminology for IATE*. Luxemburg: European Parliament. Terminology Coordination.
http://termcoord.eu/2016/03/termcoord-and-termcat-in-close-cooperation-to-enrich-iate-content/

Oliver, Antoni, and Mercè Vàzquez. 2015. «TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction». In *International Conference on Recent Advances in Natural Language Processing* (RANLP 2015), 473–479. Hissar, Bulgaria.
http://www.aclweb.org/anthology/R15-1062

Oliver, Antoni, Mercè Vàzquez, and Georgina Ubide. 2017. «Estudi de la fiabilitat de la Viquipèdia com a recurs terminològic». *Revista Tradumàtica. Tecnologies de la Traducció,* 15: 10–20. https://doi.org/10.5565/rev/tradumatica.193

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.
http://latina.phil2.uni-freiburg.de/pusch/zfk/30/16_Oliver.pdf

Pastor, Verónica, and Amparo Alcina. 2010. «Search Techniques in Electronic Dictionaries: A Classification for Translators». *International Journal of Lexicography* 23(3): 307–354. Oxford University Press.
https://doi.org/10.1093/ijl/ecq015

Rundell, Michael, and Adam Kilgarriff. 2011. «Automating the Creation of Dictionaries: Where will it all end». In *A Taste for Corpora. In honour of Sylviane Granger*, ed. by Fanny Meunier, Sylvie de Cock, Gaëtanelle Gilquin, and Magali Paquot, 257–282. Amsterdam: John Benjamins.
Tarp, Sven. 2008. *Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer Verlag.

Tarp, Sven. 2009. «Beyond lexicography: New visions and challenges in the information age». In *Lexicography at a crossroads: dictionaries and encyclopedias today, lexicographical tools tomorrow*, ed by Henning Bergenholtz, Sandro Nielsen, and Sven Tarp, 90(1): 17-32. Bern: Peter Lang.

Tarp, Sven. 2012. «Online Dictionaries: Today and Tomorrow». *Lexicographica: International Annual for Lexicography* 28(1): 253-268. Berlin, New York: De Gruyter.
https://doi.org/10.1515/lexi.2012-0013

TERMCAT, CENTRE DE TERMINOLOGIA. 2010. *El diccionari terminològic*. Vic: Eumo Editorial; Barcelona: TERMCAT, Centre de Terminologia. (En Primer Terme; 9. Criteris i Mètodes).
ISBN 978-84-9766-394-6; 978-84-393-8676-6

Trap-Jensen, Lars. 2010. «One, Two, Many: Customization and User Profiles in Internet Dictionaries». In *Proceedings of the XIV Euralex International Congress*, ed. by Anne Dykstra and Tanneke Schoonheim: 1133–1143. Ljouwert: Fryske Akademy.

Varantola, Krista. 2002. «Use and Usability of Dictionaries: Common Sense and Context Sensibility?» In *Lexicography and Natural Language Processing*, ed by Marie-Hélène Corréard, 30-44. A Festchrift in Honour of B. T. S. Atkins. Euralex.

Vàzquez, Mercè, Antoni Oliver, and Georgina Ubide. 2018. «La terminologia jurídica del IATE en català». *Revista de Llengua i Dret, Journal of Language and Law* 69: 139–153.