Bilingual Dictionary Drafting: Bootstrapping WordNet and BabelNet

David Lindemann[1,2], Fritz Kliche[2]

[1] The Bilingual Mind, UPV/EHU University of the Basque Country,
Justo Vélez de Elorriaga 1, 01006 Vitoria-Gasteiz (Spain)
[2] IwiSt Institute for Information Science and Natural Language Processing,
Universität Hildesheim, Universitätsplatz 1, 31141 Hildesheim (Germany)
E-mail: david.lindemann@ehu.eus, fritz.kliche@uni-hildesheim.de

Abstract

In this paper, we present a simple method for drafting sense-disambiguated bilingual dictionary content using lexical data extracted from merged wordnets, on the one hand, and from BabelNet, a very large resource built automatically from wordnets and other sources, on the other. Our motivation for using English-Basque as a showcase is the fact that Basque is still lacking bilingual lexicographical products of significant size and quality for any combination other than with the five major European languages. At the same time, it is our aim to provide a comprehensive guide to bilingual dictionary content drafting using English as pivot language, by bootstrapping wordnet-like resources; an approach that may be of interest for lexicographers working on dictionary projects dealing with other combinations that have not been covered in lexicography but where such resources are available. We present our experiments, together with an evaluation, in two dimensions: (1) A quantitative evaluation by describing the intersections of the obtained vocabularies with a basic lemma list of Standard Basque, the language for which we intend to provide dictionary drafts, and (2) a manual qualitative evaluation by measuring the adequateness of the bootstrapped translation equivalences. We thus compare recall and precision of the applied dictionary drafting methods considering different subsets of the draft dictionary data. We also discuss advantages and shortcomings of the described approach in general, and draw conclusions about the usefulness of the selected sources in the lexicographical production process.

Keywords: Bilingual Lexicography; Bilingual Dictionary Drafting; WordNet; BabelNet

1. Computational lexicography and WSD in multilingual settings

1.1 Starting Point

According to *Ethnologue* data, more than 400 languages have one million or more first-language speakers. If we check the availability of bilingual dictionaries for these languages, we observe that many language pairs, even those involving one of the top ten languages of the world, remain uncovered. For Basque, for instance, a European language with about one million speakers, bilingual dictionaries of significant size and

quality are available today for Spanish, French, English, Russian, and German.

Lacking suitable lexicographical resources for all other language pairs, a dictionary user may follow two main strategies: they may use more than one bilingual dictionary, i.e. retrieve the desired information via hub, and thus perform double lookups or trust their knowledge in the hub language. This we may call the 'traditional' approach. Alternatively, they may also rely on automatically built bilingual dictionary-like resources for the required language pair, or place their query in machine translation backed web portals that work with automated algorithms and use English as a hub.

For the first case, there are a number of disadvantages linked to the required availability of the respective dictionaries, and to the disposition to spend the required time for multiple lookups in one process of lexical information retrieval. Its ease and its efficacy for the user is what makes the second strategy appealing.

One fundamental problem applies to both strategies. Mistakes in the retrieval of translation equivalents due to lexical semantics issues, and different distributions in the lexicalization of concepts between languages, are doubtlessly frequent, and discussions regarding *asymmetric lexicalization* are thus a real classic in metalexicographical writing (for the cited concept, see Hartmann, 1990; cf. also Wiegand, 2002; Gouws, 2002). Furthermore, if two different bilingual dictionaries are needed for looking up possible equivalents, the risk of being misled may also be doubled.

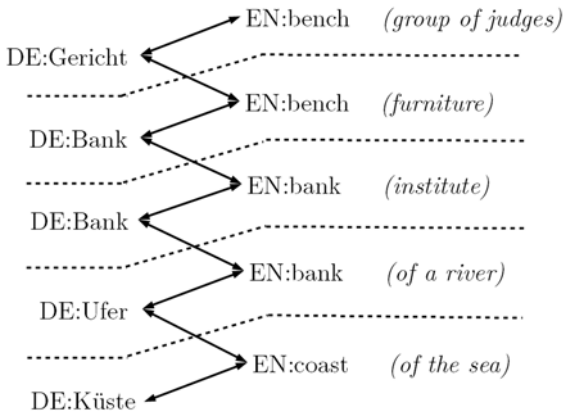

Figure 1: Asymmetric lexicalization of concepts

Asymmetric lexicalization can be illustrated by the examples given in Figure 1, where arrowed lines represent possible translation equivalences between senses that correspond to the lemma-strings preceded by the German or English language code, and dotted lines divide concepts; glosses are given in brackets to disambiguate concepts. Arrows that cross dotted lines represent mismatched translation equivalences that erroneously seem possible according to the character strings they link to each other. Without further information (dotted lines and glosses), all the equivalences between lexical items represented here are equally possible. The inventory

of senses shown here is far from complete, and the game could be continued (for example, *Gericht* may also mean an edible 'dish', while *dish* in turn also may denote a vessel used for serving food, which in German can be called 'Geschirr', which is an item that also may denote horse or ox harnesses, etc.).

The figure also does not show distinctions between (1) homonomy (German $Bank_1$ vs. $Bank_2$, English $bank_1$ vs. $bank_2$, (2) polysemy ($bench_1$ vs. $bench_2$), and (3) a splitting of senses, which is not necessary for a German monolingual but is necessary for a German-English bilingual dictionary entry (*Ufer*, 'egde' of a river vs. of the sea, a lemma that in German monolingual dictionaries is usually not marked as polysemous). Here, we see only text strings annotated as nouns of a certain language; the required condition for the mismatched equivalences to occur. A good bilingual dictionary of course provides the user with useful homonym or sense disambiguating advice in order to avoid such misleading pairings.

Problems related to a look-up process misled by asymmetric polysemy structures also may apply to the second case; in fact, this is the main shortcoming observed when employing algorithms that interlink entries of two bilingual dictionaries, using their shared language as hub (for example, as stated by Saralegi et al., 2012). Also, in parallel corpus processing, the semantic disambiguation of polysemous lexical items (WSD) has still to be regarded as a central problem; users who lack a suitable bilingual dictionary and thus stick to statistical machine translation engines, must deal with errors related to homonymy and polysemy in the results they obtain.

1.2 Bilingual Dictionary Drafting Methods: A Brief Overview

If we thus decide to develop lexicographical resources for new language pairs in order to overcome these shortcomings, we can employ 'traditional' methods: namely, the manual compilation of bilingual dictionaries starting from scratch. However, this is a very labour-intensive task, only feasible for lexicographical products that satisfy commercial criteria (which is definitely not the case for dictionaries of a language such as Basque) or grow in publicly well-funded environments. To reduce the level of manual effort required for bilingual dictionary making, a further development of (semi-)automatic dictionary drafting methods seems worthwhile.

For a rough classification of (semi-)automatic methods to obtain bilingual word pairs as candidates for an inclusion as translation equivalents into a bilingual dictionary (see Varga et al., 2009 for a survey of related work), we can distinguish between corpus-based methods on the one hand, and methods that rely on transferring data from existing lexical resources, on the other. Both approaches may be combined, e.g. as in Saralegi et al. (2012), where the equivalent pairs obtained by linking the content of two bilingual dictionaries are ranked according to distributional similarity in a bilingual text corpus.

In addition, we can group translation equivalent drafting methods according to the following qualitative feature: whether it results in bilingual word lists, i.e. lists of equivalent candidates, or whether it is capable of linking word-sense disambiguated lexical items to each other, i.e. of linking word senses, for a bilingual dictionary draft that includes WSD. Bilingual data found in the WordNet-related multilingual lexical resource MCR 3.0 (Gonzalez-Agirre et al., 2012), as shown in Table 1, demonstrates, for the instances of the noun *banku* in Basque WordNet, how equivalences may be extracted from this kind of resource and including a discrimination of word senses, i.e., a grid that avoids mismatches of the kind illustrated in Figure 1.

Basque Synset	**English Synset**	**MCR ontology classes**
banku_1; banketxe_1;	bank_9; bank_building_1	banking; artifact_1; artifact; Building+; Artifact+ Function+ Object+
aulki_3; banku_2;	bench_1	furniture; furniture_1; artifact; Artifact+; Artifact+ Furniture+ Group+ Instrument+ Object+
banku_3; banketxe_2;	depository_financial_institution_1; bank_2; banking_concern_1; banking_company_1	banking; organization_1; group; Corporation+; Function+ Group+ Human+
banku_4;	bank_3	factotum; object_1; object; LandArea+; 1stOrderEntity+ Natural+ Object+ Place+
banku_5;	bank_6	finance; assets_1; possession; CurrencyMeasure+; Function+
banku_6;	bank_5	money; income_1; possession; Keeping+; Artifact; Function+ MoneyRepresentation+ Part+
banketxe_3; banku_7;	banking_industry_1; banking_system_1	industry; industry_1; group; Corporation+; Function+ Group+ Human+

Table 1: Synsets containing *banku* in EusWN and aligned English data

On multiple occasions, lexical data have been transferred from dictionaries to build wordnets from scratch, using the Princeton WordNet concept grid as the starting point (i.e., the 'expand method'), or to enrich already existing wordnets; advantages and shortcomings of this approach have been discussed widely (Vossen, 2002; Fišer & Sagot, 2015, among others). A major problem regarding this approach is, again, a mismatched merging of word senses that belong to homonymous or polysemous dictionary headwords and WordNet concepts.

Automated drafting of bilingual dictionary content may significantly ease the manual effort required to make dictionaries from scratch. As earlier experiments have shown, even for a relatively marginal language-pair like German-Basque, one can obtain equivalent candidates for around two thirds of the initial lemma list (Lindemann et al., 2014). But, in any case, it is not only the recall on the initial word lists that automated drafting methods may offer, but it is also, of course, the precision, that is, in our case, the adequacy of the draft equivalent pairs that makes the difference: for the

production of a dictionary that deserves this name, as long as automated efforts continue to fail to achieve precision rates approaching 100%, manual editing of the draft data seems indispensable.

The English Princeton WordNet and Basque WordNet, the two resources used for the experiments described in this paper, were manually built, or at least manually validated. Thus, we should expect high precision, and experiments carried out in the past confirm this assumption even for pivoted bilingual dictionary drafting. Lindemann et al. (2014) evaluated a German-Basque dictionary drafting experiment that involved data from English and Basque WordNets, and from GermaNet (Hamp & Feldweg, 1997), version 8. They found that the rate of equivalences assessed as false did not reach 10%, and another 10% was assessed as partly correct (for the partial matching of compounds) or nearly so, i.e. fuzzily correct. These precision rates were surpassed only by the data from cross-language links attached to *Wikipedia* page titles, and by the Basque equivalents in German *Wiktionary.*[1] However, the latter two resources yielded a much lower recall on the list used as gold standard for German dictionary headwords.

WordNet	**Noun items**	**Verb items**	**Adjective items**	**Adverb items**	**Synsets**
Princeton 3.0 (PWN)	147,245	25,051	30,082	5,580	118,408
Basque 3.0 (EusWN)[2]	40,420	9,469	148	0	30,263

Table 2: Statistics from MCR 3.0

Basque WordNet (Pociello, Agirre & Aldezabal, 2011) was built by semi-automatic means following the 'extend model', i.e. by defining Basque lexicalizations for concepts present in Princeton WordNet (Miller et al., 1990). After a semi-automatic drafting by transfer from Basque dictionaries, the workflow for the construction of this resource involved a manual validation of the whole content. In Basque WordNet 3.0 (henceforth EusWN), concepts are aligned one-to-one to Princeton WordNet 3.0 (PWN) synsets. EusWN can thus be regarded as a translation of PWN. Table 2 contains the overall statistics for both resources.[3] It is clear that EusWN covers no more than about 25% of the concepts represented in PWN.

[1] Also one of the assessed parallel corpus word alignment tools led to results with a precision higher than 90%, but with a very conservative parameter setting, that allowed a recall not higher than 5%.

[2] Not all EusWN synsets contain lexical items; in the case they are not linked to any Basque lexical item, they are, however, semantically annotated. See Pociello et al. (2011) for reference.

[3] The content from both WordNets and documentation are available at http://adimen.si.ehu.eus/web/MCR/.

BabelNet (Navigli & Ponzetto, 2010) is an automatically built multilingual resource. It contains data extracted from a wide range of sources, some automatically, some manually built or manually validated. Just as in WordNet, the basic unit in the data model is the synset node, which is identified by a unique number. Just as in MCR and Open Multilingual WordNet (Bond & Foster, 2013), two of the approaches used to build a multilingual WordNet, all concepts exist in English, and as soon as lexicalizations and other item types in languages other than English that belong to these concepts are available, they become linked to one of these.

BabelNet 3.7	**Noun items**	**Verb items**	**Adjective items**	**Adverb items**	**Synsets**
English (overall)	11,303,752	58,644	112,518	19,545	6,667,885
English (English-Basque intersection)	5,010,332	15,132	1,310	190	2,469,915
Basque	2,727,673	9,558	443	54	2,469,915

Table 3: Statistics for BabelNet 3.7

One of the additional values of BabelNet is that content extracted from numerous resources[4] appears as merged to BabelNet synsets that ideally should be unique for each concept, i.e. duplicate concepts should be merged to a single synset. The extraction and merging tasks are performed by algorithms which are regularly improved and updated, along with the inclusion of more data. Table 3 contains statistics for the BabelNet content as for version 3.7, released in 2017.

2. From Bootstrapping to Evaluation

Having in mind the reasons discussed above, for the research presented in this paper we concentrate on a transfer-based method that allows for the extraction of sense-to-sense equivalences. We have been bootstrapping and evaluating bilingual lexical data from English and Basque WordNets, on one hand, and from BabelNet, on the other. The underlying approach is as simple as extracting lexicalizations in two languages for the same concept (i.e., that share a common unique synset ID), and quantitatively and qualitatively evaluating the obtained bilingual dictionary draft. The rates for the recall of the extracted data on a basic lemma list and for precision in terms of translation equivalence, give us clues related to both uses at the same time: for a possible use as draft content in dictionary making, and for what we have to expect when using web portals that present automatically gathered data as a reference dictionary.

[4] A complete list of the sources for the lexical items in BabelNet is available at: http://babelnet.org/about.

We downloaded the BabelNet 3.7 indices dump which stores the BabelNet corpus as an *Apache Lucene* index.[5] We retrieved its content using the Java API which is available for download on the BabelNet website.[6] We collected the synsets that contain at least one lexical entry for English and one for Basque, and found 2,469,915 synsets for this intersection. For each synset we collected (1) the BabelNet synset ID; (2) the English and Basque lexical items; (3) the English glosses; (4) metadata on the "type" of the synset, which is either "named entity" or "concept"; and (5) the source of the lexical item. Additionally, the synset ID includes (6) a marker for part of speech. We wrapped our scripts into a processing pipeline, both for reproducibility of the results and for an easy adaptation to other language pairs (or language sets).

For all intersection calculations, lexical items are taken into account as graphically normalized strings. All upper case letters have been converted to lower case, and all hyphens or spaces between multiword lexical units have been suppressed, in order to harmonize graphical variants found in the sources. For example, the Basque term for *death penalty* appears in the data in three graphical variants (*heriotza-zigor, heriotza zigor, Heriotza zigor*), each of which are normalized to a one-word form, *heriotzazigor*. This form is not documented in the data, but in general, noun+noun compounds in Basque also may appear as one single word (*eguzki-lore, eguzki lore* or *eguzkilore*, literally 'sunflower').[7] Some items, namely those stemming from Wikipedia and Wikidata, may contain a short sense-disambiguating gloss in brackets, in addition to the lexical item itself, as in *gotiko (hizkuntza)*, and *gotiko (estiloa)*, 'gothic language' vs. 'style'. These glosses have been suppressed for the same reason: in the respective synset, the strings *gotiko* and *Gotiko* appear, with no gloss; after normalization, all four are treated as duplicates, and therefore as one unique lexical item.

In general, we found inconsistencies regarding the initial case of lexical items. In principle, Basque orthography is more regular than English, as a range of nouns that are not considered named entities (proper names) in English have an uppercase initial letter (e.g. names of languages, days of the week, months, etc.). But, aside from this, many inconsistencies have been found in the Basque lexical items stemming from BabelNet sources other than WordNet. For instance, Basque terms in software localization (Microsoft Terminology) bear initial upper case; even verbs such as *Bidali*, 'send' or *Onartu*, 'accept', presumably because these equivalent pairs were defined to serve as localized flags for buttons on a website or software application. For items that represent Basque Wikipedia page titles, we have also found inconsistencies: around 30% have a lower case initial letter, but this feature seems not to be consistently

[5] https://lucene.apache.org/core

[6] http://babelnet.org/download

[7] Unlike the two separated variants of this compound, the merged single word is not found in the normative wordlist of Standard Basque (Euskaltzaindia, 2010), although it is frequent in corpora. In other cases, in turn, a merged compound is listed as the standard form (*aireontzi*, 'airplane'). For the experiments presented here, multiword units are merged in general.

related to the noun type.

We have not used the noun type filter built into BabelNet, that is, the tags "named entity" and "concept" present in the synsets, to evaluate the effect of that filter. Consequently, lexicalizations for named entities (proper nouns) also may appear in the counts presented in Table 4 if a common noun is homographous (e.g. Basque (and Spanish) *Lima* to *lima*, 'lime' *Gaza* to *gaza*, 'gauze'). It should also be mentioned here that Basque nouns erroneously tagged as common nouns instead of proper nouns in the corpus processing (e.g. *Praga*, 'Prague', *Polisario*) at this stage, have not been manually removed from EusLemStd, a Basque lemma inventory used for quantitative evaluation (see Section 3).

The quantitative and qualitative evaluation has been carried out using built-in features of the *TshwaneLex* software application,[8] into which we have imported all lexical data on hand. This allowed us to merge all data according to a pre-defined XML schema, and, at the same time, to keep all evaluation steps reproducible.

3. Quantitative Evaluation

In this section, we give an account of intersecting sets of (1) the extracted lexical data stemming from (a) WordNet and (b) BabelNet, and (2) the entries of EusLemStd, a frequency headword list used here as gold standard for a Basque lemma inventory. This word list is produced by computational means; it contains common nouns, verbs, adjectives and adverbs that appear as headwords in at least one of the standard reference dictionaries for Basque, as well as in at least one of the two major monolingual corpora, a hand-selected reference corpus, and a large web corpus (see Lindemann & San Vicente, 2015). The qualitative evaluation of random subsets of this intersection is presented in Section 4 below.

Headwords: intersecting sets		
EusLemStd ∩ EusWN ∩ BabelNet	18,004	(31.0%)
EusLemStd ∩ EusWN	18,122	(31.3%)
EusLemStd ∩ BabelNet	23,194	(40.0%)
EusLemStd	57,919	(100.0%)

Table 4: Intersection of EusWN, BabelNet, and EusLemStd (headword strings).

[8] http://tshwanedje.com/tshwanelex/

In Table 5, we quantify the intersection of (1) EusWN/PWN concepts, (2) BabelNet concepts and EusLemStd; that is, synsets that contain at least one item found on the Basque reference lemma list.

Concepts: intersecting sets	**Noun synsets**	**Verb synsets**	**Adjective synsets**	**Adverb synsets**	**Synsets**
EusWN ∩ EusLemStd	21,533	2,894	106	0	24,533
BabelNet ∩ EusLemStd	31,028	2,914	293	25	34,260

Table 5: Intersection of EusWN, PWN, and EusLemStd (concepts)

The Basque lexical items found in BabelNet stem from the sources listed in Table 6. The table contains the overall numbers of items, as well as the numbers of strings that also appear in EusLemStd. Lexical items homographous to each other inside or across parts of speech count here as one unique string.

Source	**EusLemStd intersection unique items**	**EusLemStd intersection total items**	**BabelNet 3.7 total Basque items**
All Sources	23,194	67,221	2,737,728
Open Multilingual WordNet	18,060	39,343	48,934
Wikidata	7,347	8,159	190,764
Wikipedia	6,646	6,849	182,967
BabelNet	2,215	3,989	2,255,355
Wikipedia Redirections	3,254	3,565	51,440
OmegaWiki	2,485	2,816	5,625
Wiktionary	1,464	1,629	2,188
Microsoft Terminology	581	735	3,887
GeoNames	75	79	1,879
WikiQuotes	29	29	218
WikiQuotes Redirections	28	28	96

Table 6: Basque lexical items in BabelNet 3.7 (concepts and named entities)

If we relate these figures to the amounts of synsets, for the intersection of the Basque BabelNet with EusLemStd, we find a distribution of Basque lexical items per synset as shown in Table 7. Note that synsets that contain a standard lemma also may contain further items not found on EusLemStd. Synsets tagged as "named entity" in BabelNet have been filtered from the subsets quantified in this table.

Source	EusLemStd intersection items/synset	EusLemStd intersection total synsets	BabelNet 3.7 total Basque synsets
All Sources	2.28	29,420	2,469,915
Open Multilingual WordNet	1.59	24,786	28,699
Wikidata	1.02	8,004	87,922
Wikipedia	0.87	7,883	81,777
BabelNet	3.44	1,161	1,755,914
Wikipedia Redirections	0.85	4,210	11,598
OmegaWiki	1.08	2,607	3,970
Wiktionary	1.09	1,496	1,656
Microsoft Terminology	1.07	689	3,108
GeoNames	0.00	0	4
WikiQuotes	0.51	57	61
WikiQuotes Redirections	1.33	21	24

Table 7: Basque concepts in BabelNet 3.7 (tagged as "concept" in BabelNet)

4. Qualitative Evaluation

4.1 WordNet

For the translation equivalences obtained from WordNet, we have carried out a qualitative evaluation for (1) a random set of noun and verb synsets that contain only monosemous Basque items; that is, items that occur only in one synset, and (2) a random set of other synsets; i.e., those that also contain polysemous Basque lexical items, as we presumed a higher degree of fuzzy or false matchings for polysemous items. For adjectives, we have not evaluated the monosemous items separately, as the number of synsets containing only these does not even reach a dozen. The adequacy of the semantic matching between Basque and English equivalents has been assessed on a scale of three values, as formerly used in similar studies (Fišer et al., 2012; Lindemann et al., 2014):

(1) OK, for a correct matching, in the sense that the Basque lexical item could be used in a dictionary entry for denoting the pertaining concept without any changes,

(2) FUZZY, for a fuzzy semantic matching, which means that the lexical item does not match the pertaining concept in a way that could be used in a dictionary entry, but that its semantic distance to the ideal equivalent is to be regarded as small; it may be a hyponym or hypernym, a meronym or a holonym of an ideal equivalent. For verbs, equivalents that are semantically very close but with incompatible valencies (e.g. regarding transitivity) are also assessed as FUZZY. A paraphrase of this value could be "the lexicographer has to intervene here,

but it is not a completely false equivalent."[9]

(3) FALSE, for a lexical item that provides nothing usable for a lexicographer when editing the entry.

For 300 synsets, the adequacy of the corresponding 546 lexical items has been assessed. The distribution of the assessment values is summarized in Table 8.

The data taken into account for assessment are the English glosses and example sentences, and the English and Basque lexical items. During the assessment process, the semantic relations or ontology classes of a synset could also be displayed. We assess the equivalents as for a translation from Basque to English, which is the direction contrary to the editing process of EusWN. Consequently, we do not assess here whether the group of English items could have been translated to Basque in a more appropriate way than via the Basque items found.[10] Critical in this context are nominal derivations from Basque verbs, often employed in EusWN as equivalent of English nouns that denote actions or results of actions, but that are not treated as lemma in Basque dictionaries, and consequently neither in EusLemStd, as for example the nominal derivations *xahutze*, *ahaitze* for the English 'wastage'.

EusWN/PWN equivalences	**Nouns**	**Verbs**	**Adjectives**	**All POS**
Total synsets intersect. EusWN/EusLemStd	21,533	2894	106	21,533
• Monosemous	6,058	201	11	6,270
• Polysemous	15,343	2,693	95	18,131
Synsets evaluated	100	100	100	300
• Monosemous	50	50	16	
• Polysemous	50	50	84	
Synsets all items OK	87%	75%	94 (94%)	85%
• Monosemous	45 (90%)	37 (74%)		
• Polysemous	42 (84%)	38 (76%)		
Synsets OK/FUZZY	98%	94%	96 (96%)	96%
• Monosemous	49 (98%)	48 (96%)		
• Polysemous	49 (98%)	46 (92%)		
Synsets 1+ FALSE	2%	7%	4 (4%)	4%
• Monosemous	1 (2%)	2 (4%)		
• Polysemous	1 (2%)	5 (10%)		

Table 8: Qualitative evaluation of equivalents extracted from EusWN

[9] The equivalence assessed in this way is not to be confused with *fuzzynymy*, which is a semantic relation encoded in EuroWordNet, that holds when the tests for synonymy, homonymy and meronymy "fail but the test *X has some strong relation to Y* still works" (Vossen, 2002: 37). Fuzziness here includes all somehow close relations apart from cross-language synonymy (in the sense of adequacy as dictionary translation equivalent), i.e. including homonymy.

[10] However, we have unsystematically annotated the assessed data with free text comments and proposals for more appropriate equivalents. These annotations may be used in the future as notes for preparing a more systematic and complete survey.

4.2 BabelNet

The qualitative evaluation of Basque-English equivalences found in BabelNet differs from the process described above in some points. As explained above, together with the Basque lexical items we have extracted the tags denoting their respective source and stored the data in the database used for evaluation. This allows the disambiguation of the evaluation results according to the source of the pertaining item, as shown in Table 10.

Since we found the rendering of the automatic sense merging carried out for building BabelNet a particularly interesting detail, we have introduced a fourth assessment value, MERGE_ERROR. This value was assigned in cases where the random synset displayed for evaluation was found to contain lexical items that denote (and glosses that describe) two different concepts. For example, one synset contains lexical items and definitions of the English noun *underground* that refer to the word sense 'tube, metro', as in "The London Underground", and to the word sense 'resistance, underground' with the definition "a secret group organized to overthrow a government…", while both senses in PWN appear in distinct synsets. As for the translation equivalence, this value has thus to be regarded a variant of FALSE.

BabelNet 3.7	**OK**	**FUZZY**	**FALSE**	**MERGE ERROR**	**(Asses-ments)**
All Sources	*1,211 (88.9%)*	*63 (4.6%)*	*44 (3.2%)*	*44 (3.2%)*	*1,362*
Open Multilingual WordNet	717 (89.2%)	49 (6.1%)	28 (3.5%)	10 (1.2%)	804
Wikidata	57 (93.4%)	0 (0.0%)	1 (1.6%)	3 (4.9%)	61
Wikipedia	194 (87.8%)	5 (2.3%)	6 (2.7%)	16 (7.2%)	221
BabelNet	3 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3
Wikipedia Redirections	13 (52.0%)	3 (12.0%)	4 (16.0%)	5 (20.0%)	25
OmegaWiki	75 (91.5%)	2 (2.4%)	0 (0.0%)	5 (6.1%)	82
Wiktionary	132 (92.3%)	4 (2.8%)	5 (3.5%)	2 (1.4%)	143
Microsoft Terminology	20 (87.0%)	0 (0.0%)	0 (0.0%)	3 (13.0%)	23
GeoNames	0	0	0	0	0
WikiQuotes	0	0	0	0	0
WikiQuotes Redirections	0	0	0	0	0

Table 10: Qualitative evaluation of BabelNet equivalences for sources

As the reader will observe, the qualitative assessments made for items stemming from different sources diverge significantly. For a dictionary draft, we may accept only items from particular sources, as the encoding of lexical data in BabelNet allows such filtered extraction. Lexical items that originally are titles of redirection pages in *Wikipedia* and *Wikiquotes*,[11] in general, should only match fuzzily or very fuzzily to the pertaining concept. This is because, in their original resource, their reason to be is that there is no other page in that resource that matches better. The redirections in *Wikipedia* that link to *turkey* in the sense of 'turkey meat', for example, include *Turkey Sandwich*, *Cooking a turkey*, *Turkey meat*, and *Turkey dinner*, i.e. two-word units, and even phrases of a different part of speech. Depending on the desired application, such fuzzy matchings may be more or less useful; as translation equivalents, most of them will not serve.

The evaluation results for BabelNet synsets, according to part of speech, are collected in Table 11. In principle, we can also relate the evaluation data disambiguated by source to the parts of speech, both for lexical items and for items grouped as synset. For space reasons, we concentrate here on giving a complete account of the outcome for synsets, as this already provides a good overview of the value a dictionary draft based on BabelNet can have in a lexicographical workflow. The assessments for the 1,184 lexical items that have been evaluated in total are distributed as follows: 1,056 OK (89.2%), 58 FUZZY, 39 FALSE, and 31 MERGE ERROR. As these items belong to 625 different synsets, the average number of Basque lexical items found per synset in this random subset of the English-Basque BabelNet is 1.89.

BabelNet 3.7	**Nouns**	**Verbs**	**Adjectives**	**Adverbs**	**Total**
Assessed synsets	200	200	200	25	625
All items OK	179 (89.5%)	163 (81.5%)	188 (94.0%)	23 (92,0%)	553 (88.5%)
1+ items OK, and 1+ items FUZZY	3 (1.5%)	14 (7.0%)	2 (1.0%)	0 (0.0%)	19 (3.0%)
1+ items OK, and 1+ items FALSE	2 (1.0%)	3 (1.5%)	0 (0.0%)	0 (0.0%)	5 (0.8%)
All items FUZZY	5 (2.5%)	9 (5.5%)	8 (2.0%)	0 (0.0%)	22 (3.5%)
1+ items FUZZY, and 1+ items	1 (0.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.5%)
All items FALSE	5 (2.5%)	8 (4.0%)	1 (0.5%)	2 (8.0%)	16 (2.6%)
MERGE_ERROR	5 (2.5%)	3 (1.5%)	1 (0.5%)	0 (0.0%)	9 (1.4%)

Table 11: Qualitative evaluation of BabelNet equivalences for synsets and part of speech

[11] As for BabelNet 3.7, there is nearly no Basque data found from Wikipedia and Wikiquote Redirections (cf. Section 3 above).

While displaying random noun synsets, in 30 cases the synset referred to a named entity, and the corresponding English lexical items were proper nouns. The reason for these to appear in our evaluation data in all cases was the fact that the Basque equivalent contained a string homographous to a EusLemStd entry, as for example the Basque common noun *datu*, 'date', homograph to "a title for chiefs, sovereign princes, and monarchs in [...] Regions of the Philippines" (*Wikipedia*), or 'materia', which also is the title of an album recorded by an Italian music band. In these cases, we skipped the evaluation of the synset and went on to the next (so that 230 noun synsets have been evaluated in total), but we also performed a second test: Whether the synset was listed as "named entity" (in opposition to "concept") in BabelNet. For all 30 cases, the result was positive, so that we may conclude that named entities are labelled properly in BabelNet. But, in principle, cross-class homograph nouns may appear merged as one in BabelNet (which was not the case in the random subset we evaluated);[12] this is the reason why we wanted to have all string homographs to EusLemStd entries evaluated.

As mentioned above, the algorithms used for concept merging, as for BabelNet 3.7, lead to some mismatched junctions. The intended lexicographic use of BabelNet data is to regard a number of translation equivalences as noisy or false. The problematic aspect for this regarding mismatches, however, is the fact that the unique ID that serves for highlighting the wrongly merged synset will not be stable: as soon as the sense merging algorithm is improved, the concept must be split again. The stability of synset IDs is a central feature for linking concepts across different resources, which we will discuss in the following section.

5. Interoperability Issues and Lexicographic Postprocessing

In this section, we want to give a brief overview of some of the issues related to data model interoperability and the representation of lexical semantic relations. We cannot discuss all issues in detail here; nevertheless, the following general comments may serve as orientation for making a transfer based dictionary drafting, with wordnet-like concept-oriented resources for bootstrapping.

Converting a concept-oriented collection of lexical data into a headword-oriented dictionary draft is a computationally trivial transformation task. As mentioned in Section 2, we are able to represent our dictionary draft datasets in XML, as illustrated in Figure 2 below. In connection with this transformation, we have to mention two issues, which are far from trivial, for lexicographers: (1) the modelling of homography,

[12] While unsystematically browsing BabelNet, we found e.g. Dexter Raymond Mills, Jr., a.k.a. *Consequence*, an American rapper from Queens, New York, merged to the common noun synset *consequence, aftermath*, which is the one the Basque equivalents found here refer to. We also have had a look at the four items extracted from GeoNames that are present in the Basque BabelNet and classified as concept (cf. Table 7); contrary to their classification, all four are place names, and thus, named entities.

i.e. on which level we distinguish between homograph headword strings that point to dictionary entries related to different parts of speech (cf. in English *sound*$_{\text{N}}$, *sound*$_{\text{V}}$, and *sound*$_{\text{ADV}}$), and (2) the modelling of a distinction between homonymy and polysemy (cf. Section 1.1).

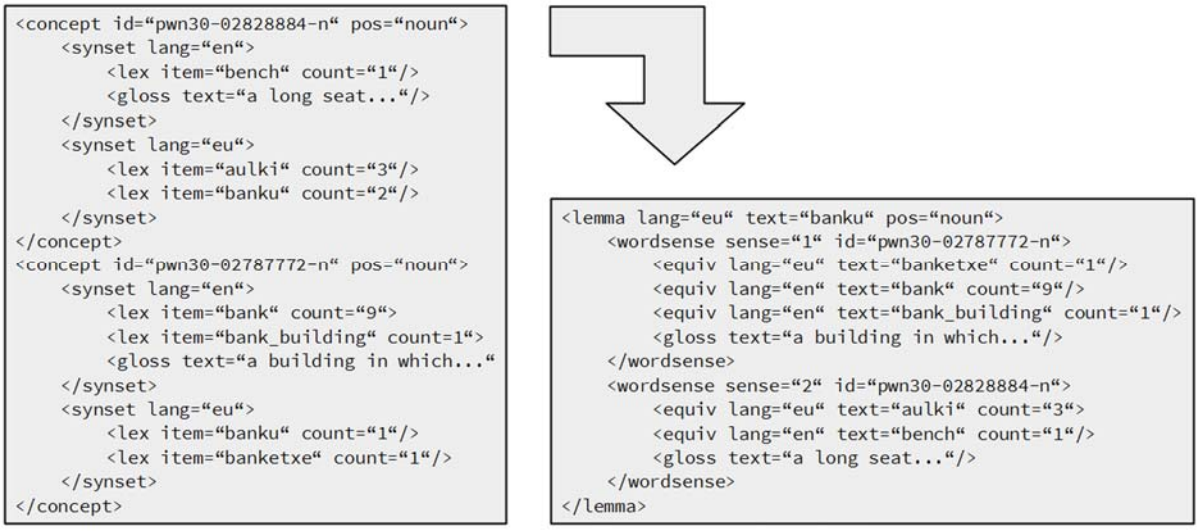

Figure 2: XML transformation

The distinction between homograph lemma-part of speech (lempos) entities is not problematic, since part of speech is encoded in the synset ID, and the transformation described here does not lead to dictionary entries with mixed-up parts of speech. On the contrary, homonymy and polysemy are treated equally in the data model of PWN and EusWN. In the case of the examples discussed in Section 1.1, as a consequence, the homonyms *bank* (institution) and *bank* (of a river) would appear in the same entry, just as do the two senses of *bench* (group of judges, furniture). If a disambiguated representation of these two different phenomena is desired, it has to be introduced in a further postprocessing step. This might work semi-automatically, e.g. by comparison to lists of items flagged as homonyms in dictionary headword lists.

Regarding the bits of XML code shown in Figure 2, we have to point out, of course, that it is a simplified presentation of what is possible. Here we just include the text attributes (alternatively representable as text values) for lexical items and abbreviated glosses. WordNet and BabelNet include more information linked to synsets, which may be used as microstructural item types in a dictionary; chiefly example sentences, domain flags, ontology classes, and semantic relations, and in BabelNet also images. For lexicographic purposes, Benjamin (2016) describes a more sophisticated cross-language mapping between lexical items, instead of (only) between synsets, in order to be able to relate every item-to-item link to more fine-grained classes of (quasi-)synonymy relations. The inclusion of more item types into a dictionary data model that is compatible with wordnet-like resources is a very attractive field to explore. Also, further item types linked to synset-IDs in a multilingual dictionary database potentially represent an extension to the source wordnet, at the same time.

A central issue which is also linked to data modelling is the internal representation of polysemy (besides its disambiguation from homonymy) that results from a transformation as illustrated in Figure 2. Two questions arise: (1) Does the draft dictionary entry contain all word senses of a lemma we want to represent? (2) Is the splitting of word senses found in the draft entry suitable for the dictionary for which we want to produce a draft, or is it (a) too fine-grained, (b) redundant, or are we (c) missing further distinctions?

Regarding question (1), we see no straightforward way to ascertain the respective answer other than via classical lexicography (i.e. manual work). However, we are preparing experiments to address that issue lemma by lemma with semi-automated quantitative comparisons to polysemy structures in existing dictionaries. Such comparisons will be helpful for question (2a,b), in case the sense splitting in the draft data significantly exceeds the number of senses found in reference dictionaries, or vice versa (2c). Before having conducted such bulk comparisons, our analysis of random subsets of the draft data suggests that the phenomena (2a,b) are frequent. One explanation lies in the 'expand' method of wordnet building and is connected to *genuine* and *false autohyponymy*, i.e. the same lexical item appearing in synsets that are hyponyms to each other (Pociello et al., 2011: 135–137). Examples of these include the translation *zahar*, 'old' for the English synset containing *moth-eaten, dusty, stale,* "lacking originality of spontaneity; no longer new", or *edan*, 'drink' for *drink, booze, fuddle*, "consume alcoholic beverages". While genuine autohyponyms should be maintained as different senses in a bilingual dictionary, for lexicographic purposes, false autohyponyms should be merged. A possible strategy for sense merging by PWN's own means is an automated classification as subsenses to one sense of homograph cohyponyms, i.e. lexical items that are graphically identical and share the same hypernym, or a common ancestor even higher in the hierarchy (cf. Miller, 1998: 42).[13]

The problems (2a-c) in computational linguistics are commonly referred to as *granularity* of word senses; different computational applications require more fine or coarse grained word senses (Prakash et al., 2007), and the same, of course, is true for dictionaries that serve different functions. In other words, requirements and strategies for a postprocessing of wordnet sense granularity will be closely related to the lexicographic project at hand. In any case, to merge senses will be technically more feasible than to introduce any splitting.

It should be clear that the problems we find for working with WordNet as a resource for lexicography are closely related to the nature of that resource, and the functions for which it was developed. Lexicography is explicitly not among these functions,

[13] In order to avoid "unmotivated cohyponyms", in other wordnet-like projects, a "crossed classification" of synsets is introduced, i.e. a classification of the same synset node in two different places in the hierarchy allowed in GermaNet, such as *banana* (a) as edible fruit and (b) as cultivated plant (Kunze, 2010, p. 507); such double classifications of the same concept could regularly be transformed into subsenses in a dictionary entry.

although WordNet has become a de-facto standard resource for monolingual and multilingual e-dictionary projects of all kinds.[14] Benjamin (2016: 28–31) mentions related problems not directly linked to data models but mostly to the original functions of WordNet: (1) The glosses linked to PWN synsets often do serve for disambiguating word senses, but not in a way that could be regarded adequate for publishing in a dictionary entry. (2) Some wordnets of languages other than English have been built automatically and contain a significant amount of errors, which is not problematic for some NLP applications, but it is, of course, for lexicography; and it becomes highly problematic if noisy data are just reproduced in a dictionary portal without being marked as possibly wrong. (3) The criterion that defines synonymy in wordnets is relatively weak in the sense that it allows too many cross-language equivalence links (between all members of a synset in language A to all members of a synset in language B). In other words, well-defined subclasses of the synonymy relation should be introduced systematically. Aside from that, the author mentions that when building a wordnet by the 'expand method', (4a) some synsets are filled with explanatory phrases instead of lexical items that serve as dictionary lemma, and (4b) a concept must exist in PWN to be expanded to the new wordnet. Finally, (5) the restrictive licensing of some wordnets makes bulk bootstrapping, and in some cases even isolated experiments, impossible.

6. Conclusions and Further Work

By bootstrapping wordnets and BabelNet, we have built a bilingual dictionary draft from scratch that includes a grid of lempos: entities and word senses, each of which furnished with one or more lexical items in two languages, and covering up to 40% of a previously defined list of Basque dictionary headwords. By the quantitative and qualitative evaluation of these draft data we have verified our initial hypothesis regarding the precision of the obtained translation equivalent pairs. Comparing the rendering of WordNet data versus BabelNet data, we come to the following two main conclusions:

(1) In terms of recall on our initial Basque lemma list, BabelNet yields significantly higher rates than EusWN alone (around 40% compared to 30%), and, at the same time, the precision we have measured by manual assessments stays on a very similar level, close to 90%. This, of course, is recall and precision regarding an English-Basque dictionary draft, and if we wanted to produce new dictionaries for uncovered language pairs with English as pivot, we would have to also take into account the data for these third languages. As an example of a lexicographically uncovered language pair, we have measured the recall for Slovene translation equivalents on Basque lemmata (EusLemStd) comparing

[14] A list of dictionary websites that use WordNet data is found at https://wordnet.princeton.edu/wordnet/related-projects/.

bootstrapped dictionary draft data from wordnets and from BabelNet, with encouraging results. By linking EusWN to SloWNet (3.0 2015 version, Fišer et al., 2012), 66% of the synsets that contain EusLemStd lemmata also contain Slovene lexical items (16,291 synsets); on the other hand, 78% of the BabelNet synsets that contain EusLemStd Basque lemmata contain also Slovene items (22,864 synsets). As we have done here for Basque-English, a qualitative evaluation of the drafted Slovene-English mappings would be necessary, in order to predict the precision of a Basque-Slovene dictionary draft.

(2) Both EusWN and BabelNet 3.7 synsets are identified by unique ID codes that may be copied into the dictionary draft, following the goals discussed in Section 5 above. There is no guarantee for the stability of BabelNet synset mergings, and consequently of the corresponding synset ID codes, at least as for the current version 3.7, as we have pointed out in Section 4.2. The same problem also applies to WordNet data, but with an announced solution. EusWN synsets are linked one-to-one to PWN synsets, and their ID numbers correspond to an Interlingual Index that has been adapted from the sense inventory of Princeton WordNet 3.0, which means that it will not necessarily be compatible with future Princeton WordNet versions nor updated versions of other wordnets. As a possible solution, we are looking forward to the implementation of a stable, version-independent Global WordNet Grid (Vossen et al., 2016), a list of unique concept identifiers that will serve as central sense index across languages and future updates of wordnets.

In any case, we have shown that if a bilingual dictionary project starts from scratch, it makes sense to include a drafting of a word sense grid and translation equivalents in the workflow, starting with wordnet-like concept-oriented resources. Apart from the more obvious and doubtlessly very important advantage of reducing the manual effort in dictionary content editing, we point out a benefit, closely linked to the data model used that underlies the resources used here for bootstrapping. As soon as the lexicographical process goes on, i.e. the lexical data obtained from the dictionary draft are being edited, enriched, and linked to other lexicographic item types, they can be used for reciprocally enriching the resources of the wordnet-family, by retro-bootstrapping and inclusion, or by the definition of cross-resource links. Necessary conditions for a continuous mutual enrichment of this kind are the stability of synset IDs on the wordnet side, and the maintenance of an interoperable data model on the dictionary side.

For the Basque language, without taking into account the licence constraints that still apply in some cases, based on wordnets today we are able to produce bilingual dictionary drafts with about 70 languages. By bootstrapping BabelNet, we can obtain drafts with many more; we would start with a quantitative analysis of the mutual coverage (intersection) of every possible language pair in this very big resource. This, as we have shown, does not significantly lower the precision of its content in comparison to its nucleus, the multilingual wordnet, in spite of growing more and

more. If we connect any two languages by the methods described here, in both cases, i.e. using wordnets and using BabelNet, the English language functions as hub. Therefore, it makes sense to first evaluate the quality of the mappings between the desired languages and English, as we have done here for Basque.

For the part of a standard Basque dictionary headword list that today can be covered by the methods described here, a manual editing would allow to discover and to fill sense gaps, to improve the description of senses, and to correct errors. For the part of the list that is not covered, links to concepts that exist in English concept-based resources will have to be set. In some cases, for a Basque word sense no matching concept is listed in the originally English-based resources; the "discovered" concept will serve as an amendment to those, and so to a human and machine readable conceptualisation of our world.

7. Acknowledgements

The research leading to these results has received funding from the Basque Government (Research Group IT665-13). Funding is gratefully acknowledged.

8. References

Benjamin, M. (2016). Problems and Procedures to Make Wordnet Data (Retro)Fit for a Multilingual Dictionary. In *Proceedings of the Eighth Global WordNet Conference*. Bucharest, Romania, pp. 27-33.

Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the The 51st Annual Meeting of the Association for Computational Linguistics*

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

Fišer, D., Gantar, P. & Krek, S. (2012). Using explicitly and implicitly encoded semantic relations to map Slovene Wordnet and Slovene Lexical Database. In *Semantic Relations-II. Enhancing Resources and Applications. Istanbul, Turkey*

Fišer, D., Novak, J. & Erjavec, T. (2012). sloWNet 3.0: development, extension and cleaning. In Proceedings oft he 6th International Global Wordnet Conference. *Matsue, Japan*, pp. 113-117.

Fišer, D. & Sagot, B. (2015). Constructing a poor man's wordnet in a resource-rich world. *Language Resources and Evaluation*, 49(3), pp. 601–635.

Gonzalez-Agirre, A., Laparra, E. & Rigau, G. (2012). Multilingual Central Repository version 3.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC'12*. Istanbul, Turkey.

Gouws, R. (2002). Equivalent Relations, Context and Cotext in Bilingual Dictionaries. *Hermes*, 28(1), pp. 195–209.

Hamp, B. & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.* Madrid, Spain, pp. 9–15.

Hartmann, R. R. K. (1990). The not so harmless drudgery of finding translation equivalents. *Language & Communication*, 10(1), pp. 47–55.

Kunze, C. (2010). Lexikalisch-semantische Ressourcen. In K.-U. Carstensen, C. Ebert, C. Ebert, S. J. Jekat, R. Klabunde & H. Langer (eds.), *Computerlinguistik und Sprachtechnologie: eine Einführung*. Heidelberg: Spektrum.

Lindemann, D. & San Vicente, I. (2015). Building Corpus-based Frequency Lemma Lists. *Procedia - Social and Behavioral Sciences*, *198*, pp. 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. EURALEX 2012*. Bolzano, Italy, pp. 563–576.

Miller, G. A. (1998). Nouns in wordnet. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database*. Cambridge MA: MIT Press, pp. 24-45.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4), pp. 235–244

Navigli, R. & Ponzetto, S. P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA

Pociello, E., Agirre, E. & Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45(2), pp. 121–142.

Prakash, R. S. S., Jurafsky, D. & Ng, A. Y. (2007). Learning to merge word senses. In *Proceedings of EMNLP-CoNLL 2007*

Saralegi, X., Manterola, I. & San Vicente, I. (2012). Building a Basque-Chinese Dictionary by Using English as Pivot. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation. LREC'12*. Istanbul, Turkey

Varga, I., Yokoyama, S., & Hashimoto, C. (2009). Dictionary generation for less-frequent language pairs using WordNet. In *Literary and Linguistic Computing*, 24(4), pp. 449–466.

Vossen, P. (2002). EuroWordNet General Document. University of Amsterdam.

Vossen, P., Bond, F., & McCrae, J. (2016). Towards a truly multilingual Global

Wordnet Grid. In *Proceedings of the Eighth Global WordNet Conference. Bucharest, Romania*, pp. 419-426.

Wiegand, H. E. (2002). Equivalence in bilingual lexicography: criticism and suggestions. *Lexikos*, 12(1), pp. 239–255.