ISSN 2007-9737

SentiTegi: Semi-manually Created Semantic Oriented Basque Lexicon for Sentiment Analysis

Jon Alkorta, Koldo Gojenola, Mikel Iruskieta

IXA NLP Group, University of the Basque Country (UPV/EHU), Vizcaya, Spain

jon.alkorta@ehu.eus, koldo.gojenola@ehu.eus, mikel.iruskieta@ehu.eus

Abstract. The creation of a semantic oriented lexicon of positive and negative words is often the first step to analyze the sentiment of a corpus. Various methods can be employed to create a lexicon: supervised and unsupervised. Until now, methods employed to create Basque polarity lexicons were unsupervised. The aim of this paper is to present the construction and evaluation of the first semantic oriented supervised Basque lexicon ranging from $+5$ to -5. Due to the lack of resources, the Basque lexicon was created translating the SO-CAL Spanish dictionary by means of two bilingual dictionaries following specific criteria and then slightly corrected with the SO-CAL English dictionary and frequency data obtained from the Basque Opinion Corpus. Evaluation results show that the correlation between human annotators is slightly better than between a gold standard lexicon (obtained from human annotation) and the translated dictionary. This shows that the quality of the translated lexicon is satisfactory, although there is a space to improve it.

Keywords. Semantic oriented lexicon, manual translation method, Basque, sentiment analysis.

1 Introduction

Sentiment analysis is a task that classifies documents according to their polarity. This research area has had a big development in the last years due to social networks and Internet, which have increased the quantity of opinions and other types of text with emotion, and is in demand of methods for automatic processing.

There are many resources for sentiment analysis for the most used languages such as English [9], Chinese [15] and Spanish [5].

Additionally, competitions like SemEval [10] have greatly contributed to the development of resources and tools for sentiment analysis. However, the development is not symmetric on lesser used languages or languages in normalization process like Basque.

The semantic oriented lexicons are related to the lexical level and, so, they are useful and important in sentiment analysis. If the semantic orientation of the words is known, opportunities open up to calculate the semantic orientation of sentences and, therefore, the semantic orientation of texts taking into account syntax and discourse constraints.

The creation of the semantic oriented Basque lexicon has been semi-manual translating from the SO-CAL Spanish dictionary, and then enriching it with corpus analysis and the English SO-CAL dictionary. In the translation process, different bilingual dictionaries have been used. We have decided to use a semi-manual procedure to create our lexicon, in order to take into account some idiosyncratic characteristics of Basque language.

The aim of this paper is to present a semantic oriented lexicon for Basque. We will emphasize the process of creating this lexicon, and particularly the solutions adopted to solve the problems encountered.

The main contributions of this work are: $i)$ the creation of a domain-specific semantic oriented Basque lexicon, $ii)$ a description of a semi-manual technique to create the lexicon and $iii)$ a thorough evaluation.

doi: 10.13053/CyS-22-4-3075

This paper has been organized as follows: after presenting related work in Section 2, Section 3 describes the methodology of the translation process. Then, Section 4 discusses the design decisions, while Section 5 describes the characteristics of the created lexicon in two stages. In Section 6 the quality of the lexicon is evaluated and, finally, Section 7 concludes the paper, also proposing directions for future work.

2 Related Work

There are various approaches for the creation of polarity lexicons, based on knowledge or on automatic methods. Each of the approaches has its advantages and drawbacks.

SO-CAL [14] is a dictionary-based tool to extract sentiment from texts. The dictionary was created manually, where words are annotated with polarity (positive or negative) and strength (semantic orientation: from ±1 to ±5). There are two versions of SO-CAL tool. The original version is the English SO-CAL and the Spanish version, the second one, is based on the previous version. The English and Spanish dictionaries (V1.11) contain 6,610 and 4,880 words, respectively.

A disadvantage of manually-created lexicons is the hard-work to make modifications. In contrast, they can be tailored to be domain-specific and, depending on the linguistic information used, they can treat a variety of different linguistic phenomena.

ML-SentiCon [6] is a multilingual polarity lexicon, where the lexicons have been automatically generated from an improved version of Senti-WordNet. It contains a Basque lexicon that contains 4,323 lemmas. The polarity values are situated between -1 and $+1$, in a continuous scale. Additionally, QWN-PPV tool [11] is able to generate multilingual polarity lexicons, including Basque. This unsupervised tool makes use of a corpus and WordNet.

The main disadvantage of these lexicons is that they are not domain-specific, so their results could vary from one domain to another. In contrast, their main advantage lies on the facility to create them.

Another characteristic of previous three works is that the sentiment value of words is in a scale, although the scale dimensions are different. However, there are works in which the sentiment value of words are not in scale. For example, in some works like [13], there are two non-numerical tags: *positive* and *negative*. Consequently, two words with different intensity are expressed with the same tag.

Methods to evaluate lexicons are different depending on each technique. Some works [3] use intrinsic methods where the result of the system is compared to a gold standard data set, predefined by evaluators. In contrast, there are other systems [4] which use extrinsic methods where the system is evaluated in an applied setting. Finally, some works [7] use both extrinsic and intrinsic methods.

The lexicon presented in this work differs from previous ones in several respects. SO-CAL dictionaries have also been manually created but, until now, they have dealt with languages which are not morphologically rich (Spanish and English) in contrast with Basque. Another relevant difference of this study has been the evaluation. We will apply an intrinsic evaluation and measure, using Pearson correlation, the agreement between two human annotators, and the reliability between the gold standard (based on human annotation) and the translated dictionary. Finally, the characteristic of the created lexicon is another interesting aspect. The words of the lexicon have the sentiment value in a scale from -5 to $+5$. This allows us to study how sentiment shifters of different linguistic levels (morphology, syntax and discourse) affect on sentiment analysis.

3 Methodology

In order to create a semantic oriented lexicon for Basque, we have adopted several decisions taking different factors into account:

i) **Time.** The creation of a semantic oriented lexicon for Basque is related to the project of linguistics-based Basque sentiment analysis and, for that reason, the time to create the lexicon is limited.

ii) **Resources.** The Basque language is still in a normalization process and this has some limitations to create corpora and to reuse computational resources. On the one hand, it is difficult to create a large opinion corpus of different topics. This situation could affect to the quality of the lexicon if the corpus is used for that. The collaboration of lexicographers would be ideal but it is a costly resource, not available. This situation adds a difficulty to create a semantic oriented Basque lexicon from zero.

iii) **Quality.** We want to develop the lexicon with the best possible quality (and in the less time possible) and with that aim we will first translate the lexicon, after that evaluate it and then improve our semantic oriented lexicon following an specific criteria.

3.1 Resources for Translation

We have used mainly four resources in the translation process.

i) **The SO-CAL Spanish Dictionary [14].** This dictionary is the source to create the Basque semantic oriented lexicon. It contains 4,880 words of five grammatical categories (noun, adjective, adverb, verb and intensifier).

ii) **Two Bilingual Dictionaries:** Spanish-Basque: Elhuyar dictionary [16] and Zehazki [12]. These dictionaries have been used to translate the Spanish SO-CAL dictionary. Moreover, they have also been used to check if the translated word is an entry of such dictionaries since we will work only with words which are entries of one of these dictionaries. Dealing with collocations and expressions is necessary but it is out of the scope of this work.

iii) **The Basque Opinion Corpus [1].** After getting the first version of the lexicon, each entry has been checked in the corpus to create a domain-based lexicon. The corpus contains 240 texts of six different domains.

iv) **The SO-CAL English dictionary [14].** This version which contains 6,610 words has been used to verify and enrich the already created domain-based lexicon.

Taking all the factors explained above into account and using the mentioned resources, we have decided to translate the SO-CAL Spanish dictionary to create the Basque SO-lexicon *Sentitegi*, following the methodology explained in Figure 1.

3.2 Translation Steps

Figure 1 shows the steps followed in the translation process. To begin with, a first version of a semantic oriented Basque lexicon has been created from the Spanish version of the SO-CAL dictionary. After that, the second version has been created enriching it with the English lexicon version (V1.11) and limiting it to the domains of Basque Opinion Corpus.

Some interesting phenomena have been detected in the translation process of SO-CAL dictionaries from Spanish and English versions (V1.11) to Basque. Table 1 shows these five phenomena.

- Phenomenon 1 (P1): the Spanish word is translated but the translation is not an entry of Elhuyar [16] and Zehazki [12] dictionaries, so we do not take it into account.
- Phenomenon 2 (P2): The Spanish word is translated, it is an entry of Elhuyar but the translation does not appear in the Basque Opinion Corpus. Consequently, it will appear in the first version (V1.0) but not in the second one (V2.O).
- Phenomenon 3 (P3): The Spanish word is translated, it is an entry, it appears in the corpus but it is not in the SO-CAL English dictionary. So, it will appear in the first version of the dictionary, but not in the second one.
- Phenomenon 4 (P4): The Spanish word is translated, it is an entry, it appears in the corpus and it is not present in the SO-CAL English dictionary. Then, it will be included in the (first and) second version.

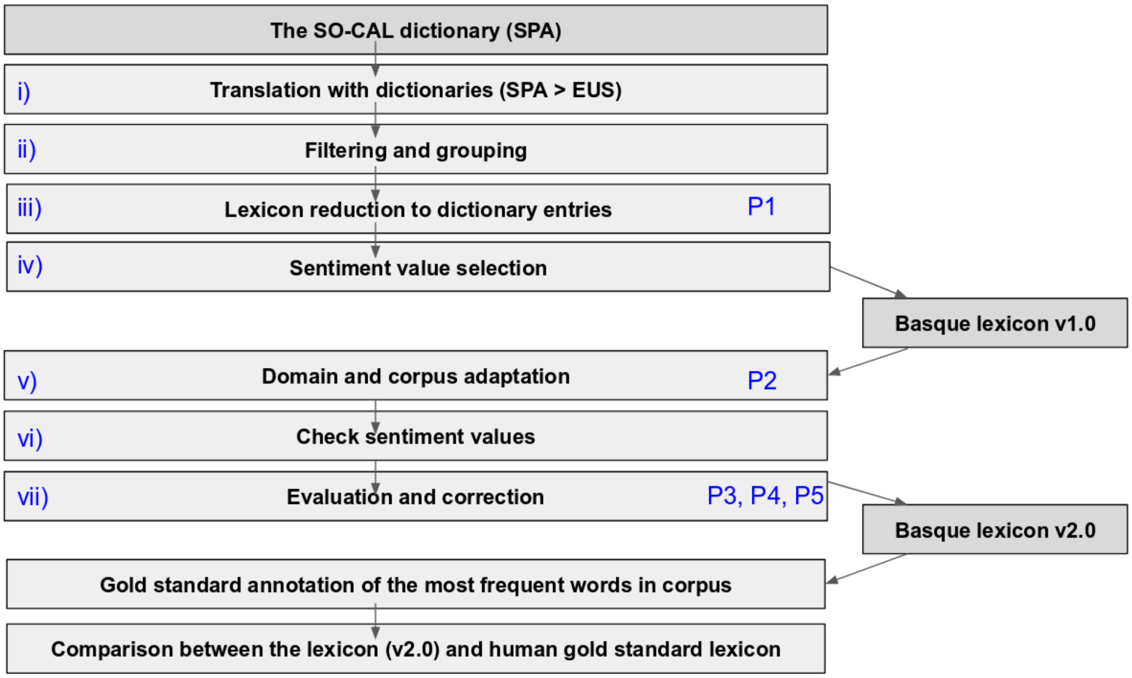
The SO-CAL dictionary (SPA)
i) Translation with dictionaries (SPA > EUS)
ii) Filtering and grouping
iii) Lexicon reduction to dictionary entries
P1
iv) Sentiment value selection
Basque lexicon v1.0
v) Domain and corpus adaptation
P2
vi) Check sentiment values
vii) Evaluation and correction
P3, P4, P5
Basque lexicon v2.0
Gold standard annotation of the most frequent words in corpus
Comparison between the lexicon (v2.0) and human gold standard lexicon

Fig. 1. Steps of the translation process. The enumeration in blue on the left indicates methodological steps. The blue code on the right (P1 to P5) indicates different phenomena in the translation process

– Phenomenon 5 (P5): The Spanish word is translated, it is an entry, it appears in the corpus and it also a word of the SO-CAL English dictionary. It will appear in the first and second versions. These last two phenomena are the same but the decision is different that depends on the characteristic of each word.

The translation process has been the following (see Figure 1):

i) **Automatic translation from Spanish into Basque.** The Spanish sentiment dictionary of SO-CAL has been translated using Elhuyar [16] and Zehazki [12] dictionaries. When one word of the dictionary has more than one entry, all the entries have been taken into account. The sentiment value of the Spanish word has been assigned to all the correlated elements in Basque.

For example, the Spanish word *desacreditar* -2 "discredit" has been translated into Basque in different forms: *izena_kendu*, *ospea_kendu* and *sona_kendu* "discredit" with the same meaning. This example shows how one Spanish word could be translated in different forms to Basque. But these translations are not entries of the dictionary. Consequently, they have not been taken into account.

ii) **Filtering and grouping.** After translating all the words and transferring their sentiment values, the repeated words in Basque have been filtered and grouped.

Table 1 shows how words in Basque (fourth column) can have one or more translations in Spanish (third column). The phenomena numbered 1, 2 and 4 have one translated word in Spanish whereas 3 and 5 have more than one.

This phenomenon occurred because those words are polysemic. There are cases where two or more words in Spanish correspond to the same word in Basque and vice versa. Consequently, in some cases, each word in

Table 1. Words that belongs to five phenomena related to translation process

Phenomenon	SPA	SPA grouping	EUS	ENG	Value
P1	desacreditar “discredit”	desacreditar -2 “discredit”	ospea_kendu -2 “discredit”	-	-
P2	atrofiar “atrophy”	atrofiar -1 “atrophy”	atrofiatu -1 “atrophy”	-	-
P3	amago “feint”	amago “feint” -1 cicatriz “scar” -2	seinale “signal” -1	-	-
P4	franquismo “Francoism”	franquismo -2 “francoism”	frankismo -2 “francoism”	-	-2
P5	correcto “correct”	acertado “correct” +3 correcto “correct” +3 decente “decent” -2	zuzen +3 “correct”	right +1 correct +3	+3

Basque has several meanings and sentiment values in Spanish.

iii) **Dictionary entry: Check if the Basque translation is an entry in the Elhuyar [16] and Zehazki [12] dictionaries.** We have only accepted the translations which are entries of Elhuyar and Zehazki dictionaries. Consequently, Phenomenon 1 in Table 1 has occurred: *ospea_kendu* “discredit” is a collocation and not an entry, so we will not take it into account. In contrast, other words in the table are entries in the dictionary and they are maintained.

iv) **Sentiment value selection.** The value (and meaning in Spanish) of each word in Basque will be selected.

In order to choose the value, we have followed the following criteria:

- If the word in Basque has one translation (and value) in Spanish and if that translation is correct, the translation is selected. This is the case of phenomena 2 and 4 in Table 1. Sometimes the translation is not “correct” or “direct” as we will observe in Section 4.
- If the word in Basque has many translations (and values) in Spanish, the translation has been selected according to which translation is the best to use in the Basque Opinion Corpus [1]. We have analyzed the context of the words in the corpus using Key Word In Context (KWIC) format for concordance. This is the case of Phenomena 3 and 5 in Table 1.
- In the creation of the first version of the lexicon, there have also been cases where the word in Basque has not instances in the corpus. In these cases, the meanings that are used more frequently have been selected.

After these four steps, the first version of the Basque lexicon (V1.0) has been created. However, we detected some inconsistencies and we have felt the necessity to feed more information and, for that reason, we followed new steps to create the Basque lexicon (V2.0):

v) **Domain and corpus adaptation: New lexicon based on the Basque Opinion Corpus [1].** We have curated the first lexicon (Basque V1.0) and created the second version of this lexicon (Basque V2.0). This new lexicon has been curated with the information obtained form word frequencies we have extracted from the Basque Opinion Corpus.

The effects of this step are showed in Phenomenon 2 in Table 1. The word *atrofiatu* “to atrophy” does not appear in the corpus, so it is not related to the domains of the corpus and, consequently, we do not take it into account. We do not take into account them because our work is limited to our corpus and we want to maintain as much as possible the coherence of SO values and avoid complexities which we do not see useful. In Table 1, Phenomena 3, 4 and 5 are not affected by this limitation while Phenomenon 2 is. With this procedure, the number of entries in the lexicon was reduced from 8,140 to 1,813 words, because it was manually checked and reviewed.

vi) **Curate and check SO values of each entry**: Find the English translations of each Basque entry in the SO-CAL English dictionary. Using the Elhuyar dictionary [16], we have translated the words in Basque to English and, after that, we have checked if the translated words are in the SO-CAL English dictionary. If the word is in this dictionary, we have maintain the dictionary entry and its value in the second version of the Basque dictionary. If the word is not in the English dictionary, almost in all cases. it was excluded from the second version in the Basque dictionary.

In Table 1, Phenomena 3 and 4 do not have any translation in the English dictionary and, consequently, their (English) column in Table 1 is empty. In contrast, Phenomenon 5 has two translations according to the English dictionary: *right* and *correct*.

vii) **Evaluation and correction**: Compare and choose the best translation and value. In this step, each word in Basque has the same value, most of the times, in Spanish and English (Basque V1.0).

There are 3 different cases in this situation:

- Phenomenon 3. There is not a word in the English version corresponding to the Basque word and the previous Spanish one is not accepted. In phenomenon 3, the word *seinale* “sign” has been assigned the value -1 (Table 1, fourth column) but there is not a corresponding value in the English version and, consequently, we have removed that value.
- Phenomenon 4. There is not a corresponding word in the English version for Basque and the previous Spanish translation and value are accepted. The word *frankismo* “francoism” is related to Spain and, for that reason, it appears in the Spanish version and not in English. In this case, we have maintained the assigned value.
- Phenomenon 5. The English translation and value are the same or better quality than the Spanish ones. Phenomenon 5 shows that the Spanish and English values agree, so we have assigned the value $+3$ to *zuzen* “correct”. In other cases, the English and Spanish values differ. When this happens we decided that the English value will prevail to the Spanish one in the second version of the Basque dictionary, because the quality is slightly better in English as we previously report.

Phenomena 3 and 5 show how we have decided to give more relevance to the English version.[1]

4 Discussion

We explain in this section how we have solve the most fundamental problems we have found during the translation process:

i) **Source language is not always the preferred language.** English and Spanish could be the source language but we have chosen Spanish due to several reasons. The overall accuracy of the English SO-CAL is 76.62% while in the Spanish version is 71.81% [2]. In other words, the difference between them is not big enough. On the other hand, there are many more resources to translate

[1] Sometimes there is not a corresponding word in the English dictionary [16], an example and the explanation of what we have done in such cases is explained in Section 4.

Table 2. Examples of translations applying the coherence criteria

Criteria	EUS	Value	EUS	Value
A	errukigabe ”ruthless”	-4	errukigabeko ”(with) ruthless”	-4
B	tonto ”stupid”	-3	tuntun ”stupid”	-3
C	arduradun ”responsible”	$+2$	arduragabe ”irresponsible”	-2

the dictionary from Spanish to Basque than to translate from English to Basque. So, the translation from Spanish is more reliable and extended as shown in Table 1, where the phenomenon numbered 4 (*frankismo* “francoism”) shows that although the English dictionary contains more items, there are some words in the Spanish dictionary that are not present in the English one.

In contrast, the English version has helped to check if the assigned value to the Basque word in the first version from Spanish is correct. In the cases where the value of the Spanish and English versions are different, we have preferred the English one as Phenomenon 3 (*seinale* “signal”) shows. Due to this decision, the number of words of the lexicon has decreased from 1,813 to 1,237 entries.

ii) **Not one to one translation**. Another problem was presented when, in the translation, a Spanish word could be translated into Basque in different forms but with the same sense. We have decided to use all the translated words in Basque so as to get the higher recall possible. The first step, the automatic translation from Spanish into Basque, shows that one or more entries have been taken in Basque.

For example, the Spanish word *aparatoso* “showy, spectacular” has been translated into Basque in two different ways: *arranditsu* “spectacular” and *deigarri* “showy”.

iii) **Domain adoptation of polysemic words**. There are some words that have opposite meanings according to their context. The best solution would be to create two entries but then it would be difficult to implement it in a system that does not distinguish between word senses. In this situation, we have decided to take only one meaning and we have used the Basque Opinion Corpus [1] to choose the meaning with the appropiate SO value.

For example, the Basque word *deigarri* “showy, spectacular” comes from Spanish *aparatoso* -3 “spectacular” or *llamativo* $+3$ “showy”. Taking the context of the word in the corpus into account, we have disambiguated the word manually and chosen the value $+3$ for this word.

iv) **Coherence consistency.** In the process of choosing the value, we have to try (when the values match) to maintain the coherence of the values taking these criteria into account. Examples of the criteria are shown in Table 2.

A) Sometimes, the same word appears in different forms. For example, in the creation of the first version of the lexicon, it is usual that one word appears sometimes with genitive *-ko* “with” and other times with an elided genitive, and in both cases is a dictionary entry. In these cases, we decided to assign the same value. One of the cases is the adjective *berehala* “immediate”. It appears with genitive suffix: *berehalako* “immediately” and without it *berehala* “immediate”. We have assigned the same sentiment value $(+2)$ to both.

B) We assign (when the values match) the same value to words with similar meanings. For example, *tonto* “stupid” is used with man while *tuntun* with the same meaning is used with woman. We assign the value -3 to both.

Table 3. The semantic oriented Basque lexicons (V1.0 and V2.0)

	V1.0		V2.0	
Grammatical category	**Words**	**%**	**Words**	**%**
Noun	2,282	28.06	461	37.27
Adjectives	3,162	38.85	446	36.05
Adverbs	652	7.98	54	4.36
Verbs	1,657	20.36	276	22.32
Intensifiers	387	4.75		
Total	**8,140**	100	**1,237**	100

C) We also assign the same intensity (1 to 5), but opposite value (positive/negative) to antonymic words when the values coincide in Basque dictionary entry. In Basque, some prefixes (*des-* and *ez-* "dis-") and suffixes (*-ezin* "impossibility" "inability" and *-gabe* "without") are used to invert the meaning of the words and we have put special attention on these ones.

v) **"Incorrect" translations.** There have been some translations which are incorrect because of different factors. The Spanish word *provinciano* "backward" (-1) is employed to refer to people of Bizkaia and Gipuzkoa provinces. The Elhuyar dictionary [16] has defined the word as "inhabitant of Bizkaia or Gipuzkoa", a translation which is not useful for our purpose.

vi) **"Indirect" translations.** There have been some translations that we have considered as indirect. They are correct translations but since they have an extensive meaning and they are used in limited situations, they are not useful for us.

For example, the word *beltz* "black" could have two meanings: $i)$ a color $ii)$ "black, sad; gloomy, depressing" (figurative meaning). The figurative use of that word is less usual, there are other words with the same meaning and, taking into account that the word could complicate the correct sentiment value assignation of texts, we have decided not to assign any SO value.

The explained problems show the difficulty to translate a semantic oriented lexicon semi-automatically. This translation process is large and very detailed where the translation of the lexicon has different phenomena.

5 Results

As a result of the translation process, two versions of the semantic oriented Basque lexicon have been created. Table 3 shows the characteristics of these two versions.

The first version (V1.0) is the result of the first four steps in the translation process (Figure 1). It is translated directly from the Spanish SO-CAL dictionary with a strict criteria. But, unlike the second version (V2.0), the first version is not subject to the restrictions of being an entry of the Basque bilingual dictionaries and it was not improved taking into account the English SO-CAL dictionary, the Basque Opinion Corpus and other kind of features that work differently such intensifiers are considered as dictionary entries.

As a result of these considerations, the first versions have 8,140 entries and the second version 1,237, respectively. In both cases, nouns and adjectives are the grammatical categories with more entries. Verbs and adverbs are least frequent entries, whereas intensifiers have not been taken into account in the second version because they affect to other words, so we think that it is better to analyze differently assigning different values that does not go from -5 to -5 values.

Table 4. Examples of parallel lexicon

Word in lexicon	Value	SPA	Value	ENG	Value
bikain	$+5$	excepcional	$+5$	excellent	$+5$
on	$+2$	buen	$+2$	-	-
eskas	-1	escaso	-2	insufficient	-1
txar	-3	adverso	-3	bad	-3

Another interesting characteristic of the created lexicon is that it is parallel. That means that each word of the lexicon has it translations in English and Spanish and the sentiment values in each language also are included. This information appears in an orderly manner in the resource.

In Table 4, there are four examples showing the parallel lexicon. Sometimes, four sentiment values do not match because the Spanish and English SO-CAL lexicons have been created in different way. But the Basque word always matches with one of them. The examples of Table 4 are adjectives and they show how the sentiment values are in a scale.

Once we have implemented this lexicon in the Basque SO-CAL preliminary version, the created semantic oriented lexicon is useful to assign sentiment value to words as well as sentences, as is shown in the following examples:

(1) [*Halere, pentsa litekeenaren aurka, gaien urritasunak eta diskurtso errepikakorrak*$_{-6}$ *ez dakarte ñabardura aberastasunik, are gutxiago argumentu-mailako sakontasunik.*]$_{-6}$
(However, contrary to what is thought, the scarcity of problems and the repetitive$_{-6}$ discourses do not imply rich nuances, much less a plot depth.)$_{-6}$

(2) [*Arazo nagusia*$_{+2}$*, nire ustez, gaien*$_{+4}$ *emankortasun zalantzazkoan eta ekintzaren bilakaera eskasean*$_{-3}$ *datza.*]$_{+3}$
(The main$_{+2}$ problem is, I believe, the uncertain fertility of the topics$+4$ and the slow$_{-3}$ evolution of the action.)$_{+3}$

(3) *(...)* *[Emaitza ezustekorik*$_{-1.5}$ *gabeko istorio bat da, irakurlea epel*$_{-1.5}$ *uzteko arrisku dezente duen tonu arras moderatu batean emana.*]$_{-3}$
(The result is an unsurprising$_{-1.5}$ story, given in a moderate tone with a risk to leave the reader cold$_{-1.5}$.)$_{-3}$

As we show in the three examples the words of the dictionary have a SO value at the end of the word. To mention one, in Example 1, the Basque version of SO-CAL tool assigns the value -6 to the word *errepikakor* “repetitive”. There is no another word with sentiment value according to lexicon, so the sentiment value of the sentence is also -6.[2] The methodology to calculate the semantic orientation of the sentence is similar in Examples 2 and 3.

6 Evaluation

In this section, we want to evaluate two aspects of the translation task. On the one hand, we want to evaluate the difficulty of the task. We think that the annotation of sentiment polarity is a difficult task because there is not a guide to follow and subjective perceptions must be, first, measured and, last, corrected if possible. On the one hand, the inter-annotator agreement of SO value annotation has been evaluated between two linguists annotators. On the other hand, we also want to measure the quality of the translated lexicon. With these in mind, a gold standard annotation has been created from the previous annotation and discussion by both annotators.

[2] In this sentence, the sentiment value of the word *errepikakor* “repetitive” in the lexicon is -4. But in SO-CAL tool, there are some mathematical operations related to linguistic phenomena that increase or decrease the sentiment value of the words. In this case, the sentiment value has increased to -6.

Table 5. Pearson correlation measurement and contingency table between two annotators

Grammatical category	Pearson 1	Pearson 2
Noun	0.87	0.59
Adjectives	0.71	0.60
Adverbs	0.93	0.82
Verbs	0.87	0.76
Total	**0.79**	**0.73**

Total categories			
	0	**NEG**	**POS**
0	187	12	27
NEG	14	42	5
POS	39	5	69

In order to evaluate these two aspects, we have extracted the most frequent 400 words (100 per each grammatical category) using Analhitza [8] from the Basque Opinion Corpus [1]. We have used Pearson correlation [17] to evaluate both tasks. Pearson correlation has been used in two different ways: $i)$ Pearson 1: the correlation is measured taking into account only the annotated words by both annotators and $ii)$ Pearson 2: the correlation is measured taking into account all words in the corpus.[3]

6.1 Correlation between annotators

We have decided to measure the correlation of two annotators to create the gold standard, taking into account the results achieved in the correlation coefficient. Table 5 shows the coefficient for each grammatical category, together with a contingency table.

Pearson 1 value shows that the correlation coefficient is high (0.79). This means that the value assigned is similar in a big percentage of the annotated words. The coefficients for different grammatical categories are situated between 0.71 and 0.93. In a similar way, Pearson 2 also shows high correlation, although it is slightly lower (0.73), with values between 0.59 and 0.82.

The contingency table of Table 5 shows that the biggest difference comes when one annotator has assigned a value to one word and the other one had not assigned any value and vice versa (90.19 % of all discrepancies 92 of 102).

After calculating this correlation, two annotators have discussed about their differences and after reaching consensus, a gold standard has been created.

6.2 Correlation between the lexicon and gold standard

The correlation between the human gold standard lexicon and the translated lexicon shows some differences compared to the correlation between two annotators as presented in Table 6.

With Pearson 1, the cases in which the dictionary and gold standard contain an annotation for the word show similar correlation when compared to the results of two annotators (0.79). The correlation is high since the coefficients for the different grammatical categories are situated between 0.69 and 0.96. In contrast, Pearson 2 shows a lower correlation (0.54) and the coefficients of grammatical categories are situated between 0.47 and 0.59.

The interpretation of these results is that the values assigned to the dictionary and gold standard are similar (Pearson 1). But the difference from the previous result in Pearson 2 is created when the semantic oriented lexicon assigns value to the word and the annotator does no do it. This situation does not occur in the correlation between two annotators.

The contingency table shows us how the gold standard and the created dictionary differ. The discrepancy here also comes from the difficulty to assign a positive or negative value to a word. The difference is similar: 89.83 % of all discrepancies (106 of 118) are related to the decision to assign sentiment polarity to words. But here, in contrast with correlation between two annotators, the last version of the lexicon is more conservative, because the gold standard

[3] This means that there are cases where one word has been annotated by one annotator or by none of the them. When it happens, the un-annotated words value is 0 in order to calculate the Pearson correlation.

Table 6. Pearson correlation measurement and contingency table between the gold standard and the Basque semantic oriented lexicon (V2.0)

Grammatical category	Pearson 1	Pearson 2
Noun	0.96	0.59
Adjectives	0.78	0.56
Adverbs	0.75	0.47
Verbs	0.69	0.54
Total	**0.76**	**0.54**

	Total categories		
	0	**NEG**	**POS**
0	195	2	15
NEG	30	34	8
POS	59	4	53

annotates much more words than the lexicon does, decreasing the correlation in Pearson 2.

To sum up, the evaluation shows a high correlation in Pearson 1 in the case of two annotators and the lexicon and gold standard. The correlation coefficient is 0.79 and 0.76, respectively. In the case of Pearson 2, the correlation between two annotators remains high (0.73) but the correlation measure falls between the lexicon and gold standard (0.54).

7 Conclusion and Avenues for Future Work

In this paper we presented the first semi-manually created semantic orientation lexicon for Basque[4]. Time factor, few resources and quality pushed us to translate the SO-CAL Spanish dictionary to Basque.

The translation process has followed several steps. To summarize the steps, the English and Spanish SO-CAL dictionaries have been translated into Basque using two bilingual dictionaries. After that, the groups of words with the same meaning have been grouped and the best sentiment values according to the context of the Basque Opinion Corpus have been chosen. Finally, the created lexicon has been adapted to the domains of the Basque Opinion Corpus. The Basque sentiment lexicon has its limitations, since polysemy and figurative meaning phenomena were not considered and therefore are not totally solved.

Pearson correlation shows that the agreement coefficient is high between both annotators with respect to the following two factors: $i)$ assigning a value and $ii)$ deciding if a word has any value. In contrast, in the case of the comparison between human gold standard and translated lexicon, the correlation coefficient is high when the value is assigned but not in the case of deciding if the word has a value or not, which results has been lower. This lower coefficient appears mainly because there are less words annotated in our translated lexicon V2.0.

At present, the second version of semantic oriented lexicon is implemented in the Basque SO-CAL. In a foreseeable future, our aim is to improve this lexicon but considering morphosyntactic and discourse phenomena. This lexicon will be the basis of this system and we will consider how to enrich the system with sentence level and text level information.

Acknowledgements

Jon Alkorta's work is financed by a Basque Government scholarship (code: PRE_2017_2_0041). Koldo Gojenola's work is partially financed by the PROSA-MED (TIN2016-77820-C3-1-R) funded by the Spanish Ministerio de Economia, Comercio y Competitividad) and UPV/EHU IXA Group (University of the Basque Country). Mikel Iruskieta's work is partially financed by the TUNER project (TIN2015-65308-C5-1-R) funded by the Spanish Ministerio de Economia, Comercio y Competitividad) and UPV/EHU IXA Group (University of the Basque Country). We would like to offer our special thanks to Maite Taboada, Arantxa Otegi and Oier Lopez de Lacalle.

[4] The semantic oriented Basque lexicon is available at: `http://ixa.si.ehu.es/node/11438`

References

1. **Alkorta, J., Gojenola, K., & Iruskieta, M. (2016).** Creating and evaluating a polarity - balanced corpus for basque sentiment analysis. *Proceedings of Fourth International Workshop on Discourse Analysis (IWoDA16)*, pp. 58–62.

2. **Brooke, J., Tofiloski, M., & Taboada, M. (2009).** Cross-linguistic sentiment analysis: From english to spanish. *Proceedings of the international conference RANLP-2009*, pp. 50–54.

3. **Chetviorkin, I. & Loukachevitch, N. (2012).** Extraction of russian sentiment lexicon for product meta-domain. *Proceedings of COLING 2012*, pp. 593–610.

4. **Chetviorkin, I. & Loukachevitch, N. (2014).** Two-step model for sentiment lexicon extraction from twitter streams. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 67–72.

5. **Cruz, F. L., Troyano, J. A., Enriquez, F., & Ortega, J. (2008).** Clasificación de documentos basada en la opinión: experimentos con un corpus de crıticas de cine en espanol. *Procesamiento del lenguaje natural*, Vol. 41, No. 0.

6. **Cruz, F. L., Troyano, J. A., Pontes, B., & Ortega, F. J. (2014).** Ml-senticon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas. *Procesamiento del Lenguaje Natural*, Vol. 53, pp. 113–120.

7. **Goyal, A. & Daumé III, H. (2011).** Generating semantic orientation lexicon using large data and thesaurus. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Association for Computational Linguistics, pp. 37–43.

8. **Otegi, A., Imaz, O., de Ilarraza, A. D., Iruskieta, M., & Uria, L. (2017).** Analhitza: a tool to extract linguistic information from large corpora in humanities research. *Procesamiento del Lenguaje Natural*, , No. 58, pp. 77–84.

9. **Pak, A. & Paroubek, P. (2010).** Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, volume 10, pp. 1320–1326.

10. **Rosenthal, S., Farra, N., & Nakov, P. (2017).** Semeval-2017 task 4: Sentiment analysis in twitter. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 502–518.

11. **San Vicente, I., Agerri, R., & Rigau, G. (2014).** Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 88–97.

12. **Sarasola, I. (2005).** *Zehazki: gaztelania-euskara hiztegia*. Alberdania.

13. **Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966).** The general inquirer: A computer approach to content analysis.

14. **Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011).** Lexicon-based methods for sentiment analysis. *Computational linguistics*, Vol. 37, No. 2, pp. 267–307.

15. **Tan, S. & Zhang, J. (2008).** An empirical study of sentiment analysis for chinese documents. *Expert Systems with applications*, Vol. 34, No. 4, pp. 2622–2629.

16. **Zerbitzuak, E. H. (2013).** Elhuyar hiztegia: euskara-gaztelania, castellanovasco. usurbil: Elhuyar.

17. **Zou, K. H., Tuncali, K., & Silverman, S. G. (2003).** Correlation and simple linear regression. *Radiology*, Vol. 227, No. 3, pp. 617–628.

Article received on 18/12/2017; accepted on 15/02/2018.
Corresponding author is Jon Alkorta.