Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 266 – 277

Procedia Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Building corpus-based frequency lemma lists

David Lindemann[a]*, Iñaki San Vicente[b]

[a]*UPV/EHU University of the Basque Country*
[b]*Elhuyar Foundation*

Abstract

This paper presents a simple methodology to create corpus-based frequency lemma lists, applied to the case of the Basque language. Since the first work on the matter in 1982, the amount of text written in Basque and language resources related to this language has grown exponentially. Based on state-of-the-art Basque corpora and current NLP technology, we develop a frequency lemma list for standard Basque. Our aim is twofold: On the one hand, to propose a primary Basque lemma list for a bilingual dictionary that is currently being worked on at UPV/EHU, and on the other, to contrast existing Basque dictionary lemma lists with frequency data, in order to evaluate the adequacy of our proposal and to compare lemma lists with each other.

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: Lexicography; Corpus Linguistics; Lemma Frequency; Basque Language

1. Introduction

Lexicography today resorts to corpora in order to find new dictionary entries; but how to decide whether a term is important enough to appear in a dictionary? Natural Language Processing techniques offer a chance to gather statistical information about words, such as information about the usage of candidate headwords, which helps us to gain evidence for a need to include it in a dictionary macrostructure. The evidence mostly referred to is frequency. For lexicographers, frequency data is important in three regards:

* Corresponding author. Tel.: +34-945-013-148; fax: +34-945-013-200.
E-mail address: david.lindemann@ehu.eus

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.445

1. The usage frequency of a lemma, which we can measure with corpus methods, is related to the look-up frequency of that lemma in dictionaries (De Schryver et al. 2010; Wolfer et al. 2014);
2. Frequency data is useful information for a lexicographer in the dictionary editing process;
3. Frequency data may be included in a dictionary microstructure (that is, the body of a dictionary entry), so that the dictionary user gets access to it. This is particularly useful for dictionary users who look up words in their L2. English learners, for instance, from the 1990s onwards find frequency information in dictionaries that have been designed for them (Kilgarriff 1997).

For this study, we have developed a frequency lemma list for Basque, following the model of state-of-the-art frequency lemma lists for German that are published under the title DeReWo. We have followed a double motivation:

1. to define a Basque lemma list for a first edition of the Basque→German part of EuDeLex, a bilingual dictionary currently being developed at UPV/EHU, and
2. to survey and prove criteria for including headword candidates in a dictionary lemma list, and to propose a methodology.

Having these goals in mind, we have assembled frequency lemma lists extracted from corpora, and examined the appropriateness of the evidence gained from these data sets. Furthermore, we have had the chance to compare the corpus-based lemma lists with the lemma lists of some Basque dictionaries. It is important to point out again that the aim of our work has not been to enrich an existing dictionary with more entries but to set up a the macrostructural content for a Basque dictionary from scratch.

This paper is organized as follows: In the following part, we give a short survey of the investigation on frequency lemma lists. In part 3, we present the language resources and the methodology used in this study. Part 4 is dedicated to the experiments we carried out and the results we have obtained, and part 5 offers reflections about these, some conclusions, and an outlook on future work on this issue.

2. The state of the art

2.1. German corpus-based frequency lemma lists

In order to provide resources for lexicography and other branches of linguistics, the Mannheim-based *Institut für Deutsche Sprache* publishes frequency word form and lemma lists under the title DeReWo, based on the DeReKo corpora (see Kupietz et al. 2010), which contain literary, scientific and press text from 1980 onwards. In 2014, these corpora counted 25 billion tokens.

Raw frequency data extracted from corpora to be valuable as headword candidate list in lexicography, some automatic, semi-automatic and handmade working steps are necessary, as frequency lists built entirely by automatic methods do not contain only accurate data. On the one hand, in order to assign a <lemma> or <lem-pos> (lemma and part of speech) pair to every word form on the list —in other words, to reach from word form frequency data to lemma frequency— the corpus has to be lemmatized and furnished with morphosyntactic tags, which is done by a linguistic tagger. On the other hand, for a lemma as headword candidate, a minimum frequency threshold has to be defined, that is, a minimal count of usage examples a lexicographer needs for defining the headword's semantic value or values (Sinclair 2005). For unigrams (single word lemmata), this threshold has been set to 20 occurrences (*ibid.*), but more factors have to be taken into account, such as the polysemy of a lemma and the number of homographs to it and their parts of speech.

In the case of DeReWo (see IDS 2009), the general method for building headword lists from frequency data can be summarized as follows:

1. The intersections of the frequency lemma list extracted from corpora with previously existing dictionary lemma lists are taken as accurate headword candidates;
2. The remaining list entries, those absent from previously published dictionaries, are classified by means of semi-automatic methods (that is, in groups) or by hand as accurate headword candidates or not.

The DeReWo-40,000 frequency lemma list (IDS 2009) has been used as raw material for editing a German lemmalist for the EuDeLex bilingual dictionary. After having edited around 10% of the list entries by hand, 95% of these could be taken directly into the dictionary lemma list without any changes, and the list entries that needed manual editing and dictionary lemmata that were included by the lexicographers despite not appearing on the frequency list sum up not more than 5% (Details in Lindemann 2013, 2014). These figures show to what extent the frequency list is suitable as a resource for defining a dictionary macrostructure, as very little editing by hand is needed.

2.2. Basque frequency dictionaries

1. Sarasola, Ibon (1982): *Gaurko euskara idatziaren maiztasun-hiztegia: 1977ko corpus batean oinarritua*. Donostia: Gipuzkoako Aurrezki Kutxa Probintziala (Sar82). This first Basque frequency dictionary is based on a 800,000-token corpus of press and literature texts published along the year 1977 that were selected, processed, digitized and pos-tagged by hand. The lemmata that were not found on the 2,000-lemma list declared by the Basque language academy *Euskaltzaindia* as standard at that time[†] were grouped under their most frequent graphical form. With 19 occurrence counts as minimum threshold, they calculated frequency data for about 3,000 lemmata, related to the whole corpus as well as to ten sub-corpora. For this study, we took into account the list of 2,945 lemmata with 19 or more occurrences.
2. Etxebarria, J.M. & Mujika, J.A. (1987): *Euskararen oinarrizko hiztegia: maiztasun eta prestasun azterketa*. Gasteiz: Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia. For their study, the authors built a corpus of 130 handmade transcriptions of private conversations and radio broadcasts. The grouping of dialectal forms of the lemmata and the lemmatization itself were also done by hand. The dictionary counts 3,154 headwords, having set the minimum occurrence threshold to 2. In order to obtain relations between dialect and frequency, they separated Bizkaian and Gipuzkoan dialect sub-corpora and calculated frequency for both. Furthermore, their work contains 12 specialised domain dictionaries showing the occurrence counts in the voice recording transcription corpora. These dictionaries are not included in the main frequency dictionary, and they don't calculate the intersections between the two groups of data. For the present paper, we have not taken this data into account, as it is not available in a digital form.
3. UZEI (2004). *Maiztasun Hiztegia*. Donostia: UZEI (UZEI04). A frequency dictionary based on the 4.5 million token corpus *XX. mendeko corpus estatistikoa* (Urkia 2002). After removing proper names, auxiliary verb forms and non-free morphemes with lexical value, among others, the amount of tokens taken into account for calculating frequency data has been about 3.7 million. Every entry contains information of two kinds: Lemma and graphical form(s). The dictionary headwords are standard Basque lemmata, and below each of them, the frequency of its dialectal forms is given with a diachronical split (1900-1939, 1940-1968, 1969-1990 and 1991-1999), so that the usage of a lemma can be compared in time, text domain and dialect. Parting from these data, the dictionary offers absolute and relative frequency figures. For the present paper, we took into account the sums of the split occurrence counts for each lemma in its standard Basque form.

3. Experiments

As remarked in chapter 1, the goal of this study is primarily to build a frequency-based lemma list that can serve as starting point for the editing of the Basque→German part of the EuDeLex dictionary. On our way towards this goal, we have worked out a methodology for setting up that lemma list, and we have compared it to the other corpus-based frequency and lexical resources we had access to, in order to get evidence for the accurateness of our results.

[1] The standardization of the Basque language, lead by the language academy *Euskaltzaindia*, began in 1968 and is an ongoing process (see Hualde & Zuazo 2007).

3.1. Frequency lemma lists

For the present study, we had access to data from the following Basque corpora, which today are the biggest corpora for Basque:

1. Egungo Testuen Corpusa (ETC) (Sarasola et al. 2013). 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. For the present study, we have used a frequency list extracted from that corpus that contains lemmata with at least 10 occurrences.
2. Elhuyar Web-corpus (Elh124) (Leturia 2012). This corpus was built in 2012 and counts 124,625,420 tokens. It contains text of all kinds and domains from the internet. For setting up this corpus, a fully automatic method was used: Starting from a set of seed words, queries consisting of combinations of these were sent to online search engines. The documents obtained by these search queries were collected.
3. Elhuyar Web-corpus (Elh200) (Leturia 2014). A corpus built by Igor Leturia for his PhD thesis by fully automatic methods. It contains 200 million tokens. It was set up by web crawling, starting from a set of Basque web documents: Following the hyperlinks found in these and saving the documents those links lead to, a new set of documents is obtained, that serves to repeat the process.

The ETC corpus was built in a controlled way; one can say it consists of text that reflects a commendable use of standard Basque, as only certain sources were taken into account. In opposition to that, the Elh124 and Elh200 corpora were built in an "opportunist" way, without controlling domain or style, so that texts that are beyond a commendable standard were also accepted. In spite of that fact, Leturia (2012) measures the accurateness of this corpus for lexicography, comparing it to the *XX. Mendeko Corpus* (Urkia 2002) and to the *Lexikoaren Behatokiko Corpus* (Euskaltzaindia 2009), which reflect standard Basque language in use. His experiments show that those web-corpora contain a large number of lexical items not found in the latter two, and that consequently they are suitable for a use in a lexicographical workflow.

In addition to these corpora, we also had access to two other resources: Sar82, the first Basque frequency dictionary, on the one hand, and UZEI04, on the other.

3.2. Frequency lemma list processing

Building a frequency lemma list from a corpus as starting point, the raw material consisting of word forms has to be processed, in order to extract frequency data that belong to lemmata, as explained in chapter 2. For this reason, the first working step has to be a linguistic tagging of each token in the corpus.

In the case of ETC, we had access to a frequency lemma list, each entry of which carried a lemma and frequency information. For Elh124 and Elh200, the linguistic tagging has been done by us, using the *Eustagger* tool (Aduriz et al. 1996) The extracted frequency lemma lists carry lemmata and a three-level part of speech information: Frequency data regarding (1) lemma signs without PoS-disambiguation, (2) the main part of speech category (noun, verb, etc.), as well as (3) a more fine-grained tagging including the part of speech subcategory (common noun, proper noun, place name, etc.). We have had our computers calculate frequency data on all of these three levels, as shown for the example lemma-sign "*alegia*", that might be a noun, a place name, or a conjunction, in chapter 4.3.1 below.

The frequency lemma lists extracted from big web corpora must not be taken as good without further post-processing, as they contain "noise", that is, inaccurate headword candidates. In relation to that, a series of problems can be named:

1. Words from other languages that are homograph to Basque lemmata (e.g. el, that appears on our Basque corpus-based frequency list, and is homograph to the very frequent Spanish article or pronoun el, as well as to the Basque el, a very unfrequent lemma that only appears in OEH, see chapter 4.5.1 below). This is particularly problematic as it leads to contaminate the frequency data for the real Basque lemma.
2. Words from other languages that are not homograph to any Basque lemma (e.g. the Spanish con)

3. Errors in automatic linguistic tagging: incorrect lemmatisations (e.g.. ona>*ona -on-) and incorrect POS-tagging (e.g. basiliko, *adjective -noun-)

As possible workarounds for these problems, we have tested the following strategies:

1. If Eustagger along the corpus has used a high number of different POS tags for the same lemma, to suppress that lemma;
2. To suppress a lemma if it appears to have a high frequency rank on our list, but has no usage examples in the corpus-based OEH dictionary, as it has been proposed before to use the amount of usage examples in OEH as indicator for frequency (Sarasola et al. 2008);
3. To suppress a lemma if its rfreq value (see equation 1) appears to be very different for the two processed big web-corpora.

When comparing lemma lists, another problem we have to deal with is the fact that the frequency lemma lists are not directly comparable with each other. On the one hand, their source corpora are of a very different size, and on the other, the granularity level of the frequency data each of them contains is also different, as explained above. In order to perform comparisons, we have carried out a “standardisation” of all lemma lists. Having in mind the limits of the lemma lists, we have applied the following methodology for this standardisation:

a) We have limited the lemma lists to one-word lexical items (unigrams);
b) We have focused on lemma frequency, that is, we have grouped POS-split-frequencies and frequencies split according to graphical variants under the same lemma;
c) Absolute frequency data is not suitable for comparisons between corpora of different sizes. In fact, it is not the same for a lemma to have 30 occurrences in a 2 million token corpus as in a 20 million token corpus. In order to solve this problem, we have calculated an occurrence percentage value, a relative frequency, calculated according to formula (1) below.

$$rfreq_i = \frac{100 * f_i}{N} \tag{1}$$

where f_i is the absolute frequency of lemma *i* in a corpus, and N is the number of tokens in that corpus.

3.3. Dictionary lemma lists and lexical resources

The corpora that were used for extracting the frequency lemma lists we have used in this study as reference (Sar81, UZEI04) are of a very different size in relation to the corpora we have used. In order to carry out a deeper analysis, we have also compared the frequency lemma lists to the content of some handmade lexical resources, that serve as reference or “gold standard”:

a) The lemma list of the Language Academy's *Hiztegi Batua* (Euskaltzaindia 2008). It counts 35,640 headwords (counting homographs only once).
b) The lemma list of the *Orotariko Euskal Hiztegia* (OEH) (Mitxelena & Sarasola 1988). It counts 89,296 headwords and 36,676 headwords for subentries (counting homographs only once).
c) 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 Natural Language Processing 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 EDBK, 64,737 are lemmata (EDBL_lemma).
d) The *Elhuyar Euskara-Gaztelania* (Spanish-Basque) dictionary's Basque lemma list (ElhDic) (Elhuyar Hizkuntza Zerbitzuak 2013), wich countains 64,459 lemmata (counting homographs only once).

e) The 26,886 lexical items (homographs counted only once) contained in *EuskalWordNet* (EusWN) (Pociello 2007; Pociello et al. 2011). Not all of these Basque lexical items, that are a part of synsets which are linked to the English Princeton WordNet (Fellbaum 1998), are canonical forms that may be taken for a lemma, as they are direct equivalents of English lexical items. As equivalents for English nouns, for instance, substantival derivatives of Basque verbs (*-t(z)e* or *-tasun*) often appear.

Before storing all data in the same database, we have carried out a graphical standardisation of lemmata: Hyphens and spaces in between words both were replaced by low hyphens ("_"), in order to get lexical items such as "lan-bulego" and "lan bulego" together, all other hyphens were suppressed, and high case letters were replaced by low case. The unified database counts 1,905,263 headwords. Below each of these, the headwords of each source that are homographs to them were stored together with the corresponding frequency and other data.

4. Results

4.1. Frequency list similarity

How divergent are lemma frequency lists extracted from different corpora? Which is the contribution they can make one to another? In order to systematically compare *Rank* values we have used two methods.

On the one hand, following the approach of Kilgarriff (2001), we have compared the ranks of frequency lists by means of the Spearman Rank Correlation. This measure computes the distance between two rankings by comparing the positions the elements have in one ranking and the other, as the equation (2) describes. The comparison is done over the *n* most frequent lemmata taken from the union of the two corpora we want to compare.

$$\rho = 1 - \frac{6\sum(x_i - y_i)^2}{n(n^2-1)} \tag{2}$$

where x_i and y_i are the position the element *i* has in the 1st and 2nd rankings, respectively.

Table (1) shows results for the *n=500* values. Corpora are sorted by publication date. We can see that Elh124 and Elh200 yield a very high similarity. Even if Elh200 contains roughly 80M tokens more that Elh124, the facts that both corpora were built by massively crawling the Internet and that both corpora were tagged using the same tools could explain the high similarity score achieved. It is also noteworthy that the highest correlation coefficients are found between corpora which are closest in time. For example, the most similar lemma list to Sar82 is UZEI04, while newer corpora show far lower correlation coefficients.

Table 1. Comparison of the 500 most frequent lemmata in the frequency lemma lists according to the Spearman Rank Correlation coefficient.

	Sar82	UZEI04	ETC	Elh124	Elh200
Sar82	1				
UZEI04	0.5542798107	1			
ETC	0.4056939783	0.4641690074	1		
Elh124	0.4150037011	0.5536959952	0.5627681191	1	
Elh200	0.3805365731	0.4675802057	0.5265724492	**0.9009107076**	1

On the other hand, we reproduced the experiment carried out by Leturia (2014), based on the approach proposed by Baroni et al. (2009). The comparison includes all the frequency lemma lists we had available. The approach compares two corpora in terms of lemma *coverage* and *enrichment*, i.e., to what extent the lemmata in a corpus are covered by the other and how many lemmata has a corpus that the other has not.

Let us say that we have two corpora C_a and C_b. *Coverage* measures how many of the lemmata that reach a minimum frequency t (t=20) in C_b occur also at least t times in C_a (see equation (3)). This would describe which is the proportion of lemmata in C_a for which C_b can also provide information. *Enrichment* measures the proportion of lemmata that occur less than t (t=20) times in C_a but more than t times C_b, with respect to the total number of lemmata below the frequency threshold t in C_a. This would indicate the proportion of lemmata for which C_b can provide information but C_a can not.

$$coverage(C_a/C_b) = \frac{N_a \cap N_b}{N_a} \tag{3}$$

where N_a is the number of lemmata in C_a whose frequency>=20 and N_b is the number of lemmata in C_b whose frequency>=20

$$enrichment(C_a/C_b) = \frac{N_b}{S_a} \tag{4}$$

where N_b is the number of lemmata in C_b whose frequency>=20 and S_a is the number of lemmata in C_a whose frequency<20

Results on table (2) show that the majority of the lemmata in the Sar82 dictionary are covered by the other lemma lists, as we had expected, and in contrast, those have a low coverage with respect to the other lemma lists. Enrichment results show the same tendency (see Table (3)): most of the other lemma lists can provide information about lemmata in Sar82 that do not occur with a minimum frequency. If we take a look at the more up-to-date corpora, we can see that Elh124 and Elh200 lemma lists have a high coverage with respect to ETC, while this does not happen the other way round: ETC only covers the 40% of the lemmata in Elh124, and 34% of the lemmata in Elh200. Again, the same tendency is reflected by the enrichment results: ETC only offers a 5-7% enrichment over the other two, while Elh124 and Elh200 lemma lists provide around 20% enrichment over ETC. This behaviour could be explained by the composition of the corpora. ETC is restricted to content gathered from specific sources which limits the heterogeneity of the corpus when compared with Elh124 and Elh200, that allow for any Basque content found in the Internet.

Table 2. coverage results. Each row shows the coverage a certain corpus obtains in the other corpora. In contrast, each column shows the coverage the other corpora offer respect to a certain corpus.

	Sar82	UZEI04	ETC	Elh124	Elh200
Sar82		26.13%	7.11%	4.13%	3.36%
UZEI04	91.76%		22.14%	14.13%	11.51%
ETC	93.78%	83.12%		41.58%	34.63%
Elh124	95.38%	93.04%	72.91%		89.80%
Elh200	95.52%	93.08%	74.61%	91.70%	

Table 3. enrichment results. Each row shows the enrichment level a certain corpus has with respect to the others. In contrast, each column shows the enrichment the other corpora offer respect to a certain corpus.

	Sar82	UZEI04	ETC	Elh124	Elh200
Sar82		0.60%	0.20%	0.02%	0.01%
UZEI04	80.41%		0.64%	0.16%	0.14%
ETC	94.85%	66.08%		5.39%	7.83%
Elh124	93.81%	82.24%	18.78%		10.84%
Elh200	94.85%	82.65%	21.82%	28.26%	0.01%

As one summarizing conclusion for this section, we can point out that Elh124 and Elh200 corpora show a high similarity in terms of lemmata. Hence, for the sake of simplicity, the following sections report results solely for the Elh200 corpus. The reason to favour Elh200 over Elh124 is that it provides information about a higher number of lemmata according to the enrichment index (28.26%). Also, it is more comparable to ETC in terms of size.

4.2. Dictionary lemma lists in the corpora

In order to compare the sets of lemmata that are headwords in one of the reference dictionaries to the sets of lemmata extracted from the corpora, we have calculated intersections and disjoint subsets for unigrams. All dictionary headwords count 135,251, 75,622 of which occur in the corpora. In figure (1) below, the distribution of the subsets is shown for the five existing lexical resources mentioned in chapter 3.3 above. As we had expected, the OEH lemma list contains the relatively highest amount of lemmata that do not occur in the corpora, as that dictionary contains lexical items and spellings from contemporary language use. EDBL_lemmata are the set containing the relatively biggest intersection with the corpus lemmata, while nearly a half of the disjoint subset (EDBL_lema [61,803] \ (ETC ∪ Elh200)) consists of proper names and place names (4,703 of 10,505 headwords).

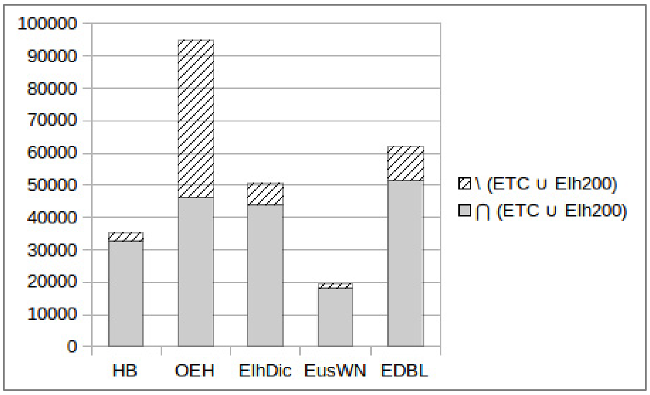


Fig. 1. Dictionary lemma lists in the corpora

4.3. A basic dictionary lemma list

4.3.1. Unigrams

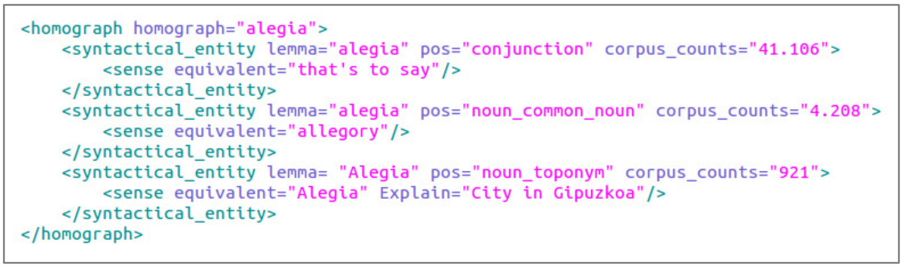
With the data and methodology mentioned in section 3 as starting point, we propose a basic lemma list for a language learner's dictionary, which is based on lemma frequency, with a goal of furnishing the entries of that dictionary with frequency data for every syntactical entity belonging to a headword. In a first step, we have filtered the unigrams that occur in either ETC or Elh200 as well as one of the handmade dictionaries. We will now have a look at this intersection between corpus and dictionary lemmata, that counts 75,481 headwords, and we will apply more filters. As for the EDBL_lemmata, we obtain the following figures: In the intersecting set with the corpora of

51,157 lemmata, 46,124 show an occurrence count of 20 or more. These lemmata may be grouped according to the EDBL syntactical tags, on two granularity levels: According to the main Part of Speech (POS) category (noun, verb, etc.) , or according to the POS-subcategory, such as common noun, proper name, and place name for nouns. In table (4), we show the data for the example *alegia* for three granularity levels:

Table 4. Frequency data from Elh200 for the lemma *alegia* on three granularity levels

	Frequency list rank	**Occurrence counts**	**POS**
NoPOS	539	46,237	(lemma)
POS	609	41,106	conjunction
	3,378	5,129	noun
	1,475,970	1	adverb
	1,507,920	1	adjective
POS_POS2	618	41,106	conjunction
	3,882	4,208	noun
	10,407	921	place name
	1,440,023	1	adjective
	1,880,968	1	adverb

In fact, a basic dictionary entry structure may be built upon the EDBL syntactical tags: Below a headword, all homograph lemmata can be placed as syntactical entities, as shown in the example *alegia* in figure (2) below. Disregarding the data for entities with an occurrence count below 20, we obtain a suitable data set for a basic dictionary entry design. This is a second good reason for our decision to use the intersection of EDBL_lemmata and the corpora as starting point for building a basic lemma list for Basque.



```
<homograph homograph="alegia">
    <syntactical_entity lemma="alegia" pos="conjunction" corpus_counts="41.106">
        <sense equivalent="that's to say"/>
    </syntactical_entity>
    <syntactical_entity lemma="alegia" pos="noun_common_noun" corpus_counts="4.208">
        <sense equivalent="allegory"/>
    </syntactical_entity>
    <syntactical_entity lemma= "Alegia" pos="noun_toponym" corpus_counts="921">
        <sense equivalent="Alegia" Explain="City in Gipuzkoa"/>
    </syntactical_entity>
</homograph>
```

Fig. 1. basic xml entry structure for *alegia* (sense group simplified)

On the list of corpus lemmata that don't occur in EDBL but in any other of the reference dictionaries (24,324 headwords), 4,398 occur 20 times or more in the corpora. Thus, that subset contains more headword candidates. In order to filter out proper headword candidates, and to supply EDBL-like syntactical tags to them, we group them according to the source of occurrence. In table (5) below, we see the distribution of this subset in the remaining four lexical resources.

Table 5. Corpus lemmata not ocurring in EDBL but in the other lexical resources

∩ ((ETC ∪ Elh200) \ EDBL_lema) [24,342]		freq>=20 [4,398]
HB	4,405	1,735
OEH	19,987	2,864
ElhDic	3,854	1,645
EusWN	1,422	464

According to our proposal for the creation of a new dictionary, we shall first supply syntactical tags to these lemmata. On the other hand, we need to introduce further filter criteria for separating the proper headword candidates from the inaccurate. In fact, this group contains a high amount of "noise" (estimated 4,000 list entries that are inaccurate headwords) in all part of the list, i.e. among the most frequent as well. We shall develop methods for noise elimination for each source subset.

More than two thirds of this group consists of OEH entries. This is the case for the top three of the list: '*duela', 'el', 'ona'*. '*Duela'*, a non-canonical form consisting of a chain of an inflected form of the verb **edun* and the non-free morpheme -*la*, is found in OEH as subentry headword of the verb '*izan'.* '*El' is* found there with a reference to two classic Basque dictionaries from long before Basque standardization (Añibarro, *Voces Vascongadas* (around 1800), and Azkue's dictionary (1905-06), and it is a Basque equivalent for the Spanish coin '*real'* used at its time. Obviously, the vast majority of the corpus occurrences of this word (if not all of them) will rather be the Spanish article and pronoun than refer to an ancient coin. The third case, the word '*ona'*, used to be a Basque equivalent for the Spanish '*Doña*', and, by mistake of the automatic tagger, it has not been identified as the adjective '*on*' in a chain with the determiner "-*a*" but as a canonical form with the range of a lemma. To avoid wrong pairings like the described, we would have to take into consideration additional methods, as, for instance, having a look at the contexts of the occurrences. But also the data we already have do give us a clue for how to filter this kind of phenomenon: The linguistic tagger, which uses EDBL as parameter file, in these cases has tagged those supposed lemmata in a way that is not in concord with the part of speech information found in OEH for them. That is to say, the wrong pairing shows an inconsistency in its morphosyntactic metadata, a fact that we could make use of for designing filter criteria. In those cases, as the tagger has not found the proper corresponding lemma (for *duela* and *ona*) or it has not found a way to link it properly (*el*), it has "invented" the tags spontaneously. The following table (6) shows these three supposed lemmata and the tags set by *EusTagger*:

Table 6. Inaccurate syntactical tags

word	POS-tags (*EusTagger*)
duela	adjective
ona	adverb; adjective; noun (place name)
el	noun; interjection

Summarizing our method for defining a basic lemma list for a first edition of EuDeLex, we point out the following: Taking the methodology for developing the DeReWo corpus-based frequency lemma lists as model to follow, we have calculated the intersection of an existing handmade lemma list and the frequency lemma lists extracted from the corpora. This handmade lexical resource is EDBL, which we have chosen for two reasons: (1) The intersection of EDBL and the corpora is the relatively biggest; (2) The EDBL data provides evidence also for a basic dictionary microstructure, i.e. syntactical entities that in the next step shall be furnished with semantic information. As the editing process of the new dictionary goes further, we will have the chance to measure the adequateness both of the lemma list itself and of the syntactical entities below each headword obtained by this automatic method, which will also allow to give helpful feedback to the EDBL developers. The next step will be the inclusion of further lemmata from among the headword candidates that are not included in EDBL but in other reference dictionaries. At last, we may also include other lemmata that are frequent in the corpora but not found in any other dictionary.

4.4. Headword candidates for a Basque dictionary

Among the 946,360 entries on the automatically lemmatized list extracted from the corpora that are not found in any dictionary, a great many are "noise", like false lemmatizations (inflected forms, forms that contain non-free morphemes or non-standard spellings that have passed the automatic lemmatization process without any change), words in other languages, and Roman numbers, among others. In this vast set, neologisms, loans from other languages, non-standard but very frequent spellings and others that may be considered proper headword candidates are "hidden". We have carried out experiments based on frequency, such as filtering list entries with very different relative frequency in different corpora, applying different thresholds. Unfortunately, the methods we have chosen do not separate the noise from the proper candidates. As predicted in section 2.1, frequency alone is not informative enough to guide headword selection, so other methods, such as manual work by lexicographers, remain necessary.

5. Conclusions and further work

Based on the largest Basque corpora available today, we have built a lemmatized frequency word list. We have compared the content of that list to some dictionary lemma lists, including two previously published frequency dictionaries. The intersecting sets of corpus lemma lists and dictionary lemma lists —that is, the dictionary lemmata whose usage is verified in corpora— may be used for a basic lemma list of a bilingual dictionary like EuDeLex. Furthermore, frequency data related to each of these lemmata may be included in the dictionary entry, in three different granularity levels.

In relation to the dictionary lemma lists we have used in our experiments, on the one hand, we have pointed out the value of the EDBL lexical database, which has been developed foremost for Natural Language Processing tasks, as it provides evidence for furnishing both dictionary macrostructure and microstructure with basic information. On the other hand, taking lexical data extracted from EDBL as starting point and completing that with data from other sources will possibly contribute to an upgrade of this database.

Very frequent lemmata from the corpora that do not appear in any of the reference dictionaries may be candidates for inclusion in a dictionary lemma list. Nevertheless, our experiments have shown that frequency alone is not enough for being able to filter the “good” candidates. It will belong to future work to develop more complex statistical models and other methods that would distinguish the “noise” in an accurate way and allow to treat it appropriately. At this stage, we find that manual work is a not replaceable resource. This is not a critical problem if human lexicographers may carry out this cleaning process while editing the dictionary.

Dictionary lemmata that do not appear in the corpora we may select for a tagging with labels like *old-fashioned* or *out of use*. A comparative analysis of corpus data from different periods can help choose the proper label. In this sense, initiatives like the *Lexikoaren Behatokia* 'Lexicon Observatory' (Euskaltzaindia 2009) will provide valuable resources.

Finally, a Basque *meta-dictionary* may be one of the applications of the methods presented in this paper. To create such a dictionary, the lemma lists of a large group of Basque dictionaries should be brought together, as done for the present work, and the unified lexical database could be published with a user interface, so that the occurrences of a lemma in the different dictionaries could be seen.

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 (IT665-13 and Ber2tek (Etortek-IE12-333). Funding is gratefully acknowledged. We also want to express our acknowledgement to the following institutions, which have provided lexical data to us: To the Institute of the Basque Language (EHU Euskara Institutua) for providing access to ETC and Sar82 frequency lists and the OEH lemma list; To the IXA group (EHU) for giving access to EDBL; and to the Elhuyar Fundation, for providing Elh200 and Elh124 frequency data and the ElhDic lemma list.

References

Aduriz, I., Aldezabal, I., Alegria, I., Artola, X., Ezeiza, N. & Urizar, R. (1996). EUSLEM: A lemmatiser/tagger for Basque. *Proceedings of EURALEX 1996*. Göteborg, Sweden, 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. *Proceedings of the IRCS Workshop on linguistic databases.* Philadelphia, USA

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*, 209–226.

Elhuyar Hizkuntza Zerbitzuak. (2013). *Elhuyar hiztegia: euskara-gaztelania, castellano-vasco*. 4. ed. Usurbil: Elhuyar. http://hiztegiak.elhuyar.org/

Etxebarria, J.M. & Mujika, J.A. (1987). *Euskararen oinarrizko hiztegia : maiztasun eta prestasun azterketa*. Gasteiz: Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia.

Euskaltzaindia. (2008). *Hiztegi batua*. Donostia: Elkar. http://www.euskaltzaindia.net/hiztegibatua

Euskaltzaindia. (2009). Lexikoaren Behatokia. Bilbo: Euskaltzaindia. http://lexikoarenbehatokia.euskaltzaindia.net/

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

Hualde, J.I. & Zuazo, K. (2007). The standardization of the Basque language. *Language Problems and Language Planning*, *31*, 142–168.

IDS. (2009). Korpusbasierte Wortgrundformenliste DEREWO, v-40000g-2009-12-31-0.1, mit Benutzerdokumentation. http://www.ids-mannheim.de/kl/projekte/methoden/derewo.html

Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, *10*, 135–155.

Kilgarriff, A. (2001). Comparing Corpora. In *International Journal of Corpus Linguistics*, *6*, 97–133.

Kupietz, M., Belica, C., Keibel, H. & Witt, A. (2010). The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010)*. Valetta, Malta, 1848–1854.

Leturia, I. (2012). Evaluating Different Methods for Automatically Collecting Large General Corpora for Basque from the Web. *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*. Mumbai, India, 1553–1570.

Leturia, I. (2014). *The Web as a Corpus of Basque* (PhD Thesis). Donostia: UPV/EHU

Lindemann, D. (2013). Bilingual Lexicography and Corpus Methods. The Example of German-Basque as Language Pair. *Procedia - Social and Behavioral Sciences*, *95*, 249–257.

Lindemann, D. (2014). Creating a German-Basque Electronic Dictionary for German Learners. *Lexikos*, 24.

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

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

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

Sarasola, I. (1982). *Gaurko euskara idatziaren maiztasun-hiztegia: 1977ko corpus batean oinarritua*. Donostia: Gipuzkoako Aurrezki Kutxa Probintziala.

Sarasola, I., Salaburu, P., Landa, J. & Ugarteburu, I. (2008). *Lexiko atzo eta gaur*. Bilbo: UPV/EHU. http://www.ehu.es/lag/

De Schryver, G.-M., Joffe, D., Joffe, P. & Hillewaert, S. (2010). Do dictionary users really look up frequent words?—on the overestimation of the value of corpus-based lexicography. *Lexikos*, 16.

Sinclair, J. (2005). Corpus and text-basic principles. In Wynne, M. (Ed.) Developing linguistic corpora: A guide to good practice. Oxford: Oxbow Books, 1–16.

Urkia, M. (2002). XX. mendeko euskararen corpus estatistikoa. Hizkuntza-corpusak. Oraina eta geroa, Donostia: UZEI.

UZEI. (2004). *Maiztasun Hiztegia*. Donostia: UZEI.

Wolfer, S., Koplenig, A., Meyer, P. & Müller-Spitzer, C. (2014). Dictionary Users do Look up Frequent and Socially Relevant Words. Two Log File Analyses. In *Proceedings of the XVI Euralex International Congress*. Bolzano/Bozen, Italy, 281–290.