Bilingual Dictionary Drafting: Connecting Basque Word Senses to Multilingual Equivalents

David Lindemann, Iñaki San Vicente

UPV/EHU University of the Basque Country, Elhuyar Foundation
e-mail: david.lindemann@ehu.eus, i.sanvicente@elhuyar.com

Abstract

This paper presents a simple method for drafting bilingual dictionary content using existing lexical and NLP resources for Basque. The method consists of five steps, three belonging to a semi-automatic drafting, and another two to semi-automatic and manual post-editing: (1), the building of a corpus-based frequency lemma list; (2) the drafting of syntactical entities belonging to a lemma-sign; (3) the drafting of word senses belonging to syntactical entities; (4) a semi-automatic detection of gaps regarding syntactical entities, and (5) manual detection of word sense gaps. The described method relies on the exploitation of existing resources for Basque, and the multilingual cross-references present in WordNet. The application of the described method follows two goals: (1) drafting of a series of bilingual dictionaries with Basque, and (2) a contribution to the updating and enrichment of two Basque NLP resources used for the drafting, EDBL and EusWN.

Keywords: bilingual lexicography; dictionary drafting; basque language; WordNet

1 Introduction

This paper is a follow-up on Lindemann et al. (2014), where we presented a set of methods for a semi-automatic drafting of translation equivalents for German-Basque, a medium-density language pair. The methods presented there rely on the exploitation of parallel corpora, the data found in Wikipedia and Wiktionary, an approach that links Basque and German lexical units using bilingual dictionaries with English as pivot, and an extraction of lexical correspondences using WordNet interlanguage indexes. We conclude that in spite of German-Basque being a medium-density language pair with limited existing resources data to be extracted from, by the combined set of methods we obtain one or more adequate translation equivalents for around 60% of the corpus-based German frequency lemma list we use as reference. On the other hand, the obtained bilingual dictionary draft does not consider any distinction between homographs with different Part of Speech (PoS) or Syntactical Entity Disambiguation (SED) nor a polysemy structuring by Word Sense Disambiguation (WSD). In this paper, we present a method for drafting SED and WSD for Basque lemma-signs and multilingual translation equivalents, using the existing NLP resources EDBL and the Basque WordNet, EusWN, which carries interlanguage links to WordNets of other languages by means of the English Princeton WordNet (PWN) as pivot.

2 Existing Resources

For the experiments described in this paper, we had access to the following Basque lexical and NLP resources:

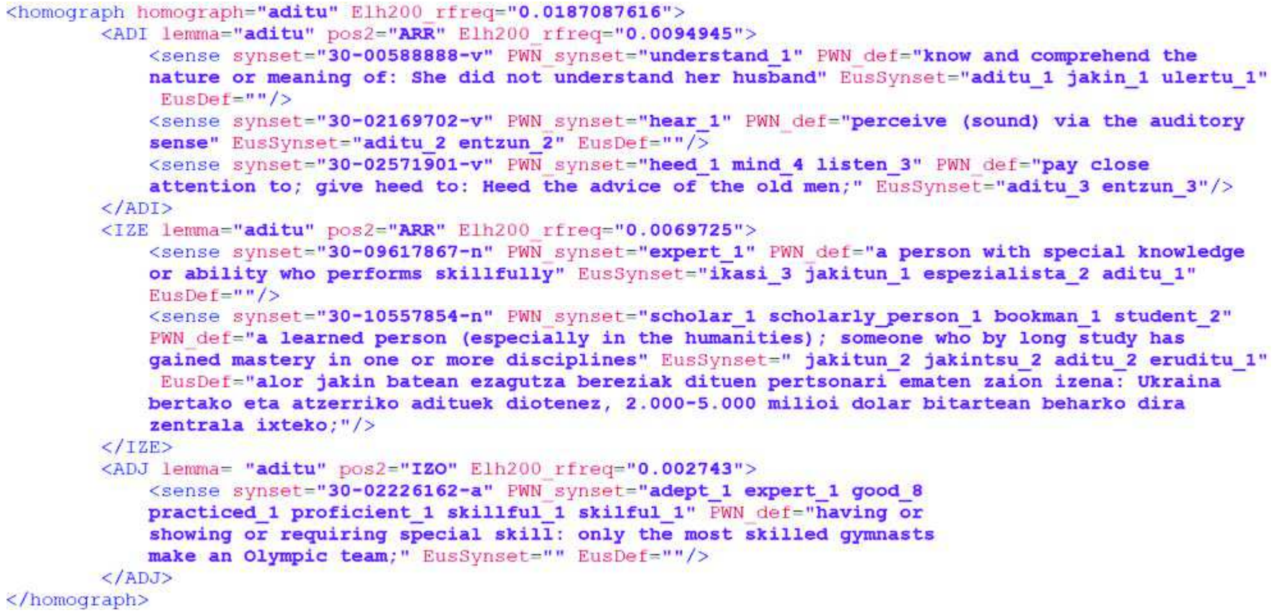
- The *Euskararen Datu Base Lexikala* ('Basque Lexical Database', EDBL) (Aldezabal et al. 2001, among others). EDBL was designed as data source for a series of NLP tools, such as syntactical taggers, lemmatizers, morphological analysers, and spell-checkers. EDBL contains canonical forms (lemmata) as well as inflected forms, and non-free morphemes. Of the 84,355 forms in EDBL, 64,737 are lemmata.
- *EusTagger* (Aduriz et al. 1996).
- *EuskalWordNet* (EusWN, version 3.0) (Pociello 2007; Pociello et al. 2011). The Basque lexical items present in this resource are a part of synsets which are linked to the English Princeton WordNet (PWN, Fellbaum 1998). EusWN contains the ID-codes of all synsets found in PWN version 3.0 which link to 49.894 Basque lexical items (word senses) which belong to 27.116 lemma-signs (homographs), and another 1.053 lexical items not linked to PWN synsets.
- *Elhuyar Web Corpus* (Elh200) (Leturia 2014), a general corpus built by means of fully automatic web crawling methods, containing 200 million tokens.
- *Egungo Testuen Corpusa* (ETC) (Sarasola et al. 2013), a corpus developed at the Basque Language Institute at UPV/EHU. In its 2013 version, this corpus counts 204.9 million tokens. It contains Basque press, literature, science and television broadcast texts that were selected by hand, and the Basque Wikipedia.
- A Basque corpus-based frequency lemma list (EuLemStd) which we propose as starting point for manually edited bilingual standard dictionaries with Basque (Lindemann & San Vicente 2015).
- Basque lemma lists of the following standard reference dictionaries: *Orotariko Euskal Hiztegia* (Mitxelena & Sarasola 1988), *Hiztegi Batua* (Euskaltzaindia 2010), *Euskal Hiztegia* (Sarasola 1996), *Elhuyar EU-ES* (Elhuyar Hizkuntza Zerbitzuak 2013).

3 Data Modelling

We build a single XML document containing the dictionary draft. On the first hierarchy level, that is, in lexicographical terminology, on macrostructure level, the entries of a corpus-based frequency list of lemma-signs for Basque (Lindemann & San Vicente 2015) are represented. In total, 53,311 one word and multiword lemma-signs reaching a minimum occurrence threshold (occurrence threshold >=20) in one of the two big monolingual Basque Corpora Elh200 or ETC, and appearing in EDBL are represented at this level as <*homograph*> elements.

As child elements of those, we list the Syntactical Entities (SE, that is, lemmata with a Part of Speech assigned) that belong to a lemma-sign as sibling elements named according to their PoS, that is in a Basque abbreviated form, noun (IZE), proper noun (IZB), verb (ADI), adjective (ADJ) and adverb (ADB), among others. These Syntactical Entities have been extracted from Elh200 corpus using *EusTagger*, a lemmatizer and PoS-tagger for Basque that uses EDBL as parameter file. Homographs and SE both are furnished with corpus frequency data.

On a third level, that is, as child elements of the respective SE, we include Basque word senses from EusWN 3.0 as <*sense*> elements siblings to each other, together with the data available for the corresponding synset. Figure 1 shows an example of a joint dataset built in the described way, here with synonyms and, where available, definitions from Basque and English WordNets.



```
<homograph homograph="aditu" Elh200_rfreq="0.0187087616">
        <ADI lemma="aditu" pos2="ARR" Elh200_rfreq="0.0094945">
            <sense synset="30-00588888-v" PWN_synset="understand_1" PWN_def="know and comprehend the
            nature or meaning of: She did not understand her husband" EusSynset="aditu_1 jakin_1 ulertu_1"
             EusDef=""/>
            <sense synset="30-02169702-v" PWN_synset="hear_1" PWN_def="perceive (sound) via the auditory
            sense" EusSynset="aditu_2 entzun_2" EusDef=""/>
            <sense synset="30-02571901-v" PWN_synset="heed_1 mind_4 listen_3" PWN_def="pay close
            attention to; give heed to: Heed the advice of the old men;" EusSynset="aditu_3 entzun_3"/>
        </ADI>
        <IZE lemma="aditu" pos2="ARR" Elh200_rfreq="0.0069725">
            <sense synset="30-09617867-n" PWN_synset="expert_1" PWN_def="a person with special knowledge
            or ability who performs skillfully" EusSynset="ikasi_3 jakitun_1 espezialista_2 aditu_1"
            EusDef=""/>
            <sense synset="30-10557854-n" PWN_synset="scholar_1 scholarly_person_1 bookman_1 student_2"
            PWN_def="a learned person (especially in the humanities); someone who by long study has
            gained mastery in one or more disciplines" EusSynset=" jakitun_2 jakintsu_2 aditu_2 eruditu_1"
             EusDef="alor jakin batean ezagutza bereziak dituen pertsonari ematen zaion izena: Ukraina
            bertako eta atzerriko adituek diotenez, 2.000-5.000 milioi dolar bitartean beharko dira
            zentrala ixteko;"/>
        </IZE>
        <ADJ lemma= "aditu" pos2="IZO" Elh200_rfreq="0.002743">
            <sense synset="30-02226162-a" PWN_synset="adept_1 expert_1 good_8
            practiced_1 proficient_1 skillful_1 skilful_1" PWN_def="having or
            showing or requiring special skill: only the most skilled gymnasts
            make an Olympic team;" EusSynset="" EusDef=""/>
        </ADJ>
</homograph>
```

Figure 1: SE extracted from corpora and WordNet senses as XML.

The word senses from EusWN are mapped to EDBL Syntactical Entities using PoS information of the synsets, consisting of the last character of each synset ID. For a distinction between common and proper nouns, which have their own PoS category in corpora, the initial case of the corresponding Basque lexical units in EusWN is analysed (proper nouns are initial upper case). Verbs and adjectives are mapped in a straightforward way, as shown in Table 1. Adverbs are not considered, as EusWN as of version 3.0 does not contain any adverbial lexical items.

EusWN POS-tag	**EDBL POS-tag**	**POS-tag dictionary draft**
n [and lexical unit initial lower case]	IZE_ARR	IZE
n [and lexical unit initial upper case]	IZE_LIB/IZB	IZB
V	ADI	ADI
A	ADJ	ADJ
R	ADB	ADB

Table 1: Mapping of POS-tags.

Due to the fact that the Basque lexicon contains a relatively high amount of homograph lemmata (around 5% of Basque lemma-signs correspond to more than one SE), a SED step is desired, as fine-grained as possible, for a dictionary draft. Corpus frequency data for every homograph SE is also a very valuable feature.

The XML <*sense*> elements containing the word senses extracted from EusWN are then completed with Basque synonyms and multilingual equivalents using the interlanguage links existent in the WordNet family using Princeton WordNet as a pivot, as it has been proposed and carried out in several projects (cf. Vossen 2004; Varga et al. 2009; Navigli & Ponzetto 2010; Lorentzen & Trap-Jensen 2011; Tavast et al. 2012, among others).

Nowadays, WordNets for a large range of languages are available. In the context of a bilingual lexicography project involving Basque, a minority language, this is particularly valuable, because there are not more than about half a dozen bilingual dictionaries that have been edited professionally, all of them involving major languages. Regarding all the other languages, and lacking suitable lexicographical products, a Basque dictionary user still needs to rely on bilingual dictionaries with, for example, Spanish, and a foreign user who wants to translate from or into Basque will also need a third language as pivot. Multilingual dictionary drafts based on WordNet may boost the edition of new dictionaries. This is exactly the motivation of the authors, together with contributing to an enrichment of the Basque WordNet.

4 Automatic SED Gap Detection

From the Elh200 Basque corpus tagged with *EusTagger*/EDBL we have extracted frequency data for 112,885 noun, proper noun, verb, adjective and adverb SE that correspond to 93,081 lemma-signs, setting for the SE a minimum threshold of 20 corpus occurrences. An example is shown in Table 2 for the Basque lemma sign *alegia*.

EDBL-PoS	frequency list rank	occurrence counts	relative frequency	PoS
LOK	618	41,106	0.020553	conjunction ['alegia']
IZE_ARR	3,882	4,208	0.002104	noun ['alegia']
IZE_LIB	10,407	921	0.0004605	place name ['Alegia']

Table 2: The Basque lemma-sign *alegia* in Elh200 corpus.

After merging EDBL with EusWN data, 14.709 of 93.081 noun, verb and adjective SE have been mapped to one or more EusWN word senses. Having in mind that the merged data unlike EusWN 3.0 also contain adverbs, we can identify 1.865 candidate SE for enriching EusWN. In total, 78.372 homographs that may appear as nouns, verbs or adjectives have not been mapped to any EusWN synset.

16.810 of these appear on our basic reference lemma list EuLemStd, and can thus be identified as a first group of candidates for an enrichment of EusWN. In addition, in 2.188 cases one or more than one SE without mapping to EusWN are homograph to a SE that has obtained a linking, as it occurs in the following example: for the aforementioned Basque lemma-sign *aditu* (Figure 1) we find three word senses for the verb, and another two for the homograph noun. The adjective *aditu*, however, is not found in EusWN 3.0, in spite of being counted 5.486 times in Elh200 corpus. Here we have an automatically obtained SE gap in EusWN, which, in this case, is to be filled manually with WordNet data, as a synset (still empty for Basque) that fits into this gap already exists in PWN. The Basque synonym *aditu* can be added to the Basque counterpart of that synset (which will carry the same ID). The synset and data relevant to it are then pasted into the dictionary draft XML as done in the manually enriched dataset for the adjective *aditu* shown in Fig. 1 above.

5 Manual WSD Gap Detection

In Table 3, we see data from WordNet belonging to six word senses of the Basque noun *adar*. The seventh one has been added manually after comparing the sense list to the polysemy listed for the noun *adar* in other Basque dictionaries and having detected a sense gap. In this case, English PWN already contained a suitable concept (synset) for the sense, with one Basque equivalent lexical unit listed: *zapata-zartzeko*, which is a terminus tecnicus for what people call a *shoehorn*, or, according to *Hiztegi Batua* used here as a reference, simply *adar* (*horn*) in Basque. The Catalan data in Table 3 has been added to show the simplicity of this multilingual dictionary drafting approach. In case the corresponding catalan synset contains any lexical units, we have hereby extended the dictionary draft to Catalan.

The detection of this kind of polysemy-gaps will be carried out manually, although automatic methods may help. As soon as the content of Basque dictionaries is fully parsed and digitally accessible, which is an ongoing process, quantitative comparisons of word senses per lemma-sign and/or per SE will be possible, and outstanding mismatches of polysemy-numbering between the resources may be highlighted.

EusWN Lexical Unit	**Definition EN**	**EU synset**	**EN synset**	**CAT synset**
adar_1	one of the bony outgrowths on the heads of certain ungulates	adar_1	horn_2	banya_1
adar_2	a railway line connected to a trunk line	adar_2	branch_line_1 spur_track_1 spur_5	enforcall_1 forcall_1
adar_3	a warning signal that is a loud wailing sound	adar_3, sirena_2 turuta_5	siren_3	
adar_4	a local branch of some fraternity or association	adar_4	chapter_3	capítol_2
adar_5	a division of a stem, or secondary stem arising from the main stem of a plant	adar_5 abar_2 besanga_1 beso_12	branch_2	branca_1 branc_1
adar_6	an alarm device that makes a loud warning sound	sirena_4 adar_6 turuta_6	horn_9	
adar_7	a device used for easing the foot into a shoe	zapata_sartzeko_1	shoehorn_1	calçador_1

Table 3: The Basque noun "*adar*" and multilingual equivalents from WordNet.

6 Conclusions and Further Work

In the presented experiments, we have treated large amounts of Basque lexical data. We have automatically built a dictionary draft for 93,081 lemma-signs extracted from a very large web corpus, including SE discrimination and frequency data for each SE. 30,478 of these lemma-signs belong to our previously defined basic dictionary lemma list for Basque (EuLemStd) which counts 53,311 entries (Lindemann & San Vicente 2015).[1] Table 4 presents the statistics of all drafted SE and word senses linked to WordNet, while table 5 shows the figures regarding only lemma-signs included in EuLemStd.

[1] As pointed out by the authors, those lemma-signs will serve as starting point for manual dictionary edition. See the referenced paper for criteria regarding the definition of further headword candidates, such as neologisms present in the corpora but not in the reference resources, and a description of the methodology and sources for the development of that lemma list.

	SE extracted from corpus	EusWN lexical items	EusWN Word senses	Lemma-signs present in both EusWN and corpora
Verbs	7,723	3,399	9,340	1,452
Common Nouns	40.403	22,853	39,523	12,369
Proper Nouns	34,144	808	885	0
Adjectives	23,509	57	148	42
Adverbs	1,856	0	0	0

Table 4: Drafted lexical data, overall outcomes.

Another set of figures to point out is the distribution of WordNet word senses. A total amount of 37,474 word senses belongs to 24,072 unique synsets or concepts. Those, in turn, correspond to 16.827 unique lemma-signs (homographs) present on the reference lemma list EuLemStd. A single SE belonging to this group of lemma-signs may have been furnished with one or more senses (monosemy vs. polysemy) in a range up to 52 senses a SE. The average amount of senses per SE is 2.2 (se detailed figures in Table 5).

On the other hand, the average amount of SE per lemma-sign (homography ratio) furnished with one or more WordNet senses is 1.01; homography applies to not more than 1.1% of the reference lemma-list entries, while the overall homography rate (including also the SE identified automatically in the corpora but not present in WordNet) is about 12%. As mentioned before, in other Basque reference resources the homography rate rounds 5%; the difference is to be explained mainly by automatic POS-tagging mistakes. In order to filter such noisy data, automatic comparison to reference resources and a posterior manual revision of the lexical data on hand is being carried out.

	Corpus-based SE	SE with one or more EusWN Word senses	Total EusWN Word senses	Polysemy ratio	SE present in corpus but not in EusWN	SE present in EusWN but not found in corpus
Verbs	4,151	1,636	6,567	2.01	2,515	279
Common Nouns	23,921	15,193	30,613	4.01	8,728	3,479
Proper Nouns	2,443	132	153	1.16	2,311	60
Adjectives	6,147	50	141	2.82	6,097	8
Adverbs	1,556	0	0	0.00	1,556	0
Total	38,218	17,011	37,474	2.20	21,207	3,826

Table 5: Drafted lexical data regarding previously defined lemma-list (EuLemStd) only.

As we have pointed out, the frequency lemma-list based dictionary draft contains links to 24.072 WordNet concepts. In those cases, the creation of a multilingual dictionary draft may be carried out immediately by means of resources such as the Multilingual Central Repository (Gonzalez-Agirre et al. 2012), where WordNets in various languages are connected by common synset-IDs, as mentioned in chapter 2 and shown in table 3.

Furthermore, we have obtained a list of 2.188 lemma-signs present in EusWN but which show one or more gaps at SE level in that resource. We consider this group of headwords as first-level candidates for an enrichment of EusWN. As for SE present in EusWN but not found in the corpus, we are about to develop a case typology. A first insight into the data reveals a prominent presence in this group of technical terms which are translations of their English counterparts, and substantive derivations of Basque verbs, which in EDBL (and in Basque lexicography in general) do not appear as headword. Lastly, it must be noted that merging the aforementioned lexical and NLP resources has led to a 32% covering of the initially proposed Basque lemma list with at least one SE *and* at least one word sense. EusWN is a manually edited resource and thus shows a very low error rate. Our next step is to study the possibility of including automatically built multilingual resources, too, in order to increase the coverage. We have turned our efforts towards BabelNet (Navigli & Ponzetto 2010), a resource that for Basque includes lexical data and multilingual equivalents extracted from WordNet, Wikipedia, Wiktionary, and Omegawiki. Such a resource will of course require a manual qualitative evaluation as the one carried out before for German-Basque (Lindemann et al. 2014).

7 References

Aduriz, I., Aldezabal, I., Alegria, I., Artola, X., Ezeiza, N. & Urizar, R. (1996). EUSLEM: A lemmatiser/tagger for Basque. In *Proceedings of EURALEX 1996*. Göteborg: Göteborg University, pp. 17–26.

Aldezabal, I., Ansa, O., Arrieta, B., Artola, X., Ezeiza, A., Hernandez, G. & Lersundi, M. (2001). EDBL: a General Lexical Basis for the Automatic Processing of Basque. In *Proceedings of the IRCS Workshop on linguistic databases*. Philadelphia.

Elhuyar Hizkuntza Zerbitzuak (2013). *Elhuyar hiztegia: euskara-gaztelania, castellano-vasco* 4. ed. Usurbil: Elhuyar

Euskaltzaindia (2010). *Hiztegi batua*. Donostia: Elkar

Gonzalez-Agirre, A., Laparra, E. & Rigau, G. (2012). Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the Sixth International Global WordNet Conference (GWC'12)*. Matsue, Japan.

Leturia, I. (2014). *The Web as a Corpus of Basque*. PhD Thesis. Donostia: UPV-EHU Lengoaia eta Sistema Informatikoak Saila.

Lindemann, D. & San Vicente, I. (2015). Building Corpus-based Frequency Lemma Lists. *Procedia - Social and Behavioral Sciences*, 198: 266–277.

Lindemann, D., Saralegi, X., San Vicente, I., Manterola, I. & Nazar, R. (2014). Bilingual Dictionary Drafting. The example of German-Basque, a medium-density language pair. In *Proceedings of the XVI Euralex International Congress*. Bolzano: EURAC, pp. 563–576.

Lorentzen, H. & Trap-Jensen, L., (2011). There And Back Again – from Dictionary to Wordnet to Thesaurus and Vice Versa: How to Use and Reuse Dictionary Data in a Conceptual Dictionary. In *Proceedings of eLex*., pp. 175–179.

Mitxelena, K. & Sarasola, I. (1988). *Diccionario general vasco - Orotariko euskal hiztegia*. Euskaltzaindia; Desclée de Brouwer

Navigli, R. & Ponzetto, S. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics (ACL '10), pp. 216–225.

Pociello, E. (2007). *Euskararen ezagutza-base lexikala: Euskal WordNet*. PhD Thesis. Donostia: UPV-EHU Euskal Filologia Saila.

Pociello, E., Agirre, E. & Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 452, pp. 121–142.

Sarasola, I. (1996). *Euskal Hiztegia*. Donostia: Kutxa Gizarte-eta Kultur Fundazioa

Sarasola, I., Landa, J. & Salaburu, P. (2013). Egungo Testuen Corpusa. UPV/EHU, Euskara Institutua

Tavast, A., Muischnek, K. & Koit, M. (2012). Cross-linking Experience of Estonian WordNet. In *Human Language Technologies: The Baltic Perspective. Proceedings of the Fifth International Conference Baltic HLT 2012*. IOS Press, pp. 96–102.

Varga, I., Yokoyama, S. & Hashimoto, C. (2009). Dictionary generation for less-frequent language pairs using WordNet. *Literary and Linguistic Computing*, 244, pp. 449–466.

Vossen, P. (2004). Eurowordnet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected Via an Inter-Lingualindex. *International Journal of Lexicography*, 17(2), pp. 161–173.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 613465, and from the Basque Government (BFI-2011-1, IT665-13 and Elkarola (KK-2015/00098, KK-2016/00087)).