CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Biblioteca Digital Biblioteca Digital de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires (Biblioteca Digital FCEN-UBA)


# Spanish DAL: A Spanish Dictionary of Affect in Language

Gravano, Agustín; Dell'Amerlina Ríos, Matías G.

2014-02

Technical Report

Reporte Técnico
Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

http://digital.bl.fcen.uba.ar/gsdl-282/cgi-bin/library.cgi?p=about&c=technicalreport

Contacto: digital@bl.fcen.uba.ar

Este documento forma parte de la colección de Reportes Técnicos del Departamento de Computación de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires. Su utilización debe ser acompañada por la cita bibliográfica con reconocimiento de la fuente.

This document is part of the Technical Reports collection of the Departamento de Computación de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires. It should be used accompanied by the corresponding citation acknowledging the source.

Fuente / source:
Biblioteca Digital de la Facultad de Ciencias Exactas y Naturales - Universidad de Buenos Aires
http://digital.bl.fcen.uba.ar

*Technical Report*

# Spanish DAL: A Spanish Dictionary of Affect in Language

Agustín Gravano Matías G. Dell'Amerlina Ríos

Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

{gravano,mamerlin}@dc.uba.ar

February, 2014

**Abstract**

The topic of sentiment analysis in text has been extensively studied in the English language for the past 30 years. An early, influential work by Cynthia Whissell, the Dictionary of Affect in Language (DAL), allows rating words along three dimensions: pleasantness, activation and imagery. Given the lack of such tools in Spanish, we decided to replicate Whissell's work in that language. This report describes the Spanish DAL, a Spanish lexicon formed by more than 2500 words manually rated by humans along the same three dimensions. We evaluated its usefulness on two sentiment analysis tasks, which showed that our lexicon managed to capture relevant information regarding the three affective dimensions. The Spanish DAL is available for download from `http://habla.dc.uba.ar`.

# Contents

# Chapter 1

# Introduction

In an attempt to quantify emotional meaning in written language, Whissell developed the Dictionary of Affect in Language (DAL), a tool for rating words and texts in English along three dimensions – pleasantness, activation and imagery [WFP+86, Whi89, inter alia]. DAL works by looking up individual words in an emotion lexicon containing 8742 words. All words in this lexicon were originally rated by 200 naive volunteers along the same three dimensions.

Whissell's DAL has subsequently been used in diverse research fields, for example as a keystone for sentiment analysis in written text [YNBN03, e.g.] and emotion recognition in spoken language [CDCT+01]. DAL has also been used to aid the selection of emotionally balanced word stimuli for Neuroscience and Psycholinguistics experiments [GBR02]. Given the widespread impact of DAL for the English language, it would be desirable to create similar lexicons for other languages.

In recent years, there have been efforts to build cross-lingual resources, such as using sentiment analysis tools in English to score Spanish texts after performing machine translation [BTT09] or to automatically derive sentiment lexicons in Spanish [PRBM12]. The purpose of the present work is to create a manually annotated lexicon for the Spanish language, replicating Whissell's DAL, aiming at alleviating the scarcity of resources for the Spanish language, and at determining if the lexicon-based approach would work in Spanish as well as it does in English. We leave for future work the comparison of the different approaches mentioned here. This report describes the three steps performed to accomplish that goal: i) creating an emotion lexicon which is likely to have a good word coverage on arbitrary texts from any topic and genre (Section 2); ii) having a number of volunteers annotate each word for the three affective dimensions under study (Section 3); and iii) evaluating the usefulness of our lexicon on simple tasks (Section 4).

The Spanish DAL emotion lexicon is available for download from the 'Resources' section of the website `http://habla.dc.uba.ar`.

# Chapter 2

# Word selection

The first step in building a Spanish DAL consists in selecting a list of content words that is representative of the Spanish language, in the sense that it will have a good coverage of the words in arbitrary input texts from potentially any topic or genre. To accomplish this we decided to use texts downloaded from *Wikipedia* in Spanish[1] and from an online collection of short stories called *Los Cuentos*.[2] Articles from *Wikipedia* cover a wide range of topics and are generally written in encyclopedia style. We downloaded the complete set of articles in March, 2012, consisting of 834,460 articles in total. Short stories from *Los Cuentos* were written by hundreds of different authors, both popular and amateur, on various genres, including tales, essays and poems. We downloaded the complete collection from *Los Cuentos* in April, 2012, consisting of 216,060 short stories.

## 2.1 Filtering and lemmatizing words

We extracted all words from these texts, sorted them by frequency, and filtered out several word classes that we considered convey no affect by themselves (and thus it would be unnecessary to have them rated by the volunteers). Prepositions, determinants, possessives, interjections, conjunctions, numbers, dates and hours were tagged and removed automatically using the morphological analysis function included in the *Freeling* toolkit [PCR+10].[3] We also excluded the following adverb subclasses for the same reason: place, time, mode, doubt (e.g., *quizás* / *maybe*), negation, affirmation and amount.

Nouns and verbs were lemmatized using *Freeling* as well, except for augmentative and diminutive terminations, which were left intact due to their potential

[1] http://es.wikipedia.org

[2] http://www.loscuentos.net

[3] http://nlp.lsi.upc.edu/freeling/

effect on a word's meaning and/or affect (e.g., *burrito* is either a small donkey, *burro*, or a type of Mexican food). Additionally, proper nouns were excluded. Names of cities, regions, countries and nationalities were marked and removed using *GeoWorldMap*,[4] a freely-available list of location names from around the world. Names of people were also filtered out. Proper names were manually inspected to avoid removing those with a lexical meaning, a common phenomenon in Spanish (e.g., *Victoria*). Other manually removed words include words in foreign languages (mainly in English), roman numbers (e.g., *XIX*) and numbers in textual form, such as *seis* (*six*) and *sexto* (*sixth*). Words with one or two characters were removed automatically, since we noticed that they practically always corresponded to noise in the downloaded texts.

## 2.2 Counting ⟨word, word-class⟩ pairs

We implemented a small refinement over Whissell's work, which consisted in considering ⟨word, word-class⟩ pairs, rather than single words, since in Spanish the same lexical form may have different senses. Thus, to each word (in its lemmatized form) we attached one of four possible word classes – noun, verb, adjective or adverb. For example, *bajo*$_{\text{prep}}$ (*under*) or *bajo*$_{\text{noun}}$ (*bass guitar*).

For each input word $w$, *Freeling*'s morphological analysis returns a sequence of tuples ⟨*lemma*, *POS-tag*, *probability*⟩, which correspond to the possible lemmas and part-of-speech tags for $w$, together with their prior probability. For example, the analysis for the word *bajo* returns four tuples: ⟨*bajo*, SPS00 (i.e, preposition), $0.879$⟩, ⟨*bajo*, AQ0MS0 (adjective), $0.077$⟩, ⟨*bajo*, NCMS000 (noun), $0.040$⟩, and ⟨*bajar*, VMIP1S0 (verb), $0.004$⟩. This means that *bajo*, considered without context, has $87.9\%$ chances of being a noun, or $0.4\%$ of being a verb.

Using this information, we computed the counts of all ⟨word, word-class⟩ pairs, taking into account their prior probabilities. For example, assuming the word *bajo* appeared $1000$ times in the texts, it would contribute with $1000 * 0.879 = 879$ to the frequency of *bajo*$_{\text{prep}}$ (i.e., *bajo* as a preposition), $77$ to *bajo*$_{\text{adj}}$, $40$ to *bajo*$_{\text{noun}}$, and $4$ to *bajar*$_{\text{verb}}$.

## 2.3 Merging *Wikipedia* and *Los Cuentos*

This process yielded 163,071 ⟨word, word-class⟩ pairs from the *Wikipedia* texts, and 30,544 from *Los Cuentos*. To improve readability, hereafter we will refer to ⟨word, word-class⟩ pairs simply as *words*. Figure 2.1 shows the frequency of each word count in our two corpora. We note that both graphics are practically

[4] http://www.geobytes.com/FreeServices.htm

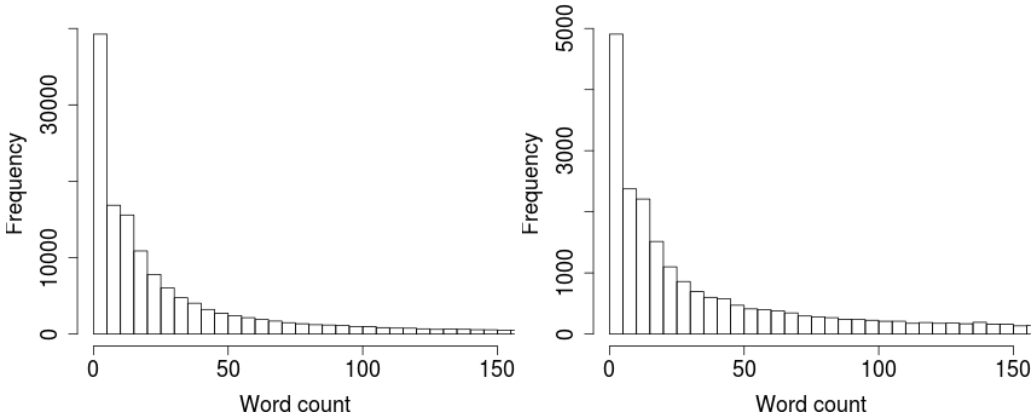


Figure 2.1: Frequency of word counts in texts taken from *Wikipedia* and *Los Cuentos*.

identical, with a majority of low-count words and a long tail with few high-count words.

To create our final word list to be rated by volunteers, we needed to merge our two corpora from *Wikipedia* and *Los Cuentos*. To accomplish this, we normalized all word counts for corpus size (*normalized_count*($w$) = *count*($w$) / *corpus_size*), combined both lists and sorted the resulting list by the normalized word count (for the words that appeared in both lists, we used its average count instead). The resulting list contained 175,413 words in total.

The top 10 words from *Wikipedia* were *más*$_{\text{adv}}$, *año*$_{\text{noun}}$, *ciudad*$_{\text{noun}}$, *población*$_{\text{noun}}$, *estado*$_{\text{noun}}$, *nombre*$_{\text{noun}}$, *vez*$_{\text{noun}}$, *municipio*$_{\text{noun}}$, *grupo*$_{\text{noun}}$ and *historia*$_{\text{noun}}$ (*more, year, city, population, state, name, time*, as in 'first time', *municipality, group* and *history*, respectively). The 10 most common words from *Los Cuentos* were *más*$_{\text{adv}}$, *vez*$_{\text{noun}}$, *vida*$_{\text{noun}}$, *día*$_{\text{noun}}$, *tan*$_{\text{adv}}$, *tiempo*$_{\text{noun}}$, *ojo*$_{\text{noun}}$, *mano*$_{\text{noun}}$, *amor*$_{\text{noun}}$ and *noche*$_{\text{noun}}$ (*more, time, life, day, so, time, eye, hand, love* and *night*).

## 2.4 Assessing word coverage

Next we studied the coverage of the top $k$ words from our list on texts from a third corpus formed by 3603 news stories downloaded from *Wikinews* in Spanish in April, 2012.[5] We chose news stories for this task because we wanted a different genre for studying the evolution of coverage.

Formally, let $L$ be a word list, $T$ any text, and $W(T)$ the set of words occurring at least once in $T$. We define the *coverage* of $L$ on $T$ as the percentage of words in $W(T)$ that appear in $L$. Figure 2.2 shows the evolution of the mean coverage

[5] http://es.wikinews.org

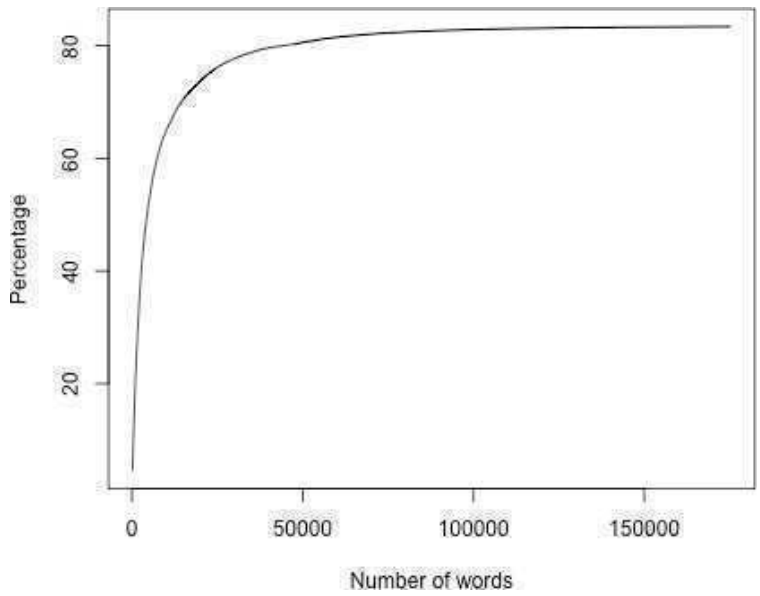


Figure 2.2: Mean coverage of the top $k$ words from our list on *Wikinews* articles.

on *Wikinews* articles of the top $k$ words from our word list. In this figure we can observe that the mean coverage grows rapidly, until it reaches a plateau at around 80%. This suggests that even a low number of words may achieve a relatively high coverage on new texts. The 20% that remains uncovered, independently of the size of the word list, may be explained by the function words and proper names that were removed from our word list. Note that news articles normally contain many proper names, days, places and other words that we intentionally discarded.

# Chapter 3

# Word rating

After selecting the words, the next step consisted in having them rated by a group of volunteers. For this purpose we created a web interface, so that volunteers could complete this task remotely.

## 3.1 Web interface

On the first page of the web interface, volunteers were asked to enter their month and year of birth, their education level and their native language, and was asked to complete a reCAPTCHA[1] to avoid bots. Subsequently, volunteers were taken to a page with instructions for the rating task. They were asked to rate each word along the three dimensions shown in Table 3.1. These are the same three dimen-

| | **Pleasantness** | **Activation** | **Imagery** |
|---|---|---|---|
| 1 | *Desagradable* (Unpleasant) | *Pasivo* (Passive) | *Difícil de imaginar* (Hard to imagine) |
| 2 | *Ni agradable ni desagradable* (In between) | *Ni activo ni pasivo* (In between) | *Ni difícil ni fácil de imaginar* (In between) |
| 3 | *Agradable* (Pleasant) | *Activo* (Active) | *Fácil de imaginar* (Easy to imagine) |

Table 3.1: Possible values for each of the three dimensions.

sions used in Whissell's work. Importantly, these concepts were not defined, to avoid biasing the judgments. Volunteers were also encouraged to follow their first impression, and told that there were no 'correct' answers. Appendix A shows the actual login and instructions pages used in the study.

[1] http://www.recaptcha.net


Figure 3.1: Screenshot of the web page for rating a word.

After reading the instructions, volunteers proceeded to judge two practice words, intended to help them get used to the task and the interface, followed by 20 target words. Words were presented one per page. Figure 3.1 shows a screenshot of the page for rating the word $navegar_{\text{verb}}$. Note that the word class (verb in this example) is indicated right below the word. After completing the first batch of 20 words, volunteers were asked if they wanted to finish the study or do a second batch, and then a third, a fourth, and so on. This way, they were given the chance to do as many words as they felt comfortable with. If a volunteer left before completing a batch, his/her ratings so far were also recorded.

## 3.2 Volunteers

662 volunteers participated in the study, with a mean age of 33.3 (SD $=$ 11.2). As to their level of education, 76% had completed a university degree, 23% had finished only secondary school, and 1% had completed only primary school. Only volunteers whose native language was Spanish were allowed to participate in the study. Each volunteer was assigned 20 words following this procedure: (1) The 175,413 words in the corpus were sorted by word count. (2) Words that had already received 5 or more ratings were excluded. (3) Words that had already been rated by a volunteer with the same month and year of birth were excluded, to prevent the same volunteer from rating twice the same word. (4) The top 20 words were selected.

Each volunteer rated 52.3 words on average (SD $=$ 34.0). Roughly 30% completed 20 words or fewer; 24% completed 21-40 words; 18%, 41-60 words; and the remaining 28%, more than 60 words.

| | Mean | SD | Skewness | Kurtosis |
|---|---|---|---|---|
| Pleasantness | 2.23 | 0.47 | −0.47 | −0.06 |
| Activation | 2.33 | 0.48 | −0.28 | −0.84 |
| Imagery | 2.55 | 0.42 | −0.90 | 0.18 |

Table 3.2: Descriptive statistics for the three dimensions.

## 3.3 Descriptive statistics

A total of 2566 words were rated by at least 5 volunteers. Words with fewer annotations were excluded from the study. Following the conventions from Whissell's work, we assigned each rating a numeric value from 1 to 3, as shown in Table 3.1. In table 3.2 we present a few basic statistics for each of the three dimensions – mean, standard deviation, skewness (a measure of the extent to which a probability distribution 'leans' to one side of the mean) and kurtosis (a measure of the 'peakedness' of the probability distribution; normal distributions have zero kurtosis).

The five **most pleasant** words, according to the volunteers, were *jugar*$_{\text{verb}}$, *beso*$_{\text{noun}}$, *sonrisa*$_{\text{noun}}$, *compañía*$_{\text{noun}}$ and *reir*$_{\text{verb}}$ (*play, kiss, smile, company* and *laugh*, respectively). The **least pleasant** ones were *asesinato*$_{\text{noun}}$, *caro*$_{\text{adj}}$, *ahogar*$_{\text{verb}}$, *herida*$_{\text{noun}}$ and *cigarro*$_{\text{noun}}$ (*murder, expensive, drown, wound* and *cigar*).

Among the **most active** words appear *idea*$_{\text{noun}}$, *publicar*$_{\text{verb}}$, *violento*$_{\text{adj}}$, *sexual*$_{\text{adj}}$ and *talento*$_{\text{noun}}$ (*idea, publish, violent, sexual* and *talent*). Among the **least active**, we found *yacer*$_{\text{verb}}$, *espiritual*$_{\text{adj}}$, *quieto*$_{\text{adj}}$, *esperar*$_{\text{verb}}$ and *cadáver*$_{\text{adj}}$ (*lay, spiritual, still, wait* and *corpse*).

The **easiest to imagine** words include *sucio*$_{\text{adj}}$, *silencio*$_{\text{noun}}$, *dar*$_{\text{verb}}$, *pez*$_{\text{noun}}$ and *pensar*$_{\text{verb}}$ (*dirty, silence, give, fish* and *think*). Finally, the **hardest to imagine** include *consistir*$_{\text{verb}}$, *constar*$_{\text{verb}}$, *morfología*$_{\text{noun}}$, *piedad*$_{\text{noun}}$ and *tendencia*$_{\text{noun}}$ (*consist, consist, morphology, compassion* and *tendency*).

We conducted Pearson's correlation tests between the different dimensions. Table 3.3 shows the correlation matrix. Correlations among rating dimensions were very weak, which supports the assumption that pleasantness, activation and imagery are three independent affective dimensions. These numbers are very similar to the ones reported in Whissell's work.

Next, we computed Cohen's $\kappa$ to measure the degree of agreement above chance between volunteers [Coh68].[2] Given that we used a three-point scale for rating each affective dimension, we used a weighted version of $\kappa$, thus taking into account the distance on that scale between disagreements. For example, the dis-

[2]This measure of agreement above chance is interpreted as follows: 0 = None, 0 - 0.2 = Small, 0.2 - 0.4 = Fair, 0.4 - 0.6 = Moderate, 0.6 - 0.8 = Substantial, 0.8 - 1 = Almost perfect.

| | Pleasantness | Activation | Imagery |
|---|---|---|---|
| Pleasantness | 1.00 | 0.14 | 0.10 |
| Activation | | 1.00 | 0.11 |
| Imagery | | | 1.00 |

Table 3.3: Correlation coefficients between the three dimensions.

tance between *pleasant* and *unpleasant* was 2, and the distance between *pleasant* and *in-between* was 1. We obtained a weighted $\kappa$ measure of 0.425 for pleasantness, 0.305 for activation, and 0.213 for imagery (see Table 3.4). Considering that these were highly subjective rating tasks, the agreement level for pleasantness and activation was quite high. The imagery task seemed somewhat more difficult, although we still observed some agreement above chance. These results indicate that our emotion lexicon managed to, at least partially, capture information regarding the three affective dimensions.

## 3.4 Comparison with expert ratings

As explained above, our word ratings were performed by *naive* volunteers who had diverse backgrounds and showed varying degrees of engagement in our labeling task. This fact poses the question of how these ratings might differ from more traditional labelings obtained via *expert* labelers – i.e. by a fixed number of people who were familiar with the task and instructed to rate words carefully and consistently.

To address this question we conducted a small experiment. We randomly selected 100 words from our lexicon and asked three people (unrelated to this study) to rate them using the same guidelines described in Section 3.1. However, in this case the labelers were asked to rate the words on a spreadsheet, and were allowed to take their time and change their ratings as many times as they wanted, until they felt confident their ratings were consistent. Additionally, words were listed to each labeler in a different random order, to prevent order bias. This way, we ran a simulation on a subset of our data of the conditions typically used for natural language labeling tasks.

First we assessed the reliability of expert ratings, and compared them against naive ratings. Table 3.4 presents the weighted $\kappa$ measure for each dimension on the 100-word subset in both conditions, and on the entire lexicon in the naive condition (as mentioned in the previous section). The expert agreement is comparable to the naive for all three dimensions – it is a bit lower for pleasantness (0.410 vs. 0.471), very similar for activation (0.303 vs. 0.271) and somewhat higher for imagery (0.354 vs. 0.204).

| | Entire lexicon | 100-word subset | |
|---|---|---|---|
| | Naive | Naive | Expert |
| Pleasantness | 0.425 | 0.410 | 0.471 |
| Activation | 0.305 | 0.303 | 0.271 |
| Imagery | 0.213 | 0.354 | 0.204 |

Table 3.4: Weighted $\kappa$ measure for naive and expert ratings.

We also looked into the similarity of expert and naive ratings. We computed Pearson's correlation coefficients for each dimension between expert and naive ratings for the selected 100 words. The results indicate that the correlation is very high for pleasantness (0.826), and high for both activation (0.609) and imagery (0.582). In all cases, the correlations were statistically significant ($t$-test, $p < 0.0001$).

These results suggest that our naive and experts ratings are indeed comparable, showing high correlations among them, as well as similar degrees of reliability. We thus conclude that our online labeling setting may function as a reasonable, fast, unexpensive proxy for more traditional expert labelings.

# Chapter 4

# Evaluation

Next we proceeded to evaluate the usefulness of our emotion lexicon. For this purpose, we developed a simple system for estimating affect along our three affective dimensions, and evaluated it on two different sentiment-analysis tasks. The first task consisted in a set of texts labeled by humans, and served to compare the judgments of human labelers with the predictions of our system. The second task consisted in classifying a set of user product reviews into 'positive' or 'negative' opinions, a common application for online stores.

## 4.1 Simple system for estimating affect

We created a simple computer program for automatically estimating the degree of pleasantness, activation and imagery of an input text, based on the lexicon described in the previous sections.

For each word in the lexicon, we calculated its mean rating for each dimension. Subsequently, for an input text $T$ we used *Freeling* to generate a full syntactic parsing, from which we extracted all $\langle$word, word-class$\rangle$ pairs in $T$. The system calculates the value for affective dimension $d$ using the following procedure:

$score \leftarrow 0$
$count \leftarrow 0$
for each word $w$ in $T$ (counting repetitions):
    if $w$ is included in $Lex$:
        $score \leftarrow score + Lex_d(w)$
        $count \leftarrow count + 1$
return $score/count$

where $Lex$ is our emotion lexicon, and $Lex_d(w)$ is the value for $w$ in $Lex$ for dimension $d$.

For example, given the sentence *"Mi amiga esperaba terminar las pruebas a tiempo"* (*"My female-friend was hoping to finish the tests on time"*), and assuming our lexicon contains the numbers shown in Table 4.1, the three values are computed as follows. First, all words are lemmatized (i.e., *mi amigo esperar terminar el prueba a tiempo*). Second, the mean of each dimension is calculated with the described procedure, yielding a pleasantness of 2.04, activation of 2.16 and imagery of 2.60.

| word | word-class | mean P | mean A | mean I |
|---|---|---|---|---|
| *amigo* | noun | 3.0 | 2.4 | 3.0 |
| *esperar* | verb | 1.2 | 1.0 | 2.8 |
| *terminar* | verb | 2.2 | 3.0 | 2.8 |
| *prueba* | noun | 1.8 | 2.4 | 2.2 |
| *tiempo* | noun | 2.0 | 2.0 | 2.2 |
| | mean: | 2.04 | 2.16 | 2.60 |

Table 4.1: Lexicon entries for the example text (P = pleasantness; A = activation; I = imagery).

It is important to mention that this system is just a proof of concept, motivated by the need to evaluate the effectiveness of our lexicon. It could be used as a baseline system against which to compare more complex affect estimation systems. Also, if results are good enough with such a simple system, this would indicate that the information contained in our emotion lexicon is useful, and in the future it could help create more complex systems.

## 4.2 Evaluation #1: Emotion estimation

The first evaluation task consisted in comparing predictions made by our simple system against ratings assigned by humans (our gold standard), on a number of sentences and paragraphs extracted from *Wikipedia* and *Los Cuentos*.

### 4.2.1 Gold standard

From each corpus we randomly selected 15 sentences with 10 or more words, and 5 paragraphs with at least 50 words and two sentences – i.e. 30 sentences and 10 paragraphs in total. These texts were subsequently rated by 5 volunteers (2 male, 3 female), who were instructed to rate each entire text (sentence or paragraph) for pleasantness, activation and imagery using the same three-point scale shown in Table 3.1. The weighted $\kappa$ measure for these ratings was 0.17 for pleasantness, 0.17 for activation and 0.22 for imagery. Consistent with the subjectivity of these

tasks, the degree of inter-labeler agreement was rather low, yet still above chance level. Note also that for pleasantness and activation the agreement level was lower for texts than for individual words, while the opposite was true for imagery.

### 4.2.2 Results

To evaluate the performance of our system, we conducted Pearson's correlation test for each affective dimension, in order to find the degree of correlation between the system's predictions for the 40 texts and their corresponding mean human ratings. Table 4.2 shows the resulting $\rho$ coefficients.

| System \ GS | Pleasantness | Activation | Imagery |
|---|---|---|---|
| Pleasantness | $0.59$ * | $0.15$ * | $-0.18$ * |
| Activation | $0.13$ * | $0.40$ * | $0.14$ * |
| Imagery | $0.16$ | $0.19$ | $0.07$ |

Table 4.2: Correlations between gold standard and system's predictions. Statistically significant results are marked with '*' ($t$-tests, $p < 0.05$).

The coefficient for pleasantness presented a high value at 0.59, which indicates that the system's estimation of pleasantness was rather similar to the ratings given by humans. For activation the correlation was weaker, although still significant. On the other hand, for imagery this simple system did not seem able to successfully emulate human judgments.

These results suggest that, at least for pleasantness and activation, our lexicon successfully captured relevant information regarding how humans perceive those affective dimensions. For imagery, it is not clear whether the information base did not capture useful information, or the estimation system was too simplistic.

### 4.2.3 Effect of word count on performance

Next we studied the evolution of performance as a function of the lexicon size, aiming at assessing the potential impact of increasing the number of words annotated by humans. Figure 4.1 summarizes the results of a simulation, in which successive systems were built and evaluated using the top 250, 350, 450, ..., 2350, 2450 and 2566 words in our lexicon.

The green line (triangles) represents the mean coverage of the system's lexicon on the gold standard texts ; the corresponding scale is shown on the right axis. Similarly to Figure 2.2, the coverage grew rapidly, starting at 18% when using 250 words to 44% when using all 2566 words.

The blue (circles), red (squares) and purple (diamonds) lines correspond to the correlations of the system's predictions and the gold standard ratings for pleasantness, activation and imagery, respectively; the corresponding scale is shown on the left axis. The black lines are a logarithmic function fit to each of the three curves ($\rho^2 = 0.90$, $0.72$ and $0.68$, respectively).

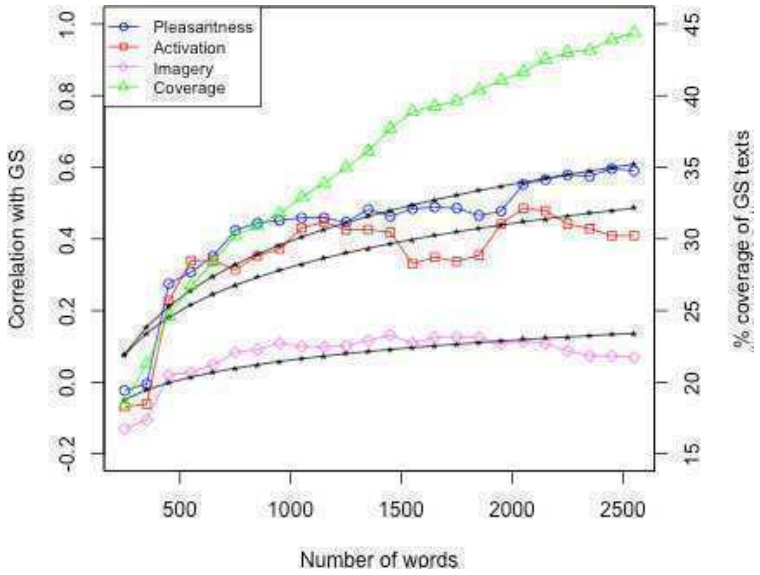


Figure 4.1: Evolution of the correlation between system predictions and Gold Standard, with respect to the lexicon size.

These results indicate that the system performance (measured as the correlation with human judgments) grew logarithmically with the number of words in the lexicon. Interestingly, the performance grew at a slower pace than word coverage. In other words, an increase in the proportion of words in a text that were known by the system did not lead to a similar increase in the accuracy of the predictions. An explanation may be that, once an emotion had been established based on a percentage of words in the text, the addition of a few extra words did not significantly change the outcome.

In consequence, if we wanted to do a substantial improvement to our baseline system, it would probably not be a good idea to simply annotate more words. Instead, it may be more effective to work on *how* the system uses the information contained in the emotion lexicon.

## 4.3 Evaluation #2: Classification of reviews

The second evaluation task consisted in using our baseline system for classifying user product reviews into positive or negative opinions.

### 4.3.1 Corpus

For this task we used a corpus of 400 user reviews of products such as cars, hotels, washing machines, books, cellphones, music, computers and movies, extracted from the Spanish website *Ciao.es*.[1] This is the same corpus used by [BTT09], who employed sentiment analysis tools in English to score Spanish texts after performing machine translation.

On *Ciao.es*, users may enter their written reviews and associate a numeric score to them, ranging from 1 to 5 stars. For this evaluation task, we made the assumption that there was a strong relation between the written reviews and their corresponding numeric scores. Following this assumption, we tagged reviews with 1 or 2 stars as 'negative' opinions, and reviews with 4 or 5 stars as 'positive'. Reviews with 3 stars were considered neutral, and ignored.

### 4.3.2 Results

We used our system in a very simple way for predicting the polarity of opinions. First we computed $M$, the mean pleasantness score on 80% of the reviews. Subsequently, for each review in the remaining 20%, if its pleasantness score was greater than $M$, then it was classified as 'positive'; otherwise, it was classified as 'negative'.

After repeating this procedure five times using 5-fold cross validation, the overall accuracy was 62.3%. Figure 4.2 shows the evolution of the system's accuracy with respect to the number of words in the lexicon. The green line (triangles) represents the mean coverage of the system's lexicon on user review texts; the corresponding scale is shown on the right axis. The blue line (circles) corresponds to the classification accuracy; the corresponding scale is shown on the left axis. The black line is a logarithmic function fit to this curve ($\rho^2 = 0.80$).

Notably, with as few as 500 words the accuracy is already significantly above chance level, which is 50% for this task. This indicates that our emotion lexicon managed to capture information on pleasantness that may aid the automatic classification of positive and negative user reviews.

Also, similarly to our first evaluation task, we observe that the accuracy increased as more words were added to the lexicon. However, it did so at a logarith-

[1] http://ciao.es

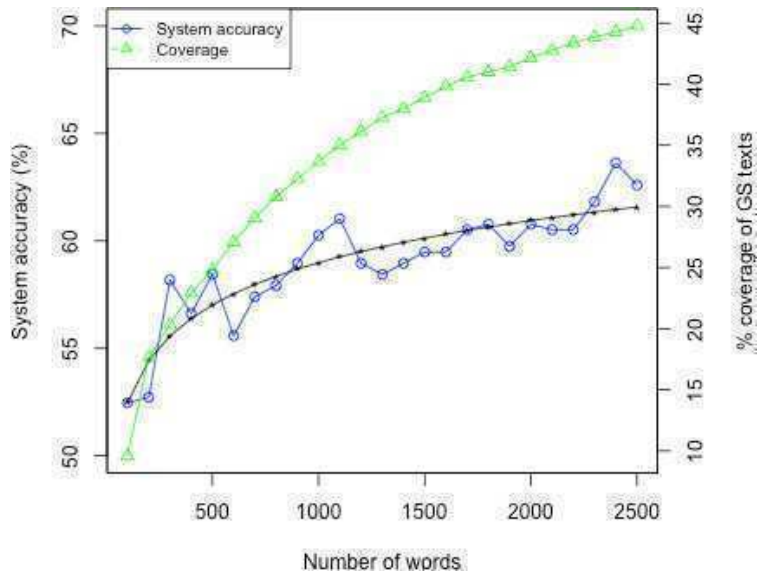


Figure 4.2: Evolution of the classification accuracy with respect to the size of the lexicon.

mic pace slower than the growth of the word coverage on the user reviews. This suggests that adding more words labeled by humans to the lexicon would only have a limited impact on the performance of this simple system.

### 4.3.3 Comparison with previous work

As mentioned above, Brooke et al. (2009) experimented with a number of cross-lingual approaches to sentiment analysis on the same corpus of product reviews in Spanish taken from *Ciao.es*. These approaches were substantially more complex than the simple system described in the previous sections. They included lexicon-based approaches capable of handling negations, amplifications and what the authors call *irrealis expressions* that should be ignored (e.g., conditional expressions such as *"sería hermoso"* / *"it would be beautiful"*), as well as language-independent approaches purely based on machine-learning techniques (e.g. support vector machines that used unigram counts and other text-based attributes). All of these approaches yielded accuracies ranging from 66 to 74.5%.

With those results as a reference point, the 62.3% accuracy achieved by the simplest-possible system based on our lexicon looks very promising. It encourages further research for investigating the effectiveness of more complex systems which take advantage of emotion lexicons specifically built for Spanish, such as the one presented in this work, both in comparison to and in addition to using cross-lingual resources.

# Chapter 5

# Conclusion

In this work we presented an emotion lexicon in Spanish, with words labeled by human volunteers for three affective dimensions – pleasantness, activation and imagery, inspired by the English DAL created by Whissell (1986, 1989). The annotations of these three dimensions were weakly intercorrelated, indicating a high level of independence of each other. Additionally, the agreement between volunteers was quite high, especially for pleasantness and activation, given the subjectivity of the labeling task.

To evaluate the usefulness of our lexicon, we built a simple emotion prediction system. When used for predicting the same three dimensions on new texts, its output significantly correlated with human judgments for pleasantness and activation, but the results for imagery were not satisfactory. Also, when used for classifying the opinion polarity of user product reviews, the system managed to achieve an accuracy better than random. These results suggest that our lexicon successfully captured useful information of human perception of, at least, pleasantness and activation. For imagery, either it failed to capture any significant information, or the system we created was too simple to exploit it accordingly.

Regarding the evolution of the system's performance as a function of the size of the lexicon, the results were clear. When more words were included, the system performance increased only at a logarithmic pace. Thus, working on more complex systems seems to be more promising than adding more human-annotated words.

In summary, this work presented an emotion lexicon that may come handy to researchers and developers of sentiment analysis tools in Spanish. Additionally, it may be useful for disciplines that need to select emotionally balanced word stimuli, such as Neuroscience or Psycholinguistics.

# Acknowledgments

This work was funded in part by ANPCYT PICT-2009-0026 and CONICET. The authors thank Carlos 'Greg' Diuk and Esteban Mocskos for valuable suggestions and comments; Julian Brooke, Milan Tofiloski and Maite Taboada for kindly sharing the Ciao corpus; and Facundo Carrillo, Viviana Cotik and Ramiro Gálvez for their help with data labeling.

# Bibliography


[BTT09] J. Brooke, M. Tofiloski, and M. Taboada. Cross-linguistic sentiment analysis: From English to Spanish. In *International Conference on Recent Advances in NLP*, pages 50–54, Borovets, Bulgaria, 2009.

[CDCT+01] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in hci. *Signal Processing Magazine, IEEE*, 18(1):32–80, 2001.

[Coh68] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[GBR02] J.R. Gray, T.S. Braver, and M.E. Raichle. Integration of emotion and cognition in the lateral prefrontal cortex. *Proceedings of the National Academy of Sciences*, 99(6):4115, 2002.

[PCR+10] L. Padró, M. Collado, S. Reese, M. Lloberes, and I. Castellón. Freeling 2.1: Five years of open-source language processing tools. In *Internat. Conf. on Language Resources and Evaluation (LREC)*, 2010.

[PRBM12] V. Pérez-Rosas, C. Banea, and R. Mihalcea. Learning sentiment lexicons in spanish. In *Int. Conf. on Language Resources and Evaluation (LREC)*, 2012.

[WFP+86] C. Whissell, M. Fournier, R. Pelland, D. Weir, and K. Makarec. A dictionary of affect in language: Iv. reliability, validity, and applications. *Perceptual and Motor Skills*, 62(3):875–888, 1986.

[Whi89] Cynthia Whissell. The dictionary of affect in language. *Emotion: Theory, research, and experience*, 4:113–131, 1989.

[YNBN03] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using NLP techniques. In *3rd IEEE International Conference on Data Mining*, pages 427–434. IEEE, 2003.

# Appendix A

# Login and instructions pages

Figures A.1 and A.2 show the screenshots of the login and instructions pages of our web interface for rating words.


Figure A.1: Screenshot of the login page.

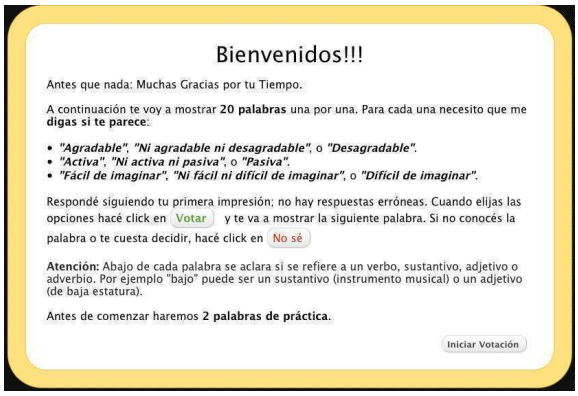


Figure A.2: Screenshot of the instructions page.