Bilingual Dictionary Drafting. The Example of German-Basque, a Medium-density Language Pair

David Lindemann[1], Iker Manterola[2], Rogelio Nazar[3], Iñaki San Vicente[2], Xabier Saralegi[2]
[1]UPV-EHU University of the Basque Country, [2]Elhuyar Foundation,
[3]Pontificia Universidad Católica de Valparaíso
david.lindemann@ehu.es, i.manterola@elhuyar.com, rogelio.nazar@ucv.cl,
i.sanvicente@elhuyar.com, x.saralegi@elhuyar.com

Abstract

This paper presents a set of Bilingual Dictionary Drafting (BDD) methods including manual extraction from existing lexical databases and corpus based NLP tools, as well as their evaluation on the example of German-Basque as language pair. Our aim is twofold: to give support to a German-Basque bilingual dictionary project by providing draft Bilingual Glossaries and to provide lexicographers with insight into how useful BDD methods are. Results show that the analysed methods can greatly assist on bilingual dictionary writing, in the context of medium-density language pairs.
Keywords: bilingual dictionary drafting; comparable corpora; Natural Language Processing; open lexical resources; parallel corpora

1 Introduction

For a bilingual dictionary project that starts from scratch, from no or little previous lexicographical work and no or little bilingual glossaries (BG) existing on their language pair, a lexicographer lacks a useful set of guidelines for Bilingual Dictionary Drafting (BDD) strategies. A Dictionary Draft, i.e., lexicographical data obtained by automatic or semi-automatic methods, is useful in the lexicographical process as it may ease the editing of macro- and microstructural lexicographical data and save human resources.

In this article, we present a set of BDD methods and their evaluation on the example of German-Basque as language pair: Direct extraction of bilingual glossaries from existent lexicographical databases and Corpora based Natural Language Processing extraction methods. The evaluation of the obtained glossaries is done (1) quantitatively for the covering of German lemmata against a corpus based frequency lemma list adapted from DeReWo-40.000 (IDS 2009) as gold standard, (2) quantitatively for the amount of Basque Translation Equivalents (TE) obtained, and (3) qualitatively for the adequateness of the TE pairings, against manually edited German-Basque dictionary entries from EuDeLex, the lem-

malist of which is adapted from DeReWo-40.000, as gold standard (3406 German lemmata starting with A), and for the adequateness of the TE's part of speech (POS)[1].

Our aim is twofold: to give support to the bilingual dictionary project EuDeLex by providing draft BG and to provide useful information related to BDD methods for lexicographers working on medium-density language pairs.

A Word about Density

Density, understood as "the availability of digitally stored material" in a language (Varga et al. 2005) is a factor not to be neglected in corpus-based lexicography. In most cases, the number of speakers of a language and its size on the web serve as approximation indicators for density, and the availability of electronic language resources is also a factor to be considered. Following Varga et al. (*op. cit.*), we group languages according to density as follows:

(1) High-density languages: languages with a hundred million speakers or more (about 12)
(2) Low-density languages: small languages with less than half a million speakers (more than 5000)
(3) Medium-density languages: languages that lie between these two extremes (about 500)

Basque is one of the latter ones; Table 1 carries a comparison of density approximation indicators for German, one of the high-density languages, and Basque.

In the bilingual context, it is the density of the smaller language of the pair that determines by which methods Dictionary Draft data can be gathered and to what extent those methods lead to useful results.

Approaches for obtaining BG that rely on statistical Natural Language Processing methods (part 2.2) and that provide reliable results in higher density language pairs, in our case may lead to a much more limited success, and we shall ask whether the reasons for a more limited success of NLP methods are the quantitative and qualitative limitations of parallel and comparable corpora available for our language pair, or whether a lack of performance is explained also by the employed NLP tools themselves, which do lead to good results for the (high density) languages they were designed for. In the case of Basque, as it is not official in EU, we can not recur to parallel corpora based on EU legal documents, as we would in the case of other medium-density European languages (cf. Steinberger et al. 2006). Independently from this fact, German-Basque parallel corpora compiled from movie subtitles and software localization files may reach considerable sizes in a near future, as promised in the OPUS Corpus project (cf. Tiedemann 2012).

1 EuDeLex is currently being developed at UPV-EHU (cf. Lindemann 2014). The manual editing of German letter A (around 10% of the planned lemmalist based on DeReWo) has been finished. The intersecting set of EuDeLex and DeReWo (German Letter A) covers more than 90% of both. EuDeLex is available at http://www.ehu.es/eudelex/.

	German	Basque
Speakers	98 million	0,8 million
Biggest Corpus (token counts)	5,4 billion	0,12 billion
Wikipedia Pages	4,5 million	0,37 million
Web contents	5.7%	< 0.1%
ELRA Products	444	6

Table 1: Some density approximation indicators for German and Basque.

Approaches that rely on lexicographical databases maintained by human lexicographers (part 2.1) also presumably suffer from a density-bias: Wikimedia content is crowd-edited by the collaborating communities of volunteers, and those of a high-density language like German largely outnumber voluntaries in Basque[2]. On the other hand, lexical databases maintained in an academic context by human lexicographers like WordNet, may be on a par in terms of quantity and quality, disregarding density across languages.

2 Bilingual Dictionary Drafting Methods

2.1 Extraction of BG from existent lexicographical databases

2.1.1 Wikimedia

For the page titles that match to our gold-standard lemmalist, we extract the whole Wikipedia / Wiktionary page content if existent. Redirect pages are also taken into account. From the page content, we extract the Interlanguage-Link for Basque (Wikipedia) and the Basque translation links (Wiktionary).

2.1.2 WordNet

In this experiment, we align German WordNet lexical units (GermaNet 8.0, see Hamp & Feldweg 1997) with Basque WordNet (EusWN 3.0, see Pociello 2007) lexical units using Princeton WordNet (PWN 3.0, see Fellbaum 1998) as pivot. GermaNet synsets (to which *n* lexical units belong) are referred to PWN synsets in the GermaNet Interlingual Index records (ILI). On the other hand, the EusWN datasets carry links to PWN. By parsing the WordNet data into the same XML file, we get a structure like the one shown in Fig. 1. From each of those aligned datasets, German-Basque glossary entries are extracted by pairing all German lexical units to all Basque lexical units present in the synset.

2 For instance, German *wiktionary* counts with 78634 user accounts and 199 active members over the last month, while the Basque *wiktionary* only has 1982 accounts with 11 active members (statistics from 15.09.2013).

```
<EngSynset Synset="eng-30-00042757-n">
     <PWN30 EngLexUnit="departure_1"></PWN30>
     <PWN30 EngLexUnit="going_1"></PWN30>
     <PWN30 EngLexUnit="going_away_1"></PWN30>
     <PWN30 EngLexUnit="leaving_1"></PWN30>
     <GermaNet80 GerLexUnit="Abfahrt"></GermaNet80>
     <GermaNet80 GerLexUnit="Weggang"></GermaNet80>
     <GermaNet80 GerLexUnit="Aufbruch"></GermaNet80>
     <GermaNet80 GerLexUnit="Distanzierung"></GermaNet80>
     <EusWN30 EusLexUnit="irteera_6"></EusWN30>
     <EusWN30 EusLexUnit="joanaldi_1"></EusWN30>
     <EusWN30 EusLexUnit="joate_1"></EusWN30>
</EngSynset>
```

Fig. 1: Aligned data of 3 WordNets.

2.2 NLP Methods

In this work, three different tools were used in order to extract lexical correspondences. Each tool depends on different NLP methods and resources. On the one hand, we applied a tool called *Pibolex*, which relies on pivoting over existing bilingual dictionaries and combines their structure and comparable corpora based methods for selecting correct translations of source words. On the other hand, we made use of two tools for bilingual lexicon extraction from parallel corpora: Giza++ and Bifid. The following sections describe those tools and the experiments we conducted with them.

2.2.1 Pibolex: Pivot techniques + comparable corpora word-alignment

Pivot-based bilingual dictionary building is based on merging two bilingual dictionaries which share a common language (e.g. LA-LB, LB-LC) in order to create a dictionary for a new language pair (e.g. LA-LC). In our case, we merged the German-English *Beolingus*[3] dictionary (Lde-en) with the English-Basque *Elhuyar*[4] dictionary (Len-eu), obtaining Lde-eu. However, this process may include wrong translations due to the polysemy of words. A pivot word can lead to wrong translations corresponding to senses not represented by the source word. These senses can be either completely different or related but with a narrower or wider meaning. For pruning the wrong translations, in this work we apply the Pibolex tool (Saralegi, Manterola & San Vicente 2011) which uses two different methods adequate for medium-density language pairs because they depend on resources that can be easily obtained:

(a) Inverse Consultation (IC1) (Tanaka & Umemura 1994): this algorithm uses the structure of the source dictionaries to measure the similarity of the meanings between a source word and its translation candidates. The IC1 method counts the number of pivot words in language B between a source word in LA and its TE candidate in LC. The more pivot words found, the stronger is the evidence for the candidate to be correct.

(b) A pruning method based on cross-lingual distributional similarity (DS) computed from a bilingual comparable corpus. Different authors (e.g. Fung 1995; Rapp 1999) have proposed to extract bilingual

3 http://dict.tu-chemnitz.de
4 http://hiztegiak.elhuyar.org

equivalents from monolingual or comparable corpora because, despite offering lower accuracy than those extracted from parallel corpora, they can be an alternative for medium and low density language pairs where parallel corpora are scarce. The underlying idea is to identify as TEs those words which show similar distributions or contexts across two corpora of different languages, assuming that this similarity is proportional to the semantic distance. The method we apply here is described in detail in Saralegi, San Vicente & Gurrutxaga (2008). Following the "bag-of-words" paradigm, a word *w* is represented by a vector composed of weighted collections of words. Those words are extracted from the contexts where the word *w* appears in the corpus. The context words are weighted with regard to *w* according to the Log-likelihood ratio measure. Once we have vector representations of the words in both languages, the algorithm computes for each source word in LA the cosine similarity between its context vector and the context vectors of all TE candidates in LC. However, we can not directly compare vectors in different languages. In order to overcome this problem, we translate vectors of words in LA to LC by means of the noisy bilingual dictionary Ldeeu, which is the only bilingual dictionary available at this stage of the process.

The IC1 algorithm suffers from low recall, which makes it rather inadequate for the task at hand. But the combination of it with the DS based method may be a way to tackle this problem. DS results vary depending on the corpora used for computing the cross-lingual similarities. The more comparable the corpora, the better. With that in mind, experiments were conducted over two different comparable corpora:

(4) News comparable corpus: the first experiment was conducted using a comparable corpus composed of news articles extracted from Die Zeit[5] newspaper in German (29M tokens) and from Berria[6] newspaper in Basque (36M tokens). No effort was done to match news topics or publications dates.

(5) Wikipedia comparable corpus: Wikipedia has been extensively exploited with NLP methods a comparable corpus (eg, Tomás et al. 2008; Paramita et al. 2012). In this case we constructed a corpus by gathering all articles that have both Basque and German versions, connected through wiki interlanguage links. The corpus has 61.484 articles per language and 91M tokens (72.5M tokens in German and 18M tokens in Basque). Although this corpus is highly comparable with respect to the topics (each article has its counterpart), it is important to note that the amount of tokens per language is unbalanced. This can lead to a decrease in the comparability degree of the corpus, because the German part holds much more information.

5 http://www.diezeit.de
6 http://berria.info

Table 2 shows the dictionaries used in the process and their statistics.

	#entries	#pairs
Lde-en (A) - Beolingus	146,451	171,775
Len-eu (B) - Elhuyar	17,672	43,201
Lde-eu (A+B, no pruning)	12,939	48,097
Lde-eu (IC1)	4,305	7,211
Lde-eu (IC1+DS wiki)	7,878	18,641
Lde-eu(IC1+DS news)	7,821	20,014

Table 2: Pivot dictionary process.

2.2.2 Parallel Corpus word-alignment

There is a large tradition of parallel corpus processing in computational linguistics, starting with the work of Gale & Church (1991), Brown, Lai & Mercer (1991), McEnery & Oakes (1995) and others (see Véronis 2000 for an overview). Different methods and tools have been proposed to align parallel texts and extract lexical correspondences from them. In this investigation, we used two word alignment tools based on these methods: Giza++ (Och & Ney 2000) and Bifid (Nazar 2012).

The fact that Basque is a medium density language unlike German represents an added difficulty for any attempt in the line of Resnik (1999), who proposed to download parallel corpora by mining the web for translated pages. In our experiments we used two parallel corpora of different sizes: a German-Basque Literary Corpus created in the context of recent research (Sanz Villar 2013; Zubillaga 2013) and another built by aligning Basque and German translations of the Bible[7]. The first one was compiled using the content of 81 digital or digitized and OCR-ed literary German originals and their official direct translations into Basque (146,457 segment pairs). In the case of the second, after removing the books not included in both Bible versions, a parallel corpus of 30,440 segment pairs was obtained using the verse as segment unit. This is an easily built resource for medium-density language pairs because the Bible is available for a wide variety of languages (Lardilleux, Gosme & Lepage 2010; Resnik, Olsen & Diab 1999) and is, therefore, an adequate baseline parallel resource for evaluating other extraction methods over this kind of language pairs.

The aforementioned extractions tools (Giza++ and Bifid) were used in order to obtain translations pairs. The resulting figures are shown in table 3.

7 Basque (Elizen Arteko Biblia 1994) and German (1984 revision of the Luther Bible).

	# of seg.	# of EU tokens	# of DE tokens	# of candidates GIZA++ p(b|g)>0.1	# of candidates BIFID
Literary Corpus	146,457	1,948,504	2,203,307	266,678	4,838
Bible Corpus	30,440	639,581	810,671	49,443	2,926

Table 3: DE-EU parallel corpora.

For word alignment with Giza++, the default sequence of models were used (IBM model 1, HMM-based model, IBM model 3 and IBM model 4). The German corpus was lemmatised and POS tagged by using Treetagger (Schmid 1995) and the Basque one using Eustagger (Ezeiza et al. 1998). Then, each word of the source corpus was substituted by a chain including the corresponding lemma and POS category. Punctuation marks and words with POS regarded as possible source of noise in the alignment process were removed from both corpora. Specifically, words excluded from the German part were those with POS tags such as APPR (preposition), APPRART (preposition with article), ART (article), KOKOM (particle of comparison), KOUS (subordinating conjunction), PRELS (relative pronoun), VAFIN (auxiliary verb, finite form) and VAINF (auxiliary verb, infinitive). From the Basque corpus, in turn, the excluded units were those with the ADL tag (auxiliary verb)[8].

Giza++ returns two files of word alignments (DE-EU and EU-DE) including a probability for each word alignment. For the draft BG, BibleGiza and LitGiza word alignments with a probability p(b|g) greater than 0.1 were selected from both Bible and Literary corpora.

In order to reduce noise, the BG obtained by Giza++ was then submitted to a filtering process using a stoplist consisting of the 150 most frequent Basque words[9]. Two versions of these BG have been evaluated: (1) Giza++ BG after stoplist filtering, and (2) after stoplist and a filtering that only allows BG entries with the POS-tags mapping to each other in one of the following ways (see table 4):

TreeTagger flag	**Eustagger flag**
NN (noun)	IZE (noun)
VV (verb)	ADI (verb)
ADV (adverb)	ADB (adverb)
AD (adjective)	ADB (adverb)[1]
AD (adjective)	ADJ (adjective)

Table 4: mapping of POS-tags TreeTagger (German) and Eustagger.

The other alignment tool, Bifid, is part of a larger project comprising the analysis of language pairs where no prior knowledge is available, which means that all forms of external resources are excluded from the processing. This tool incorporates modules for the integral process of analysing a set of do-

8 VAFIN and VAINF would be German equivalents of Basque ADL. The other removed German POS would be a morpheme in Basque words in almost all cases.

9 The Basque stoplist has been obtained from Basque ETC Corpus data (UPV-EHU, Sarasola et al. 2013).

cuments in unknown languages with the only assumption that such set consists of a parallel corpus in two languages. In its original version, this tool separates the set of documents in the two languages, aligns each document with its most probable translation and then proceeds to align the segments inside the documents (assuming that the newline character is the segment separator). Finally, from this segment alignment, it extracts an initial bilingual vocabulary which is then used for a realignment of the corpus at the segment level. The process is iterated in this way *n* times to improve the quality of the alignment at all levels.

In the case of this paper, however, we only used the bilingual lexicon extraction module because our parallel corpora were already aligned at the sentence level. The corpora were also lemmatized with the above mentioned tools, but no mapping exploiting the POS tags was used because this information is not used by the algorithm[10]. The bilingual vocabulary extraction module of Bifid uses a combination of strategies that include co-occurrence statistics as well as length and orthographic similarity metrics. As in its original version this extracted vocabulary was intended to be used for realignment, the program is very conservative in its lexical alignment in order to avoid the reproduction of errors in subsequent steps. As a consequence, it favors precision over recall, with fewer aligned pairs having a higher probability of being correct. Further experimentation will determine the right thresholds for the best compromise between noise and silence, meaning larger sets of aligned pairs with the maximum possible purity.

3 Evaluation

3.1 Comparison of German Lemma lists

Germanet offers the best recall on DeReWo lemmata. Wikipedia and Wiktionary on their own offer a similar recall; the intersections of both of these with DeReWo also reaches a very high level (see Table 5 below):

10 As the motivation behind Bifid is to be a language independent alignment tool, it does not use any kind of language-specific resources such as lemmatization or POS-tagging.

	Derewo		GermaNet		Wikipedia		Wiktionary	
∩ Derewo			32,199	33,73%	19,461	2,22%	22,028	7.01%
∩ GermaNet	32,199	80,50%			47,588	5,43%	34,309	10.93%
∩ Wikipedia	19,461	48,65%	47,588	49,86%			46,968	14.96%
∩ Wiktionary	22,028	55,07%	34,309	35,94%	46,968	5,36%		
∩ WikiORWikt	29,164	72,91%	59,995	62.86%				
Lemma total	39,998		95,449		876,309		314,016	

Table 5: DeReWo and existing lexicographical databases: German lemma counts (A-Z) and intersecting sets.

The best recall is offered by LitGiza, with a notable difference regarding the rest of drafts, even Bible-Giza. Pibolex recall is second best but far from LitGiza. BibleBifid and LitBifid offer a very low recall (see table 6).

	Derewo	Bible Giza Stop	Bible Giza StopPos	LitGiza Stop	LitGiza StopPos	Bible Bifid	Lit Bifid	Pibolex Wiki	Pibolex News
Derewo		4,639 (34.97%)	3,500 (38.69%)	15,775 (23.30%)	12,846 (24.93%)	1,007 (40.84%)	2,995 (67.27%)	5,812 (77.34%)	5,753 (77.14%)
BibleGiza Stop	4,639 (11.60%)		9,047 (100.00%)	5,001 (7.39%)	3,879 (7.53%)	2,372 (96.19%)	1,122 (25.20%)	1,868 (24.86%)	1,851 (24.82%)
BibleGiza StopPos	3,500 (8.75%)	9,047 (68.19%)		3,763 (5.56%)	3,248 (6.30%)	1,699 (68.90%)	957 (21.50%)	1,518 (20.20%)	1,504 (20.17%)
LitGiza Stop	15,775 (39.44%)	5,001 (37.70%)	3,763 (41.59%)		51,533 (100.00%)	1,125 (45.62%)	4,389 (98.58%)	4,571 (60.83%)	4,549 (60.99%)
LitGiza StopPos	12,846 (32.12%)	3,879 (29.24%)	3,248 (35.90%)	51,533 (76.12%)		935 (37,92%)	3,864 (86.79%)	3,960 (52.69%)	3,955 (53.03%)
Bible Bifid	1,007 (2.52%)	2,372 (17.88%)	1,699 (18.78%)	1,125 (1.66%)	935 (1.81%)		526 (11.81%)	544 (7.24%)	546 (7.32%)
Lit Bifid	2,995 (7.49%)	1,122 (8.46%)	957 (10.58%)	4,389 (6.48%)	3,864 (7.50%)	526 (21.33%)		1,404 (18.68%)	1,398 (18.74%)
Pibolex Wiki	5,812 (14.53%)	1,868 (14.08%)	1,518 (16.78%)	4,571 (6.75%)	3,960 (7.68%)	544 (22.06%)	1,404 (31.54%)		7,309 (98.00%)
Pibolex News	5,753 (14.38%)	1,851 (13.95%)	1,504 (16.62%)	4,549 (6.72%)	3,955 (7.67%)	546 (22.14%)	1,398 (31.40%)	7,309 (97.26%)	
Lemma total	39,998	13,267	9,047	67,699	51,533	2,466	4,452	7,515	7,458

Table 6: DeReWo and BG German entries intersections (A-Z, NLP methods).

3.2 Quantitative and Qualitative Evaluation of BG

Table 7 shows results of the qualitative evaluation carried out manually by a human lexicographer. EuDeLex is set as gold standard for comparison, regarding lemmalist and evaluation of TE appropriateness, in terms of (a) a full matching as suitable Basque TE for one of the word senses of the German BG headword (OK), (b) a semantic mismatch (FALSE), (c) a semantic (fuzzy) matching without being the TE a valuable equivalent to cite in a dictionary entry (NEAR), or (d) as PART, when a BG entry is a correct TE for a lemma as part of a Multi Word Expression in the other language. A second variable, part of speech (POS) is evaluated as matching (OK) or mismatching (FALSE).

The BG obtained from aligned WordNet synsets offers a relatively high recall on GS lemmata and, in absolute figures, the largest proportion of correct TEs. Wikipedia and Wiktionary extraction is less effective in terms of recall, but more effective with regard to TE adequateness.

Among the NLP methods, Giza and Pibolex BG offer the largest number of TE evaluated as correct, far ahead of Bifid, which on the other hand returned very little false TE and POS among the results. Giza BG, and, to a lower extent, Pibolex BG, suffer from a high percentage of inadequate, noisy TE.

	LitGiza Stop	LitGiza StopPos	BibleGiza Stop	BibleGiza StopPos	LitBifid	BibleBifid	Pibolex News	Pibolex Wiki	Wikipedia	Wiktionary	WordNet	ANY DRAFT
DE Letter A Lemma with EU TE	5,361	3,826	1,349	818	265	230	682	730	5,457	396	1,076	
Intersection with EuDeLex GS	1,276	696	434	323	177	49	559	567	265	147	774	2,928
Recall on 3406 GS lemmata	**35,41%**	**19,31%**	**12,04%**	**8,96%**	**4,91%**	**1,36%**	**15,51%**	**15,73%**	**7,35%**	**4,08%**	**21,48%**	**81,24%**
Ø TE per Lemma	2,27	1,73	2,35	1,46	1,07	1,22	2,64	2,82	1,00	1,33	2,57	2,28
Evaluated Lemma	1,276	696	434	323	177	49	559	567	265	147	774	4,248
Evaluated TE	2,901	1,206	1,019	470	189	60	1,476	1,601	265	195	1,988	9,694
• OK	**746**	**537**	**256**	**215**	**178**	**43**	**939**	**1,007**	**236**	**189**	**1,654**	**5,248**
• FALSE	1,923	519	728	230	3	15	414	513	6	3	188	3,793
• PART	186	122	11	6	6	1	14	7	4		17	246
• NEAR	46	28	24	19	2	1	109	74	19	3	129	407
Evaluated POS	2,901	1,206	1,019	470	189	60	1,476	1,601	265	195	1,988	9,694
• OK	1,071	965	448	422	186	52	1,321	1,286	245	195	1,983	6,787
• FALSE	1,830	241	571	48	3	8	310	315	21		5	3,063
Lemma with 1+ TE OK	609	439	213	189	168	39	443	454	236	147	700	2,081
Lemma with all TE OK	315	294	111	149	168	35	316	322	236	142	601	1,123
Lemma with something usable	444	232	129	62	7	6	165	164	23	5	143	1,056
Lemma with all TE FALSE	517	170	194	112	2	8	78	82	6		30	371
TE OK	**25,72%**	**44,53%**	**25,12%**	**45,74%**	**94,18%**	**71,67%**	**63,62%**	**62,90%**	**89,06%**	**96,92%**	**83,20%**	**54,14%**
POS OK	36,92%	80,02%	43,96%	89,79%	98,41%	86,67%	89,50%	80,32%	92,45%	100,00%	99,75%	70,01%
Lemma with 1+ TE OK	**47,73%**	**63,07%**	**49,08%**	**58,51%**	**94,92%**	**79,59%**	**79,25%**	**80,07%**	**89,06%**	**100,00%**	**90,44%**	**48,99%**
Lemma with all TE OK	24,69%	42,24%	25,58%	46,13%	94,92%	71,43%	56,53%	56,79%	89,06%	96,60%	77,65%	26,44%
Lemma with something usable	34,80%	33,33%	29,72%	19,20%	3,95%	12,24%	29,52%	28,92%	8,68%	3,40%	18,48%	24,86%
Lemma with all TE FALSE	40,52%	24,43%	44,70%	34,67%	1,13%	16,33%	13,95%	14,46%	2,26%	0,00%	3,88%	8,73%

Table 7: Manual Qualitative Evaluation.

4 Conclusions

4.1 Discussion

Combining all BDD methods presented here, we obtain a BG that covers more than 80% of a dictionary lemmalist based on DeReWo, and provides one or more correct TE for about a half of those.
For BDD purposes, correct TE must be separated from not suitable (noisy) BG entries; TE adequateness has to be the key criterion for lexicographical needs, before the amount of gathered data. In the ongoing editing process of EuDeLex, draft data will be divided in three groups, (1) methods with no or very little results evaluated as FALSE; the data obtained by those may be included in dictionary entries and published, without manual post-editing, (2) methods with high precision results (low degree of noise); the BG obtained by these methods could be pasted into the bilingual lexicographical database for manual post-editing, and (3) methods with a larger proportion of noisy results; the BG obtained by those will have to be post-processed in order to reduce false TEs; the POS-mapping approach presented here for Giza is a first step in that direction, enriching a Basque stoplist for results proposed by Giza from the list of Basque TEs that repeatedly have been evaluated as FALSE will be the next.
We propose to group the methods presented in this paper according to the criteria mentioned above as follows:

(1) Wiktionary, Wikipedia, LitBifid
(2) BibleBifid, WordNet
(3) LitGizaStop, BibleGizaStop, LitGizaStopPos, BibleGizaStopPos, Pibolex News, Pibolex Wiki

As we found out in this investigation, more than two thirds of the DeReWo list, on which a lemmalist for EuDeLex that covers the whole alphabet will base on, are linked to a dataset in Wiktionary and/or Wikipedia. The high rates in our qualitative evaluation reached by these methods encourage us to make use of them, and thanks to their open licence, it is possible. While other draft data needs human post-editing before inclusion in published bilingual dictionary entries, relevant data from those sources may be directly included in a dictionary search result webpage[11]. The recall these sources offer for German-Basque TE is still limited; it will mainly depend on the growth and activity of the Basque editor communities to increase it, which is, supposedly, a matter of time. Measurements like those proposed in this investigation may serve to monitor that process.
The Basque WordNet EusWN has been actively developed by human lexicographer teams at UPV-EHU, and it is today the largest and trustworthiest Basque lexical resource available with open data sources. The approach to align its synsets with GermaNet synsets using Princeton WordNet as pivot has been the one which delivered the largest proportion of correct TE for more German lemmata.
Pibolex overall results are similar to those obtained for other language pairs, confirming the tool performs robustly across languages. With respect to the corpora used in the experiments, the news corpus achieves slightly better results. This means that the Wiki corpus, although more comparable in terms of topics, suffers from the difference in amount of text between languages. The results obtained by both word alignment tools from two parallel corpora of different size show, as was to be expected,

11 Data found in these sources relevant to a bilingual dictionary is not only a TE, but also encyclopedic, phonetic, morphological, syntactic and pragmatic information or audiovisual material about a lemma.

that recall rates relate to corpus size, and the same is true for result precision. A further development of German-Basque parallel corpora is strongly desired.

In spite of German-Basque being a medium-density language pair with limited bilingual lexical and corpus resources, the amount of adequate BG entries gathered during the presented experiments is considerably high, and it will help saving human resources in dictionary writing. There is no need to say, however, that human lexicographers are still the key factor for a German-Basque dictionary writing that would meet acceptable quality standards.

4.2 Future Work

A future line of work would be to create higher comparability degree corpora, taking care of maintaining balance in terms of size, topics and genres across languages, without decreasing the overall size of the corpora. Further research about BDD would also include rendering optimization of the applied corpus based methods for the language pair German-Basque (enhancements of corpus tagging, word alignment stoplists and parameter tuning), a sophistication of these (e.g., by making use of syntactic information), as well as reproducing these experiment sets for other language pairs, which would allow for a comparison of results. Other goals not achieved at the present stage are the inclusion of multiword expressions in the experiments and measurements of polysemy covered by draft BG. We are now centering our efforts in developing a new method for the exploitation of Wikipedia as a comparable corpus using the frequency distribution of lexical units in the articles. We are representing the relative frequency of words in the articles as curves, and then comparing the curves in a purely geometrical fashion using Euclidean distance. We assume that German and Basque words with similar frequency curves will be equivalent, however we still need to find the way to make up for the already mentioned asymmetry in the amount of text in the corresponding articles in both languages.

5 References

Brown, P.F., Lai, J.C. & Mercer, R.L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 91, Stroudsburg, PA.: Association for Computational Linguistics, pp. 169–176.

Ezeiza, N., Alegria, I., Arriola, J.M., Urizar, R. & Aduriz, I. (1998). Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 17th international conference on Computational linguistics*. Association for Computational Linguistics, pp. 380–384.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. MIT Press.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the 3rd Workshop on Very Large Corpora*. Boston, MA., pp. 173–183.

Gale, W.A. & Church, K.W. (1991). Identifying Word Correspondences in Parallel Texts. In *Proceedings of the ACL Workshop on Speech and Natural Language*. ACL 91, Stroudsburg, PA: Association for Computational Linguistics, pp. 152–157.

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: Association for Computational Linguistics, pp. 9–15.

Lardilleux, A., Gosme, J. & Lepage, Y. (2010). Bilingual lexicon induction: Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*. LREC 2010, Malta, pp. 252–256.

Lindemann, D. (2014). Zweisprachige Lexikographie des Sprachenpaares Deutsch-Baskisch. In Domínguez Vázquez, M.J., Mollica, F. & Nied, M. (eds.) Zweisprachige Lexikographie im Spannungsfeld zwischen Translation und Didaktik, Lexicographica Series Maior. De Gruyter.

McEnery, A.M. & Oakes, M.P. (1995). Sentence and word alignment in the CRATER project: methods and assessment. In *Proceedings of the EACL-SIGDAT Workshop: from texts to tags, Issues in Multilingual Language Analysis (ACL)*. Dublin, pp. 77–86.

Nazar, R. (2012). Bifid: un alineador de corpus paralelo a nivel de documento, oración y vocabulario. In *Linguamática*, 4, 45–56.

Och, F.J. & Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 00, Hongkong: Association for Computational Linguistics, pp. 440–447.

Paramita, M.L., Clough, P., Aker, A. & Gaizauskas, R.J. (2012). Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In *Proceedings of the Eighth conference on International Language Resources and Evaluation*. LREC 2012, Istanbul, pp. 790–797.

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

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 99, College Park, MD.: Association for Computational Linguistics, pp. 519–526.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 99, College Park, MD.: Association for Computational Linguistics, pp. 527–534.

Resnik, P., Olsen, M.B. & Diab, M. (1999). The Bible as a parallel corpus: Annotating the "Book of 2000 Tongues." In *Computers and the Humanities*, 33, 129–153.

Sanz Villar, Z. (2013). Hacia la creación de un corpus digitalizado, paralelo, trilingüe (alemán-español-euskera). In Sinner, C. & Van Raemdonck, D. (eds.) Fraseología contrastiva del alemán y el español. Traducción y lexicografía, Études linguistiques Linguistische Studien. München: Peniope, pp. 43–58.

Saralegi, X., San Vicente, I. & Gurrutxaga, A. (2008). Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *Proceedings of the 1st workshop on Building and using Comparable Corpora (BUCC)*. LREC 2008, Marrakech.

Saralegi, X., Manterola, I. & San Vicente, I. (2011). Analyzing Methods for Improving Precision of Pivot Based Bilingual Dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP 2011, Edinburgh.

Sarasola, I., Landa, J. & Salaburu, P. (2013). Egungo Testuen Corpusa. UPV-EHU University of the Basque Country

Schmid, H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*. Dublin, pp. 47–50.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. LREC 2006, Genoa.

Tanaka, K. & Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 297–303.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth conference on International Language Resources and Evaluation*. LREC 2012, Istanbul, pp. 2214–2218.

Tomás, J., Bataller, J., Casacuberta, F. & Lloret, J. (2008). Mining wikipedia as a parallel and comparable corpus. In *Language Forum*, 34.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L. & Trón, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*. Borovets, pp. 590–596.

Véronis, J. (2000). Parallel Text Processing: Alignment and use of translation corpora. Kluwer.

Zubillaga, N. (2013). Alemanetik euskaratutako haur- eta gazte-literatura: zuzeneko nahiz zeharkako itzulpenen azterketa corpus baten bidez. PhD Thesis. Vitoria-Gasteiz: UPV-EHU.

Acknowledgements

This study has been supported by the following projects: IT665-13, Zubiak (Saiotek-SA-2013/00308) and Ber2tek (Etortek-IE12-333), funded by the Basque Government; and project EC FP7/SSH-2013-1 AThEME (613465), funded by the European Commission. Funding is gratefully acknowledged.