# Mexican Spanish Affective Dictionary

Adriana Peña Pérez Negrón(✉), Luis Casillas, Graciela Lara, and Mario Jiménez

Computer Science Department, CUCEI, Universidad de Guadalajara, Blrvd. Marcelino García Barragán 1421, 44870 Guadalajara, Mexico
{adriana.pena, luis.casillas}@cucei.udg.mx,
{graciela.lara, mario.jimenez}@academicos.udg.mx

**Abstract.** In the study of Affective Computing, the lexicon-based approach represents a useful mechanism that consists on using rated words to understand their affective role in a conversation. One of the most used lists of affectively rated words is the Affective Norms for English Words (ANEW), which evaluates the dimensions of pleasure, arousal and dominance for the English language. This list has been translated for other languages such as German or Spanish with effective results; however, there is not an affective lexicon for Mexican Spanish, rated by Mexicans. Based on the ANEW methodology, but using the most frequently words in Mexican Spanish language, similar to emoticons figures for the evaluation and an *ad hoc* app to collect the data, a list with means and standard deviation for Mexican Spanish words was obtained. Results and main differences with the ANEW study are here discussed.

**Keywords:** Emotional rating · Affective Computing · Lexicon-based approach ANEW

## 1 Introduction

Emotions influence our behavior and therefore all of what we do. The comprehension of this phenomena inspired Affective Computing, a growing interdisciplinary research area aimed to the study of the relation that involves emotions and computers [1, p. 50], conjoining sentiment analysis and emotion recognition [2]. For both, the classification of emotions is required, two main taxonomies have been applied in this area: the discrete model, which considers a small set of core or basic universal emotions (e.g. happiness, anger, sadness) and the dimensional model. According to the dimensional theory, the affective connotation of words can be described through a reduce number of dimensions [3]; this study is based on this model.

In this context, according to Poria et al. [2], lexicon-based approaches using dimensional representations for affect, provide a way of describing emotional states in a more manageable way, better dealing with non-discrete emotions and variations in emotional states over the time. Although, this approach does not consider the fact that more than one emotion can be experienced at a time.

One of the best-known and more used affective dictionaries is the Affective Norms for English Words [4], which contains 1,034 words rated in three dimensions. Based on


Á. Rocha et al. (Eds.): WorldCIST'18 2018, AISC 745, pp. 63–71, 2018.
https://doi.org/10.1007/978-3-319-77703-0_6

the Osgood [5] study two main factors, valence and arousal, explain most of the variance in affective meaning. Valence indicates the way a person evaluates a stimulus in a degree from pleasant to unpleasant, and arousal the degree of activation from calm to exciting. A third less-strongly-related dimension denominated dominance or control was included, with a range from in control to out of control [4].

The ANEW database has been adapted to different languages through translation. Among which are, for Spanish by Redondo et al. [6], they included five objective and three subjective psycholinguistic indexes. Redondo et al. [6], found that the translated words were similarly understood in Spanish compared with English, although with some sex and cross-cultural differences. For European Portuguese by Soares et al. [3], exploring differences between European Portuguese. And for German by Schmidtke [7], for 1,003 words, they also included several objective psycholinguistic variables.

Even that the ANEW database has been broadly accepted, there should be words in Mexican Spanish that are not commonly or frequently used, and also the emotional connotations of words can vary from culture to culture [8], as resulted in the study conducted by Wei et al. [9] with Chinese words. In such a way that we explored in this study a database specifically built with the most frequently used words in Mexico.

### 1.1 Related Work

Affective dictionaries can be found in the endeavors leading to the production of the General Inquirer [10], intensely used up to mid 90s. It was developed through diverse stipends from the USA National Science Foundation, and the Research Grant Councils of Great Britain and Australia. The main goal of this tool was the content analysis of text, including disambiguation, frequency detection and a sort of assessment for the words. Such text analysis was supported by a taxonomic structure organizing words in groups. The General Inquirer is still available with novel implementations using modern programming languages [11].

For the General Inquirer the selection of words was made creating random sequences of letters not related to concepts. This random sequence of characters eventually created words. The words were then linked to documents to map the terms. This was a first approach to artificially create a dictionary oriented to the content analysis [10]. It has been criticized for the lack of association with the communication among people [12].

Kučera and Francis [13] developed a collection of words based on the frequency of use, starting with a million words. These words were compared with written documents, which included stories by students with topics of abuse for emotions, and essays of science fiction. The result was a total of 8,700 words.

A direct spin-off from the General Inquirer [10] is the Lasswell Value Dictionary [14], where the General Inquirer [10] terms were reorganized. This word collection was a breakthrough regarding the affective orientation. Words were organized in categories such as: power, rectitude, respect, affection, wealth, well-being and enlightenment. This dictionary added awareness elements referring to transitions and the sentiments involved, categories created with communication goals. It also includes notions of contextual time/space, transactional indicators and force indicators.

Inheritor of previous efforts regarding affective analysis the Dictionary of Affect in Language [15], also known as DAL, is a compilation of 8,742 terms in English, where each word was rated in three dimensions: pleasantness, activation and imagery. Where the pleasantness index refers to the amount of pleasure produced by the word, the activation index refers to the amount of response produced by persons at emotional level, and the imagery index implies the visual description or figurative awareness given by persons to words.

As mentioned, one of the most influencing collection in the affective analysis over words, text and document's content is the ANEW [4]. The construction of this dictionary involved the assessment of words using a dimensional approach, such as the one used in the DAL. The ANEW dimensions are: pleasure, arousal, and dominance. These dimensions conform the Self-Assessment Manikin (SAM), graphical elements that were applied in the questionnaires to rate the words.

From the included words in the ANEW are 150 words used in the study conducted by Mehrabian and Russell [16]. They identified relevant variables (e.g. color, heat, light and sound) that involved the environmental psychology. Their proposal is that the environmental stimulus is linked to the conductive answers, based on the primary emotional mechanic of arousal, pleasure and dominance. The ANEW also includes the 450 used in Bellezza, Greenwald, and Banaji [17] taken from a frequency study of Thorndike and Lorg [18].

Grefenstette et al. [19] provided a refreshing view for affective definition over words. Known as the Clairvoyance Affect Lexicon, this study considers the context of words, as well as the syntactic roles in the sentence. Hence, adjectives and adverbs may influence the emotive load linked to words.

Among the alternative approaches for affective dictionaries is the developed by Tsai et al. [20], a two-step method that combines iterative regression with bounding normalization to build the sentiment dictionary, based on concepts and coherently involving commonsense knowledge. This approach implies the construction of auto-adjusted complex networks, through non-supervised learning mechanisms. It was adapted for Spanish by Bermudez-Gonzalez et al. [21], with complex algorithms to define emotive loads for words according to their use in the analyzed documents. The self-adjustment approach is an interesting choice that requires surveying stages; nevertheless auto-adjustment could miss some specific aspects bound to deep operation of humans.

Regarding the construction of affective dictionaries for the Spanish language, it can be highlighted Sidorov [22], a collection containing a Spanish Emotion Lexicon, also known as SEL. This dictionary contains 2,036 words with the Probability Factor of Affective (PFA) involving at least one basic emotion from: *joy, anger, fear, sadness, surprise, and disgust*. Manually marked by 19 annotators (scale: null, low, medium, high) certain thresholds on agreement were included. Unfortunately this approach has constrains regarding the data gathering through non-automated stages.

Due to the huge volume of data and information available nowadays, the use of affective dictionaries is a useful resource, which emerged as an affective evaluation alternative in business, marketing, education and security, among others. An exhaustive review with comparisons among dictionaries and strategies was made by Ribeiro et al. [23], providing a clear perspective of their current situation for affective and sentiment analyses.

As some of these dictionaries, we decided to use the most frequently used words scheme to construct a first approach to a corpus of Mexican Spanish words, as explained in the next section.

### 1.2 Words Selection

Languages involve diversity, and therefore collecting the words used by a social group is a demanding resources task. Such effort was concreted by Lara et al. [24] in the Dictionary of the Usual Spanish in Mexico, one of the most complete collections of the words used by Mexicans that includes the lingo from territorial regions, as well as the Mexicans' lingo at different historic periods. This collection was crossed with other sources to build our database.

A second source used is the collection of words from Varela et al. [25], which contains the 5,000 most frequent words in the textbooks of the Mexican Education Secretariat (SEP), for the basic education level. These textbooks are freely distributed to both public and private schools in Mexico.

Then a third source used is the most frequent words obtained in the Wiktionary [26]. This resource of the most frequently used words for different languages was developed throughout automated mechanisms. The sources of the words are movies and TV series scripts, subtitles, and electronic books. It is worth to mention that the Spanish collection from Wiktionary [26] is not specific for Mexican Spanish, nevertheless it was included due to the presence of global communication skills, already included in most languages.

These three threads were crossed to get a first set of words for the Mexican Spanish Affective Dictionary; that is to say, the words that were in all the three sources were extracted.

From this set of words were kept those that were considered might have an emotional relation without context following this steps:

- Following Wallace and Carson [27] adjectives and nouns were included. However, for them verbs and adverbs have an alternate noun or adjective form in English, which does not entirely apply for Spanish and therefore verbs were included. Adverbs were left out because they complement verbs, nouns, other adverbs or phrases.
- The names of the letters "a, b, …z" were discarded.
- The words *si* (yes) and *no* (no) were eliminated.
- Words without meaning in context like: *un* (a or an), *este* (this) or *ese* (that), were eliminated.
- The months, days and colours were also eliminated.
- Words related to quantity were eliminated such as: *mucho* (a lot) or *grande* (big).
- Words related to a position in space were also removed such as: *arriba* (up) or *encima* (on).
- Other homonymous words that might represent ambiguity during evaluation were also excluded. For example *banco*, which means both bench and bank in Spanish.

A total of 858 words compose the final set of words to be evaluated, the list can be consulted at:

https://drive.google.com/drive/folders/1TwqLUvkJmGTJReo3T6l1nJeamke4XD7M?usp=sharing.

These words were simply and automatically translated from Spanish to English using Google Translate getting only 227 words that matched the ANEW.

## 2 Rating the Words

The dimension of the words to be evaluated are, same as the ANEW, valence, arousal and control. An Android application was developed with this aim, an easy to share mechanism; the application is available in Google Play Store under the name RAPEM (stands for Affective Rate for Mexican Spanish Words). An open call was made to the undergraduate students and professors of the Computer Science and Informatics Engineering Schools to evaluate the words.

In the app, the first screen is a brief explanation of the study. At the bottom of it the user can or not "accept" to participate on it. A second screen collects the users' data name, age, gender, and if he/she is a professor or a student. Although the application is open to any one who wants to fulfil the study, only students and professors were asked to evaluate the words. As of today we keep collecting data. At this moment 192 persons had participated, most of them were men with only around 5% of female, with ages between 20 to 22 years; and an average of 4.8 evaluations by word.

Once the user captures his/her data, (i.e. name, age and gender), the words start to appear with graphics to evaluate them on each dimension. ANEW uses the graphical elements called the Self-Assessment Manikin (SAM) with a nine scale for each word. Because we are using an app, we decided to use more like emoticons figures with a five

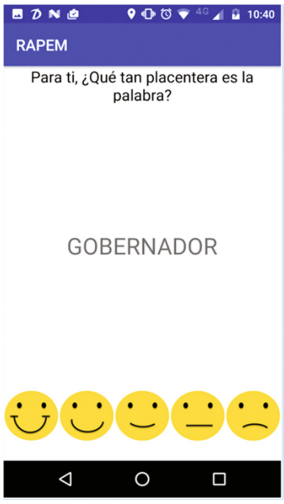


**Fig. 1.** Application screen for the word governor in Spanish

scale, eliminating the inter graphics scale used in ANEW. On the top of the screen there is a question, which helps to remain to the user, in case he/she needs it, the dimension they are rating. For example "For you, how pleasant is this word?" Then the word in capital letters appears in the middle of the screen, and at the bottom they can rate the word by touching one of the emoticon faces, see Fig. 1.

The words are randomly assigned by the app that operates on-line. A server receives the data that is scattered over a normalized database. There are two additional scripts running independently, designed to produce the statistics for each word.

### 2.1 Results

Table 1 shows a set of 15 words with results. In the first column is the word in Spanish and in parenthesis its translation to English. Second, third and fourth columns present the obtained mean of the rate in the three dimensions with its standard deviation (SD) in parenthesis. The mean was obtained by giving a value in the range of 1 to 5 for each of the 5 emoticons: 5 for the most pleasant word to 1 for the unpleasant; 5 for in control and 1 to out of control; 5 for calm to 1 for exciting.

**Table 1.** Set of rated words.

| Word Spanish (English) | Valence mean (SD) | Arousal mean (SD) | Dominance mean (SD) |
|---|---|---|---|
| *animal* (animal) | 2.30 (1.21) | 2.10 (0.69) | 2.20 (1.36) |
| *colegio* (college) | 3.10 (1.69) | 2.70 (1.21) | 3.30 (2.01) |
| *culpa* (fault) | 3.70 (1.61) | 4.10 (0.69) | 3.50 (2.25) |
| *directo* (direct) | 3.20 (1.36) | 3.00 (1.40) | 3.60 (1.24) |
| *esfuerzo* (effort) | 2.00 (1.20) | 2.00 (1.20) | 2.30 (2.81) |
| *fe* (faith) | 2.55 (1.16) | 2.27 (1.47) | 2.82 (2.51) |
| *fin* (end) | 2.10 (0.89) | 2.40 (0.24) | 3.10 (1.49) |
| *fortuna* (fortune) | 2.25 (1.52) | 2.75 (2.02) | 2.75 (2.69) |
| *informe* (report) | 2.60 (1.04) | 3.00 (1.00) | 3.30 (1.41) |
| *libro* (book) | 2.60 (0.84) | 2.30 (0.61) | 2.50 (1.65) |
| *lograr* (achieve) | 1.83 (1.31) | 2.25 (1.52) | 2.33 (0.89) |
| *necesidad* (need) | 3.40 (1.04) | 2.80 (0.76) | 3.80 (0.96) |
| *pelota* (ball) | 2.20 (1.96) | 2.20 (1.96) | 2.30 (2.21) |
| *permiso* (permission) | 3.21 (1.45) | 2.57 (1.24) | 2.57 (1.67) |
| *tratar* (treat) | 2.80 (1.96) | 2.40 (0.84) | 2.60 (1.64) |

Just as an example, three words of this set that match in the ANEW are presented in Table 2. The ANEW nine-scale was adjusted to a five-scale in order to make a comparison. It can be observed that the numbers are not only different, but that they present a different order. That is to say, the word "book" has a higher valence than "fault" in English, but in Spanish it has a lower value in valence.

**Table 2.** Rated words in the set and in the ANEW.

| Word Spanish (English) | ANEW | | ANEW | | ANEW | |
|---|---|---|---|---|---|---|
| | Valence mean | Valence mean | Arousal mean | Arousal mean | Dominance mean | Dominance mean |
| *culpa* (fault) | 3.70 | 2.45 | 4.10 | 1.76 | 3.50 | 2.87 |
| *libro* (book) | 2.60 | 4.09 | 2.30 | 0.99 | 2.50 | 3.79 |
| *tratar* (treat) | 2.80 | 5.26 | 2.40 | 1.03 | 2.60 | 4.13 |

## 3 Discussion and Future Work

We selected the most frequently used word in Mexican Spanish from the textbooks for basic education in Mexico [25] and the Spanish words in movies and TV series scripts, subtitles, and electronic books [26]. Those words were crossed with those in the Dictionary of the Usual Spanish in Mexico [24] to get the most frequently used words in Mexican Spanish.

A depuration was made to avoid words with no affective meaning out of context, creating a list of words that were evaluated in three dimensions: valance, arousal and control through an app, in order to build a Mexican Spanish Affective Dictionary with a corpus of 858 words. Because this list is not a translation, it does not imply words interpretations and includes words that might not be used in other languages.

In a simple automatic translation some words matched with the ANEW list. Compared with the ANEW, results show differences in the values for these dimensions. A factor that might generate these difference could be cultural or because of the number of evaluations. A higher scope of evaluations requires to be collected in order to discard the second possibility. Also, data was collected only from students and professors of a university. However, data should include not only professors and students, but also a varied population.

Future work, besides getting more evaluations by word, will be to make a statistical evaluation of the words that match with the ANEW. Also, increase the corpus of the Dictionary with lingo not used in textbooks.

In any language, there are not polite words with a high affective value, for their nature they will not appear in any of the selected sources; however, we intend to include them in the future in a not open to public study. Finally, following Grefenstette et al. [18], we will consider the context words and their syntactic role, to create roles to deal with sentences written in Mexican Spanish.

## References

1. Picard, R.W. (ed.): Affective Computing. MIT Press, Cambridge (1997)
2. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017)
3. Soares, A.P., Comesaña, M., Pinheiro, A.P., Simões, A., Frade, C.S.: The adaptation of the Affective Norms for English words (ANEW) for European Portuguese. Behav. Res. Methods **44**(1), 256–269 (2012)

4. Bradley, M.M., Lang, P.J.: Affective Norms for English Words (ANEW): instruction manual and affective ratings. Technical report C-1, The Center for Research in Psychophysiology, University of Florida, pp. 1–45 (1999)
5. Osgood, C., Suci, G., Tannenbaum, P.: The Measurement of Meaning. University of Illinois, Urbana (1957)
6. Redondo, J., Fraga, I., Padrón, I., Comesaña, M.: The Spanish adaptation of ANEW (Affective Norms for English Words). Behav. Res. Methods **39**(3), 600–605 (2007)
7. Schmidtke, D.S., Schröder, T., Jacobs, A.M., Conrad, M.: ANGST: affective norms for German sentiment terms, derived from the affective norms for English words. Behav. Res. Methods **46**(4), 1108–1118 (2014)
8. Triandis, H.C., Osgood, C.E.: A comparative factorial analysis of semantic structures in monolingual Greek and American college students. J. Abnorm. Soc. Psychol. **57**, 187–196 (1958). https://doi.org/10.1037/h0046061
9. Wei, W.L., Wu, C.H., Lin, J.C.: A regression approach to affective rating of Chinese words from ANEW. In: Affective Computing and Intelligent Interaction, pp. 121–131 (2011)
10. Stone, P.J., Dunphy, D.C., Smith, M.S.: The General Inquirer: A Computer Approach to Content Analysis. M.I.T. Press, Oxford (1966)
11. Stone, P.J.: Welcome to the General Inquirer Home Page. Harvard College, Cambridge (2002). Accessed 15 June 2017
12. Paltoglou, G., Gobron, S., Skowron, M., Thelwall, M., Thalmann, D.: Sentiment analysis of informal textual communication in cyberspace. In: Proceedings Engage 2010. Springer LNCS State-of-the-Art Survey, pp. 13–25 (2010)
13. Kučera, H., Francis, W.N.: Computational Analysis of Present-Day American English. Dartmouth Publishing Group, Sudbury (1967)
14. Lasswell, H.D., Namenwirth, J.Z.: The Lasswell Value Dictionary. Yale University Press, New Haven (1969)
15. Whissell, C.M.: The dictionary of affect in language. In: Emotion: Theory, Research, and Experience, vol. 4 (1989). The Measurement of Emotions
16. Mehrabian, A., Russell, J.A.: An Approach to Environmental Psychology. MIT Press, Cambridge (1974)
17. Bellezza, F.S., Greenwald, A.G., Banaji, M.R., Mahzarin, R.: Words high and low in pleasantness as rated by male and female college students. Behav. Res. Methods Instrum. Comput. **18**, 299–303 (1986)
18. Thorndike, E.L., Lorge, I.: The Teacher's Word Book of 30,000 Words. Teachers College, Columbia University (l944)
19. Grefenstette, G., Qu, Y., Evans, D.A., Shanahan, J.G.: Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes. In: Computing Attitude and Affect in Text: Theory and Applications, pp. 93–107. Springer (2006)
20. Tsai, A.C.R., Wu, C.E., Tsai, R.T.H., Hsu, J.Y.J.: Building a concept-level sentiment dictionary based on commonsense knowledge. IEEE Intell. Syst. **28**(2), 22–30 (2013)
21. Bermudez-Gonzalez, D., Miranda-Jiménez, S., García-Moreno, R.U., Calderón-Nepamuceno, D.: 28 generating with supervised learning techniques. New perspectives on teaching and working with languages in the digital era, p. 327 (2016)
22. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Gordon, J.: Empirical study of machine learning based approach for opinion mining in tweets. In: Mexican International Conference on Artificial Intelligence, pp. 1–14. Springer, Heidelberg, October 2012

23. Ribeiro, F.N., Araújo, M., Gonçalves, P., Gonçalves, M.A., Benevenuto, F.: SentiBench-a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci. **5**(1), 1–29 (2016)
24. Lara, L., Peña, G., Gordillo, L., Segovia, F., Valadez, C., Urrea, A., Cuahtecontz, N., Trujillo, E.: Diccionario del Español de México (DEM). El Colegio de México, A.C. (2012). http://dem.colmex.mx. Accessed 20 Aug 2016
25. Varela Barraza, J.A., Cabrera González, F., Zarabozo Enríquez de Rivera, D., de Rivera, D., Larios Villa, Y., González Ortiz, M.: Las 5000 palabras más frecuentes en los libros de texto oficiales de la educación básica en México. Revista electrónica de investigación educativa **15**(3), 114–120 (2013)
26. Wiktionary (n.d.): Apéndice:Palabras más frecuentes del español. https://es.wiktionary.org/wiki/Apéndice:Palabras_más_frecuentes_del_español. Accessed 10 Aug 2016
27. Wallace, A., Carson, M.: Sharing and diversity in emotion terminology. Ethos **1**(1), 1–29 (1973)