Textual genre based approach to use wordnets in language-for-specific-purpose classroom as dictionary

Itziar Gonzalez-Dios, German Rigau
Ixa Group
University of the Basque Country (UPV/EHU)
{itziar.gonzalezd,german.rigau}@ehu.eus

Abstract

When teaching language for specific purposes (LSP) linguistic resources are needed to help students understand and write specialised texts. As building a lexical resource is costly, we explore the use of wordnets to represent the terms that can be found in particular textual domains. In order to gather the terms to be included in wordnets, we propose a textual genre approach, that leads us to introduce a new relation *term_used_in* to link all the possible terms/synsets that can appear in a text to the synset of the textual genre. This way, students can use wordnet as dictionary or thesaurus when writing specialised texts. We explain our approach by means of the logbooks and terms in Basque. A side effect of this works is also enriching the wordnets with new variants and synsets.

1 Introduction

Language for specific purposes (LSP) is a subfield of applied linguistics that studies language in different contexts e.g. language for business, language for engineering, etc. The work in this area has been mainly done in the field of terminology, but nowadays theory-building data analysis and classroom/workplace practice have an important role in the development of the field (Gollin-Kies et al., 2015).

In this paper, we propose to use wordnets as a lexical/terminological references to consult in LSP teaching. Exactly, we present a method that combines textual genre analysis together with classroom practice in order to compile terms to be included in wordnets. The final aim is to provide students with a multilingual and semantically rich consult resource that will gather of the terms that can be used in a specific textual genre.

We have decided to use wordnets as a basis resource because they offer rich semantic information liked to different languages and we think it is appropriate to centralise all the resources. Moreover, its relations are helpful for students when looking for similar words, related concepts, etc. That is why, we propose a new relation: the *term_used_in*. This relation will link all the terms/synsets that can be used in a textual genre, without altering the hierarchy of wordnet.

The context of this research is Basque as LSP for sea studies. Currently, many subjects are taught in Basque at university level, but it is still a language under normalisation (the standard variant was officially created in 1968) and this fact influences the corpus and the resources we can use: there is no specialised corpus on some fields of knowledge and lexicographic/terminological data is sparse. Moreover, as in the case of fishing or farming, the specialised variant has been oral. That is why we propose to base on textual genres, standard models of text types. Following Cabré (1999), we also think that specialised texts meet certain norms that vary depending on the domain. Indeed, textual genres are a key component on specialised discourse (Gotti, 2008). Moreover, During this work, as side-effect we are also enriching wordnets, in our case, Basque WordNet (BWN) (Pociello et al., 2011).

To illustrate our approach, we report on case study about logbooks, a nautical textual genre that compiles terms from different domains such as metrology, meteorology, geography among others. We will work on terms on Basque language, a language under normalisation that is developing its specialised languages.

This paper is structured as follows: in Section 2 we sum up the context of our work; in Section 3 we present our approach and we show an example of its practical application in Section 4. After that, in Section 5 we discuss some issues relating

the process and we conclude and outline the future work in Section 6.

2 Domains, specialised knowledge and textual genres

Domains are usually defined as unitary areas of knowledge (specialised or not) and are related to semantic fields, subject matters, broadtopics, subject codes, subject domains, categories... In WordNet (Fellbaum, 1998) we can find the *Domain of synset/Member of this domain*, where synsets are linked with a category, region or usage pointer (domains) and the domains are linked with synsets. In WordNet Domains (Magnini and Cavaglia, 2000; Bentivogli et al., 2004) synsets have been semi-automatically annotated with one or more domain labels from a set of 165 hierarchically organised labels, contrasted to the Dewey Decimal Classification (DDC) system. In eXtended WND (González-Agirre et al., 2012) a graph-based approach was carried out to improve WordNet Domains by means of a simple inheritance process through the nominal and verbal hierarchies and applying UKB to propagate the domain information. BabelDomains (Camacho-Collados and Navigli, 2017) are automatically created by combining distributional and graph-based approaches and are based on Wikipedia categories for the featured articles. These hierarchical approaches are related to classical terminology work.

Specialised knowledge is the principles and techniques that are acquired in a particular discipline. According to Cabré (2003), specialised knowledge is transferred by terminological units (terms) and this transfer occurs during the specialised communication, in the discourse produced in each situation by the experts (communicative approach to terminology). A way of studying the specialised communication is through the corpus analysis. Indeed, many works dealt with terminology extraction from corpus e.g. Alegria et al. (2004).

A key component of specialised discourse is the textual genre, a prototypical type of discourse. Cabré (2005) points out that documents corresponding to textual genres are used in every professional domain, and that students should know their standard features and characteristics in order to be able to write them. These standards include format, phraseology and vocabulary. In other words, each textual genre will be marked by its own terms.

3 Approach for gathering terms

The approach we propose is conceived for environments where no corpus or few texts exist and the lexicographic/terminological resources are sparse and scattered. Next, we explain the proposed approach.

- **Critical overlook of the existing and referential resources**: before we start working on any target field it is important to know which are the lexical/terminological resources we can consult and reuse. Moreover, it is also convenient to analyse how the terms in the target domain are represented in general-purpose dictionaries/ terminological databases.

- **Analysis of the communicative needs and textual genres**: in order to choose a textual genre, we need to make an analysis of the the communicative needs, that is, we need to know which textual genres are the most used. Classroom practice is important in this step, getting to know which texts are most used and most difficult to write can be decisive to choose the textual genre. Another option is the one presented by da Cunha and Amor Montane (2019) where they make questionnaires to domain experts to know which are the most used and most difficult texts to write. This step could also be automatised if specialised corpora were available.

- **Term compilation and representation in wordnets**: in order to compile the terms, we need to consult in the existing and previously analysed resources the terms that can be used in the target textual genre. Then, we will include the terms in Basque WordNet because of its reusalibity as variants in their respecting synset. We will add the relation *term_used_in* to the hypermyn of the synsets to link it to the text genre.

We propose to create the *term_used_in* relation in order to offer students LSP students help when writing and consulting specialised vocabulary and terminology.

4 Practical application of the approach: Basque nautical terms and logbooks

In this section we describe a practical application of the above presented methodology. In this case study we will report on the logbooks and Basque terms.

4.1 Critical overlook of resources

In this section we present the resources where we can find nautical terminology in Basque.

Relating the general resources, Euskalterm[1] is the main terminological database for Basque. When looking for nautical terminology in Euskalterm, they appear under other subjects such as 1) Fishing, 2) Sports, Games and Leisure, 3) Industry, 4) Law 5) Geology and Meteorology or 6) Education and pedagogy. For instance, the term *ziaboga* (turning, a basic manoeuvre) can be found in the sports and leisure domain, because it has been compiled in the rowing dictionary.

Another general resource is *Zientzia eta Teknologia Hiztegi Entziklopedikoa*[2], a dictionary of Science and Technologies. In this dictionary, there are three categories where nautical terms can be found: sea, oceanographic and meteorology. Moreover, terms related to sea engineering can be found in categories such as general technology, electric technology and mechanic technology.

The last general resource we want to mention is WordNet. Using the synset *seafaring* and the *domain term category* relation we can find nautical terms and in WordNet Domains we also do find the nautical category. But due to the size of the BWN not all the English words are covered by the Basque version. For example, if we look for the hyponyms of the word *itsasontzi* (stands for vessel, watercraft), there are four synsets in Basque (*belaontzi* sailing ship, *galera* galley, *arrantzaontzi* fishing boat and *yate* yacht) whereas there are fourteen for English.

Relating the maritime specific resources, the most important is "*Itsasontziaren Eskuliburua*" (The Manual of the Vessel) (Sotés et al., 2015), a manual that has been written by professors of sea studies and it is conceived as a photo-dictionary. It is divided in three topics: the vessel, the port and the containerisation and it includes four term lists (Basque-Spanish, Spanish-Basque, Basque-English and English-Basque). In the book, the terms related to the previously mentioned topics are explained and illustrated with figures. This resource is so far the best for the nautical terminology and it is being integrated in *Terminologia Zerbitzurako Online Sistema (TZOS)* (Arregi et al., 2013), the terminological database of academic Basque.

Moreover, there are some resources in Basque related to navigation and the sea e.g. dictionaries such as the "Fishing dictionary", "Transport and Logistics dictionary", "Maritime Law dictionary or "Astronomy dictionary" included in EuskalTerm or independent dictionaries such as the "Dictionary of the Port of Pasaia", "the Activity Book of the Port of Bilbao", "Regatta dictionary", "Biscayan fishermen dictionary", or fishmongers dictionaries. There are also PhD theses on the fishermen speech and vocabulary of certain towns.

Finally, MARITERM (Marinelli et al., 2004) is a maritime lexical database structured as WordNet that contains the specialised lexicon of navigation and maritime transport. It can be considered a domain adaption of WordNet, with its peculiarities to the nautical domain. The lexicon includes also terms of other domains such as meteorology, geography, cartography, astronomy, law related to the sea and maritime contracts, sailing races or publications.

In conclusion, the shortfalls of the general resources are that a) nautical terms are spread in different categories (terms are scattered) and b) the coverage is low. The main problem of the maritime resources is that c) their texts and wordlists are difficult to process computationally due to their format (some of them are not even digital) and, that some of them are not available or have the reusable licenses.

4.2 Analysis of the communicative needs and textual genres

In order to analyse the communicative needs we have examined the documentation that needs to be carried on the ships. In the case of vessels with Spanish ensign, the documentation is specified by the law 14/2014, in the articles 78-87 of the chapter second chapter. According to this law, the documents that must be carried on the ship are the certificate of enrolment, navigation certificates, ensign, crew list, logbook and the bell book (logbook concerning the machines). In our opinion, the linguistically and terminologically most inter-

[1] http://www.euskadi.eus/euskalterm/
[2] https://zthiztegia.elhuyar.eus/

esting documents are the logbooks. Moreover, this textual genre is one of the most used by students and professionals.

Logbooks are the documents that the captain writes and must compile every eight hours with all the important events relating the nautical and meteorological incidents in the navigation. So, in this textual genre we will find terms about measures, size, coordinates, directions, meteorologic phenomena and places, which, in our opinion, makes it to be a very rich textual genre on nautical terminology.

As a curiosity, we want to mention that in Paleoclimatology based on the logbooks from Catalan seafarers dating from the 17th century, Prohom (2002) have rebuilt the Atlantic ocean climate. Therefore, this textual genre is not only interesting for linguistic studies but also for historical and climatological ones.

4.3 Term compilation and representation in BWN

Even though logbooks are symbolically written in vessels, the purpose of the term compilation and representation is to provide students how can they use the terms in Basque. To that end, we have looked for the terms that can be used in the logbooks in the Basque referential resources. Following, we list the the hypernyns of the terms we have gathered. The list of all Basque terms is shown in Gonzalez-Dios (2019).

- Magnitudes (4 terms)
- Cardinal and intercardinal directions (16 terms)
- Meteorological phenomena:
  - Wind: Beaufort scale, wind oscillation and wind speed (24 terms)
  - Sea: Douglas scale, form of the waves and *galerna* types (16 terms)
  - Clouds: types, forms, distribution and moisture (28 terms)
  - Precipitations: types, intensity, amount of liquid (different for rain and snow), types of storms (21 terms)
  - Temperature (10 terms)

We have included all the terms that were not already covered and have an equivalent synset in English e.g. *abiadura_angeluar* linked to *angular_velocity* in BWN. We have decided not to include geographic terms, because so far entities have not been added to BWN. Dates and hours have also not been added.

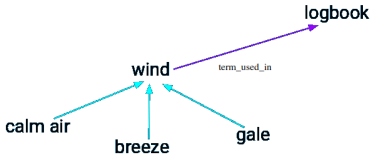

Figure 1: Example of the *term_used_in* relation

Finally, the hypernyns are linked to the synset *logbook* via the *term_used_in*. An example of this is shown in Figure 1, where a synset (*wind*) is connected to the synset of the textual genre where it is used (*logbook*). This way, students can consult which terms can appear in this textual genre.

5 Discussion

Following the presented approach, we have gathered terms and included in a semantically rich resource such as BWN, and tried to avoid the dispersion of terms, an important problem with Basque nautical terminology as shown in Section 4.1. In addition, we have provided LSP students an improved and centralised resource to help write the specialised texts.

However, when trying to represent these terms in BWN we have found some issues we will like to discuss. The first is about the conceptualisation: in logbooks and referential resources some terms are organised in a different way from WordNet and sometimes that classification was more detailed than the WordNet hierarchy e.g the types of the clouds (Figure 2) were organised in our resource taking the levels low, mid, high... (in green) into account whereas in WordNet all of them are together (in black) .

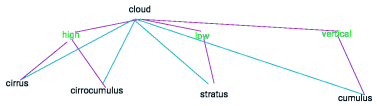

Figure 2: Example of categorisation of clouds

Secondly, many terms are not in English.

For instance, local meteorologic phenomena like *galerna*, or international conventions such as the Douglas scale, shapes of waves, etc. So the need of new language dependant concepts are necessary. That is, the need of CILI (Bond et al., 2016) is remarkable in this work. In fact, many of these terms are international and other wordnets would profit from these new synsets.

Thirdly, as we have seen, several domains are linked by gathering terms/synsets approach. In the case of the logbooks, moreover, it is remarkable that, although it is a text from the nautical domain, most of its words are not included in the *nautical* domain of WordNet Domains hierarchy. This makes us think of bigger domains, domains where knowledge from different areas meets. Indeed, this is related to communicative bottom-up approaches (Zabala et al., 2018).

Finally, we would like to encourage the use of the proposed relation *term_used_in* so that all these variants can be related. Indeed, we think it can be a step towards the characterisation of professional textual genres in wordnets. Moreover, as textual genres are *international* models, this approach can help to improve the recall of wordnets, since it allows to detect missing synsets, that is, words that are in certain texts, but not yet in WordNet.

6 Conclusion and Future Work

In this paper we have presented a method to get specialised knowledge by gathering terms and to include it in wordnets. Moreover, we want to encourage the use of wordnets in LSP classrooms as a dictionary, that can be useful for less-resourced specialised languages. To that end, we rely on textual genres as basis for term/synset gathering to be included in BWN. Indeed, textual genres have been proven to be useful to compile terms that would not appear in traditional hierarchies since they belong to different domains. We have explained our approach by means of the case of logbooks in Basque, a professional textual genre with terms from different domains and a language which is developing its specialised languages. Moreover, we have proposed a new relation called *term_used_in* for wordnets through which students can consults terms that can be used in a certain textual genre. As future work, we plan to analyse other textual genres from the engineering domain and keep on adding terms to Basque WordNet and, thus, enriching it.

Acknowledgments

This work has been partially funded by the the projects DeepReading (RTI2018-096846-B-C21), CROSSTEXT (TIN2015-72646-EXP) and BigKnowledge – *Ayudas Fundación BBVA a Equipos de Investigación Científica 2018*.

References

Iñaki Alegría, Antton Gurrutxaga, Pili Lizaso, Xabier Saralegi, Sahats Ugartetxea, and Ruben Urizar. 2004. Linguistic and Statistical Approaches to Basque Term Extraction. *GLAT-2004: The Production Of Specialized Texts*.

Xabier Arregi, Ana Arruarte, Xabier Artola, Mikel Lersundi, and Igone Zabala. 2013. TZOS: An On-Line System for Terminology Service. *Centro de Lingüística Aplicada*, pages 400–404.

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the Wordnet Domains Hierarchy: Semantics, Coverage and Balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pages 101–108. Association for Computational Linguistics.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference*, volume 2016.

María Teresa Cabré. 1999. *La terminología: representación y comunicación*. Institut Universitari de Lingstica Aplicada. Universitat Pompeu Fabra.

María Teresa Cabré. 2003. Theories of terminology: their Description, Prescription and Explanation. *Terminology*, 9(2):163–199.

María Teres Cabré. 2005. Recursos lingüísticos en la enseñanza de lenguas de especialidad. In *V Jornada-Coloquio de la Asociación Española de Terminología (AETER): Comunicar y enseñar a comunicar el conocimiento especializado*.

Jose Camacho-Collados and Roberto Navigli. 2017. Babeldomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 223–228.

Iria da Cunha and M Amor Montane. 2019. Textual Genres and Writing Difficulties in Specialized Domains. *Signos*, 52(99):4–30.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Sandra Gollin-Kies, David R Hall, and Stephen H Moore. 2015. *Language for Specific Purposes*. Palgrave Macmillan.

Aitor González-Agirre, German Rigau, and Mauro Castillo. 2012. A Graph-based Method to Improve WordNet Domains. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 17–28. Springer.

Itziar Gonzalez-Dios. 2019. Nautikako terminologia biltzen testu-generoak abiapuntu: nabigazio-egunerokoen eredua. In Itziar Aduriz and Ruben Urizar, editors, *Hizkuntzalari euskaldunen III. topaketa. Zer berri?*, pages 79–91. Udako Euskal Unibertsitatea.

Maurizio Gotti. 2008. *Investigating specialized discourse*. Peter Lang.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of LREC*, pages 1413–1418.

Rita Marinelli, Adriana Roventini, and Alessandro Enea. 2004. Building a maritime domain lexicon: a few considerations on the database structure and the semantic coding. In *LREC 211 Fourth International Conference on Language Resources and Evaluation, held in Memory of Antonio Zampolli*. Citeseer.

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and Construction of the Basque WordNet. *Language resources and evaluation*, 45(2):121–142.

Marc J. Prohom. 2002. El uso de los diarios de navegación como instrumento de reconstrucción climática. *Investigaciones Geográficas*, 28:89–104.

Iranzu Sotés, Iñaki Alcedo, Imanol Basterretxea, Aingeru Basterretxea, and Xabier Sotés. 2015. *Itsasontziaren Eskuliburua*. Euskal Herriko Unibertsitateko Argitalpen Zerbitzua.

Igone Zabala, Izaskun Aldezabal, María Jesús Aranzabe, Jose Maria Arriola, Itziar Gonzalez-Dios, and Mikel Lersundi. 2018. Corpus-driven Terminology Work for Describing Basque Academic Terminology: the Weaving Terminology Networks programme (TSE programme). In *EAFT Summit*.