Basque e-lexicographic resources: linguistic basis, development, and future perspectives

Authors: Izaskun Aldezabal, Xabier Artola, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios, Gorka Labaka, German Rigau, Ruben Urizar

Institution: Ixa Group - University of the Basque Country (UPV/EHU)

email: itziar.gonzalezd@ehu.eus

In this work, we present the two major Basque lexico-semantic resources developed at the Ixa group[1]: 1) *Euskararen datu-base lexikala - Lexikoaren Behatokiaren datu-base lexikala* (Lexical Database of Basque - Lexical Database of the Lexical Observatory), a monolingual lexical database with morphological information and 2) *EuskalWN* or Basque WordNet, a knowledge base which collects word senses and sense relations in Basque linked to other languages and other resources in the Multilingual Central Repository[2] (Atserias et al., 2004). Specifically, we show how these resources were designed from a linguistic point of view, how they are maintained, which linguistic issues arise when new entries are coded, and future perspectives. It is important to note that, as the standarisation of Basque began officially 50 years ago, this recent and ongoing normalisation process makes it challenging to develop and maintain the e-lexicography resources.

Euskararen datu-base lexikala (EDBL) is a lexical database that was created in 1992 as a lexical support for the morphological analysis of Basque, but evolved into a general-purpose lexical database used for processing Basque texts in general (Aldezabal et al., 2001). In 2010, EDBL became *Euskararen datu-base lexikala - Lexikoaren Behatokiaren datu-base lexikala* (EDBL-LBDBL) when it was populated with dictionaries from the Elhuyar foundation[3] and the UZEI lexicographic center[4]. Nowadays, EDBL-LBDBL includes standard dictionary entries, non-standard variants linked to their standard equivalents, finite verb forms and other irregular inflected word forms, dependent morphemes, compounds, multi-word entries, abbreviations, etc. Each entry contains, apart from some morphological information, other interesting data such as whether it is included in the dictionary of the Academy of the Basque Language *Euskaltzaindiaren Hiztegia*[5] (the normative orthographic dictionary that now includes definitions). EDBL-LBDBL is updated whenever the Academy releases a new version of their online dictionary. At the moment (14th October 2018), the database consists of 135,062 entries, 113,682 of which are dictionary entries. The first application of EDBL-LBDBL was the spell checker *Xuxen* (Agirre et al., 1992), but it is mainly used for the automatic processing of Basque texts e.g. IxaKat (Otegi et al., 2016).

[1] www.ixa.eus/

[2] http://adimen.si.ehu.es/web/MCR

[3] https://www.elhuyar.eus/

[4] http://www.uzei.eus/

[5] https://www.euskaltzaindia.eus/index.php?option=com_hiztegianbilatu&view=frontpage&Itemid=410&lang=eu

Basque WordNet (BWN) is the Basque version of WordNet (Fellbaum, 1998) a.k.a. as Princeton Wordnet (PWN). BWN was created following the expand approach based on 1.6 version available at that moment and it was developed in two main steps: 1) translating the upper part and 2) developing BWN together with the sense-annotated corpus EPEC-EuSemcor (Pociello et al., 2011), the Basque reference corpus EPEC annotated with senses. As new versions are released, we update BWN following mainly automatic approaches: in the update to the 3.0 version (Gonzalez-Agirre et al., 2012), each version's synsets was matched one by one, and when multiple intersections collapsed to the same synset, the set of variants was joined into one synset. At the moment, BWN includes 30,697 synsets and 50,735 variants. It is used as the basis for UKB, the word-sense disambiguation tool for Basque (Agirre & Soroa, 2009).

Being EDBL-LBDBL a Basque monolingual morphosyntax-oriented database and BWN a semantics-oriented knowledge base, the procedure for inserting new entries is obviously different, but the referential Basque resources used and the lexicalisation issues concerning Basque word forms are the same in both. In contrast, in BWN it is necessary to face with the representation of concepts that are lexicalised in a language but not in the other.

As for EDBL-LBDBL, whenever a new version of *Euskaltzaindiaren Hiztegia* is available, it is automatically checked against the version in EDBL-LBDBL in order to detect changes. These changes include 1) adding new entries, 2) adding subentries, 3) changing the standardisation mark or level, and 4) deleting entries. For the new entries that have to be added to EDBL-LBDBL, if their PoS is given by *Euskaltzaindiaren Hiztegia,* some heuristics based on their phono-morphology are applied to get their corresponding information: lemma, two-level form... Then, the information obtained automatically is revised manually before the entries are incorporated into the database. Entries without assigned PoS are introduced manually.

As for BWN, the variants related to all the upper concepts in version 3.0 have been added. Moreover, we have concentrated on adding the variants of the Base Level Concepts (Izquierdo et al., 2007) and the general concepts (genlex) that are introduced to better organise the hierarchy and the epinonyms or semantic classes (Gómez Guinovart and Solla Portela, 2018). In order to incorporate the variants, we have followed the methodology defined by Pociello (2008), but this time using more referential dictionaries and corpora. Specifically, we have used the general-purpose Elhuyar dictionary[6], the Science and Technology dictionary[7], the terminology bank EuskalTerm[8] and the academic terminology database TZOS (Arregi et al., 2008), and the definitions of the already mentioned *Euskaltzaindiaren Hiztegia.* Regarding the corpora, we have used the general corpus *Lexikoaren*

[6] https://hiztegiak.elhuyar.eus/

[7] https://zthiztegia.elhuyar.eus/

[8] http://www.euskadi.eus/euskalterm/

Behatokia (Artola et al., 2017), the academic corpus Garaterm (Zabala et al. 2013) and the Science and Technology corpus (Areta et al., 2007). We have also used the Basque Wikipedia: its entries as dictionary and its texts as corpus.

In the case of single-unit variants and lexicalised multi-word variants, we proceeded as follows: for each variant in the PWN, 1) we looked for Basque variants in the above mentioned dictionaries and terminological databases; 2) we checked in the *Euskaltzaindiaren Hiztegia* that the sense of the Basque variant corresponds to the English one; 3) if correct, we added it to BWN. If the variant was not found in the reference dictionaries, we also used the variants of the Spanish and Galician wordnets as reference and pivot to get the Basque variants.

The major linguistic issues found when updating these resources are those referring to lexicalisation. In EDBL-LBDBL we deal with issues such as deciding which subentries from *Euskaltzaindiaren Hiztegia* deserve an entry. In principle, we should consider them as lexicalised, because, otherwise, they would not be treated such as, but, in some cases, considering their morphological composition, we do not see any special distinction/notation from the ones formed by simpler constituents. Those cases are e.g. lexical suffixes relating ordinals (*bi* 'two' -> *bigarren* 'second'), intensity markers (*hau* 'this one' -> *hauxe* 'just this one'), possessive pronouns (*ni* 'I' -> *nire* 'mine'), nouns used as postpositions or complementisers (*aurre* 'front' -> *aurrean* 'in front of'), words with many spatio-temporal case markers (*meza* 'mass' -> *mezan, mezatan, mezetan* 'in mass') or modal case markers (*hotz* 'cold' -> *hotzez* 'be/feel cold'), verbal nouns (*egin* 'do' -> *egite* 'doing'), causative verbs (*egin* 'do' -> *eginarazi* 'make someone do')... However, if the form that can be generated has a special or a specialised meaning, we do add these forms: for instance, *erdiratze* 'centering' is a verbal noun, but it is also a term in football. This way, we are also providing specialised vocabulary, although it is not specifically coded.

In BWN, however, we deal with issues related to the representation of concepts. As Pociello et al. (2011) pointed out, there are conceptual level imbalances and expression level imbalances when adding Basque variants. The former are related mainly to cultural concepts (Basque and foreign) while the latter are related to concepts that are known in both languages but differ in their PoS, in the need of numeral markers (Basque plural word form *altzariAK* vs English singular word form *furniture*) or inflection markers (Basque inflected *hotzez* vs non-inflected English cold), which sometimes coincide with those mentioned in EDBL-LBDBL. Other common problem is the representation of the gender, since there is no gender marking e.g. in many job titles...

To solve the representation conflict, we have systemised three operations when relating PWN to BWN synsets: 1) merging PWN synsets e.g. the synset [actor, histrion, player, thespian, role_player] and the synset [actress] should be only one synset in Basque [*aktore, antzezle, komediante, antzezlari*] since there is no gender marking, 2) splitting PWN synsets e.g. [terrorist_organization, terrorist_group, foreign_terrorist_organization], and 3) adding Basque synsets such as [*enbata*] (sudden rough weather in the Bay of Biscay and in the Cantabrian Sea).

Eventually, the representation of not lexicalised multi-word variants also has to be dealt with. We have defined this procedure in order to cover the multi-word variants in PWN that were not found in Basque dictionaries: 1) we manually create variant proposals based on translations of each unit; 2) we look for them in the corpora; 3a) if found, we add each multi-word variant to BWN with a special label (ixalex) in order to mark that it is not fully lexicalised e.g. *animalia-birus* 'animal_virus'; 3b) if not, we label the synset as non-lexicalised (nonlex) e.g. craniometic_point.

Future challenge of both resources involves the incorporation and codification of Basque terminology. In BWN, a first systemised attempt was done with WNTERM (Pociello et al., 2008) by incorporating terms from the above-mentioned Science and Technology dictionary, but in some cases it is difficult to decide its place in BWN because words are too specialised. In order to overcome this problem, we are exploring to use TZOS (Arregi et al., 2013), where professors and lecturers group the terms used in their subjects into semantic classes. However, this work is in its beginnings. We have also worked on nautical terminology by adding the terms found in the logbooks (Gonzalez-Dios, 2017). The main problem when incorporating terms to BWN is that their English equivalents are not yet in PWN. So, we are preparing an experiment to use CILI (Bond et al. 2016). Right now, we are also analysing the possibility to add the terms that are frequently used in most of the science areas, in both EDBL-LBDBL and BWN.

Finally, we are also working on semi-automatic approaches to update the resources. For example, in order to detect in EDBL-LBDBL proper names with spellings that are no longer considered standard, we use similarity measures to compare the entries in the database with those in a standardised list. This way we can find pairs of entries referring to the same entity such as *Philadelphia* and **Filadelfia*. In WordNet, we test black-box techniques by cross-checking different ontologies in order to detect knowledge discrepancies (Álvez et al., 2017).

Both EDBL-LBDBL and BWN are available at our website.[9] EDBL-LBDBL has non-commercial license and can be accessed though its interface; BWN as part of the MCR has CC BY license and can be accessed by means of two different graphical interfaces. BWN is also available at the LLOD cloud.[10]

ACKNOWLEDGMENTS

[9] https://ixa.si.ehu.es/produktuak?language=en
[10] http://linguistic-lod.org/llod-cloud

This work is supported by the following projects: TUNER: Automatic domain adaptation for semantic processing (TIN2015-65308-C5-1-R), funded by Spanish Ministry of Economy, and Lexikoaren Behatokia IX, funded by Euskaltzaindia (Academy of the Basque Language).

REFERENCES

Aduriz, I., Agirre, E., Alegria, I., Arregi, X., Arriola, J.M., Artola, X., Díaz de Ilarraza, A., Ezeiza, N. Maritxalar, M., Sarasola, K. & Urkia, M. (1992). A Morphological Analyzer for Basque on Two-level Morphology. Proceedings of the 5th Int. Morphology Meeting. Austria.

Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Maritxalar, M., Sarasola, K. & Urkia, M. (1992). XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology. In *Proceedings of the third conference on Applied natural language processing* (pp. 119-125).

Agirre, E., & Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 33-41).

Aldezabal, I., Ansa, O., Arrieta, B., Artola, X., Ezeiza, A., Hernández, G., & Lersundi, M. (2001). EDBL: A general lexical basis for the automatic processing of Basque. In *Proceedings of the IRCS Workshop on linguistic databases*. IRCS Workshop on linguistic databases.

Álvez, J., Lucio, P., & Rigau, G. (2017). Black-box Testing of First-Order Logic Ontologies Using WordNet. *arXiv preprint arXiv:1705.10217*.

Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Diaz de Ilarraza, A., Ezeiza. N.., & Sologaistoa, A. (2007). ZT Corpus: Annotation and tools for Basque corpora. *Proceedings of Corpus Linguistics 2007* (pp. 1-19).

Arregi, X, Arruarte, A., Artola, X., Lersundi, M., & Zabala, I. (2013). TZOS: An On-Line System for Terminology Service. *Centro de Lingüística Aplicada*, 400-404.

Artola, X., Ezeiza, N., Gurrutxaga, A, Sagarna, A., & Urkia, M. (2017). Lexikoaren Behatokia: leiho bat XXI. mendeko hedabideetako euskarari. *Senez: itzulpen aldizkaria*, (48), 16.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., & Vossen, P. (2004). The meaning multilingual central repository. In Proceedings of the Second International Global WordNet Conference (GWC'04).

Bond, F., Vossen, P., McCrae, J. P., & Fellbaum, C. (2016). CILI: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference* (pp. 50-57).

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT press.

Gómez Guinovart, X., & Solla Portela, M. A. (2018). Building the Galician wordnet: methods and applications. *Language Resources and Evaluation*, *52*(1), 317-339.

Gonzalez-Agirre, A., Laparra, E., & Rigau, G. (2012). Multilingual central repository version 3.0. In *Proceedings of* 8th international conference on Language Resources and Evaluation (*LREC)* (pp. 2525-2529).

Gonzalez-Dios, I. (2017). Nautikako terminologia biltzen testu-generoetan oinarrituta: nabigazio-egunerokoen kasua abiapuntu gisa. Communication at Hizkuntzalari Euskaldunen III. Topaketa: Zer berri?. Baiona. Udako Euskal Unibertsitatea.

Izquierdo, R., Suárez, A., & Rigau, G. (2007). Exploring the automatic selection of basic level concepts. In *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP)* (pp. 1-8).

Otegi, A., Ezeiza, N., Goenaga, I., & Labaka, G. (2016). A Modular Chain of NLP Tools for Basque. In *International Conference on Text, Speech, and Dialogue* (pp. 93-100). Springer.

Pociello, E. (2008) . Euskararen ezagutza-base lexikala: Euskal WordNet. PhD thesis, University of the Basque Country (UPV/EHU).

Pociello, E., Agirre, E., & Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. *Language resources and evaluation*, *45*(2), 121-142.

Pociello, E., Gurrutxaga, A., Agirre, E., Aldezabal, I., & Rigau, G. (2008). WNTERM: Enriching the MCR with a Terminological Dictionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluations (LREC),* (pp. 1778-1784).

Zabala, I., Lersundi, M., Leturia, I., Manterola, I., & Santander, G. (2013). GARATERM: euskararen erregistro akademikoen garapenaren ikerketarako lan-ingurunea. *Ugarteburu terminologia jardunaldiak (V). terminologia naturala eta terminologia planifikatua euskararen normalizazioari begira*, 98-114.