International Journal of Computational Linguistics and Applications, Vol. 6, No. 1, 2015, pp. 63–79
Received 26/02/2015, Accepted 21/05/2015, Final 16/06/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

What do our Children Read About? Affect Analysis of Chilean School Texts

CLAUDIA MARTÍNEZ
JORGE FERNÁNDEZ
Universidad Católica de la Santísima Concepción, Chile

ALEJANDRA SEGURA
CHRISTIAN VIDAL-CASTRO
CLEMENTE RUBIO-MANZANO
Universidad del Bío-Bío, UBB, Concepción, Chile

ABSTRACT

We present a study of the affective character of 1st to 8th year Chilean school texts, to which we applied lexicon-based affect analysis techniques to identify 6 basic emotions (anger, sadness, fear, disgust, surprise and happiness). First, we generated a corpus of 525 documents, 18176 paragraphs and 137516 words. Then, using the affective words frequency, we built a classifier based on Emotion Word Density to detect emotions in the texts. Our results show that the predominant affective states are happiness (58%), sadness (16%) and fear (12%). The 6 basic emotions are present in most literary forms with uniform relative density except for songs, where anger is absent. Classifier performance was validated by comparing its results against the opinions of experts in the field, and its results show an above-average conformity (accuracy = 63%), above-average predictive capacity (precision = 69%) and good classifier sensitivity (recall = 80% and f-measure = 93%).

Keywords: Affect analysis, sentiment analysis, subjectivity analysis

1. INTRODUCTION

Affect analysis is a branch of natural language processing focused on the development and application of techniques that estimate the emotive aspects of a text [1]. Several psychological theories commonly used in affect analysis classify emotions into between six and nine basic classes. For example, Ekman [2] defines the six basic emotions as anger, disgust, fear, joy, sadness and surprise. In [3], Plutchik and Kellerman add guilt, interest and shame to the basic emotion classes. These approaches assume that all other emotions can be assigned to these basic groups, on which other theories and classifications are based. This study aims to detect emotions in school texts provided by the Chilean Education Ministry for 2014 based on a lexical resource.

The rest of the article is organized as follows: in the next section, it reviews the relevant bibliography about lexicon-based affect analysis techniques. The next two sections describe the various techniques, tools and metrics used in this study. Experiments and their results are described in sections 5 and 6, while section 7 and 8 present the expert validation stage and a discussion of the results.

2. BACKGROUND

Affect analysis is usually applied using an approach based on machine learning, where automated or supervised learning algorithms detect and classify texts into affective categories. Another common approach uses lexical structures, specifically lexicon containing sets of words categorized according to their affective content [4], [5].

Affect analysis and sentiment analysis have been applied to literary forms to gather information about the emotions that these texts might provoke the reader. Most works that have applied the lexicon-based strategy to this purpose have taken advantage of the availability of English-language lexicons. For example, Neviarouskaya in [6] applied the ATtitude Analysis Model to the classification of fairy tales using the Izard's theory of 9 emotions

[7] to detect attitudes, and using polarity analysis to detect judgments and appreciations. Later, Mohammad [8] proposed a technique for visualizing emotion-related words based upon the frequency of the emotions. This study does not focus on the detection of emotions themselves, but rather the visualization of the presence of emotionally-loaded words in certain children's stories, specifically 192 fairy tales from the Brothers Grimm. Similarly, Ptaszyinki et al. present a similar approach for the Japanese language in [9], where a similarity metric was used to expand a rather small group of emotionally-charged words (containing 503 nouns) into an emotions dictionary (containing 15612 verbs), which was later used in the syntactic rules applied to the Snow White story. The study´s results were not encouraging mainly due to dictionary quality and syntactical rules incoherence. In [10], Sugimoto and Masahide experimented with an automatic reading system that was able to read aloud texts and novels with emotional emphasis. Texts were classified according to their emotion word distribution. At the same time, a sentence's emotion was classified according to the emotions associated to its nouns, adjectives and/or verbs. In [11], Ptaszynski et al. propose a method for extracting emotion information from narrative texts by analyzing anaphoric expressions, from which the emotional state of each character in the each stage of the narrative can be determined. Finally, [12] compares studies about the use of lexicons for affect analysis in tweets.

In this study, in contrast to previous studies, school texts in Spanish are analyzed by applying the metric proposed by Mohammad [8], using an affective lexicon also in Spanish.

3. METHODOLOGY

Figure 1 presents this study´s work methodology, from text retrieval to the gathering of results and their evaluation.

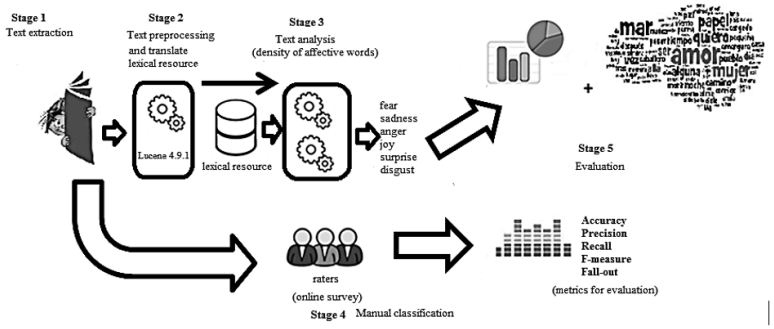


Figure 1. *Study methodology*

The methodology included four stages; the first stage was an exploratory search of official 2013 Ministry of Education texts published by Santillana Editions, next we extracted and classified the school texts from 1st to 8th grades into poems, stories, novels, songs, articles, news, interviews, and biographies. In the third stage, texts were preprocessed using the API Lucene 4.9.1, removing stopwords, whitespaces and special characters; this stage ends with the application of the SnowBall stemming algorithm. Next, an expert translated the lexical resource from English to Spanish and then detected the affective words frequency based on the emotion word density metric [8]; at the same time, in the fourth stage, it applied a survey to two raters so as to do a comparison between manual and automatic classification.

4. DEVELOPMENT

4.1. *Text extraction stage*

Texts for this study were obtained from the catalog of school texts provided free of charge by the Chilean Ministry of Education to students and teachers of all public and publicly-subsidized schools in Chile. In particular, texts were extracted from the 1st year to 8th year Language and Communication textbooks published by Editorial Santillana. These texts are categorized into 15 literary forms: Story, Reading, Poem,

Dramatic work, Fable, Legend, Myth, Article, Novel fragment, Interview, News report, News item, Biography and Other. We generated a corpus of 525 documents, 18176 paragraphs and 137516 words.

4.2. *Text preprocessing stage*

In this phase, texts were cleaned and prepared for the classification phase, so as to improve the classifier´s performance and speed up the classification process, thus aiding the affect analysis process. This phase includes several transformation stages: online text cleaning, white space removal, abbreviation expansion, stopwords removal and stemming. Finally, the preprocessed text undergoes a filtering stage, during which several pattern selection functions are applied [13].

Text cleaning involved removing any characters no considered useful for this study (stopwords, white space, special characters, etc.). The unit of analysis chosen for this work is the paragraph. Thus, periods were not removed as they act as paragraph separators. Both lexicon words and text words were stemmed prior to the matching phase. The Apache Lucene[1] 4.9.1 API was used, which includes the following modules:

a. **StopAnalyzer**: This analyzer is used to remove articles, pronouns, conjunctions, etc., as they are considered irrelevant for the purposes of this study.
b. **WhitespaceAnalyzer**: This analyzer is used to remove extra whitespace, yielding a text free of extra whitespace and concatenated words.
c. **SimpleAnalyzer**: This analyzer is used to transform all text to lowercase and to eliminate all punctuation marks, yielding a text that is case consistent.
d. **SnowballAnalyzer**: This analyzer is used to stem both the text words and the lexicon using the Snowball algorithm.

R-Project package version 3.0.1 was used in the final preprocessing stage was done using, in particular its RCurl, Rjson

[1] https://lucene.apache.org/core/

and TM packages were utilized for corpus generation and preprocessing. Likewise, the Rcpp, RColorBrewer andWordcloud R-Project packages were used to generate frequency diagrams and word clouds. Based on Hassan-Montero's work [14], we chose to use tagcloud visualization techniques to represent the most frequently occurring words in the texts.

An emotion dictionary based on WordNetAffect was used as the lexical resource [15]. As this dictionary was available only in English, it was translated into Spanish by an expert. This resource was built by selecting a subset of synsets suitable for representing affective concepts; additional information was also added to the affective synsets without defining new ones. We kept the structure of the original lexicon, thus generating a set of infrequently used words in Spanish for doing affective analysis. This dictionary includes a total of 1523 words (nouns, adjectives and verbs) organized into 6 emotion classes: anger, disgust, fear, sadness, joy and surprise. Additionally, some synonyms were added to expand the knowledge base. Table 1 shows that, after the preprocessing stage the corpus was reduced in size from 799033 words to only 137516 words.

Table 1. *Number of words per literary form*

Literary form	Including stopwords	Not including stopwords	Percentage of total
Story	147720	24737	18%
Reading	396126	62641	46%
Poem	56811	9637	7%
Dramatic work	39103	6536	5%
Legend	26740	4339	3%
Myth	25653	4183	3%
Novel fragment	14327	2424	2%
Article	39302	5717	4%
Biography	5721	860	1%
Song	1633	891	1%
News report	10096	1546	1%
Interview	17061	5996	4%
News	4503	1171	1%
Fable	3545	1735	1%
Others	10692	5103	4%
Total	799033	137516	100%

The "Other" category includes texts that cannot be categorized as belonging to any other literary form.

4.3. *Text analysis stage (density of affective words)*
For this phase, we wrote a Java script to calculate the Emotion Word Density [8], which represents the relative percentage of a word over the universe of emotion words analyzed, as shown in Equation 1. Through this parameter, we can determine the frequency of emotions present in each literary form.

$$F(e_1,t_1) = \frac{count(e_1,t_1)}{N_1} \quad (1)$$

where $F(e_1, t_1)$ is the relative frequency of affect word e_1, given that $e_1 \in t_1 \wedge e_1 \in$ lexical resource R_1, $e_1 \in t_1$, $count(e_1\ t_1)$ is the occurrence of e_1 in text t_1 y N_1 the total volume of affect words found in text t_1. According to Mohammad [8], this analysis strategy might not be always trustworthy to determine if a given phrase is expressing a certain emotion, but can be used to determine whether a particular section of text has more expressions of emotion than other sections. As mentioned before, the school texts under study are categorized into 15 literary forms, ranging from narrative texts to argumentative texts, including descriptive texts and dialogues.

5. RESULTS

Figures 2 and 3 present the main results of the text analysis phase (automatic labeled). From Figure 2, it can be seen that the predominant emotion is joy (58%), followed by sadness (16%) and fear (12%). Also, Figure 3 shows that the six basic emotions are present in most literary forms with a uniform relative density except for the case of songs, where all emotions except anger are present. However, their absolute densities are very different, with the longer literary forms such as stories or reading having higher emotion word densities.

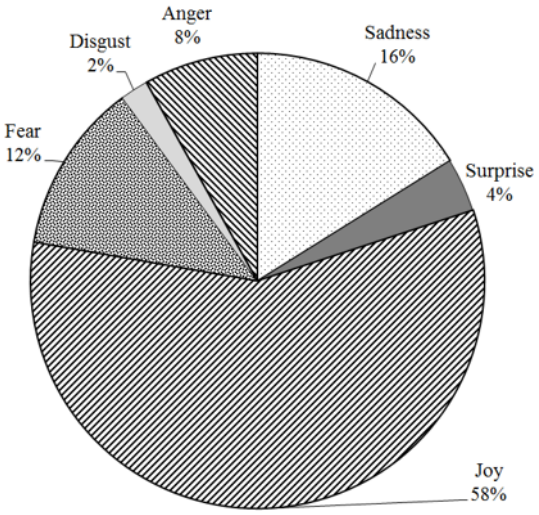


Figure 2. *Emotion density for the corpus*

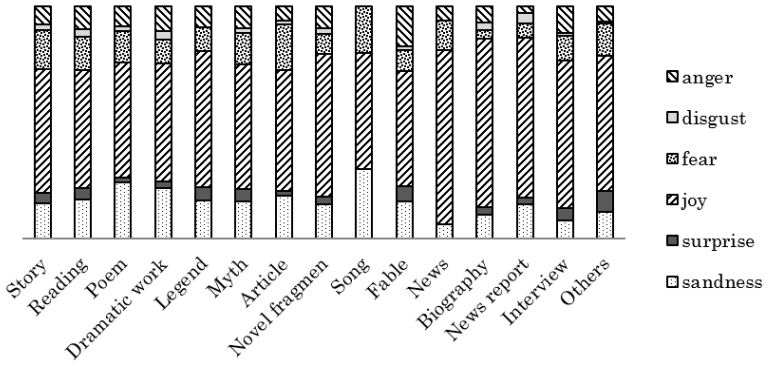


Figure 3. *Emotion relative density for each literary form*

As an example, we present detailed results for an instance of the Fable literary form: the 2nd year school text "The rainbow fish" in its Spanish translation. Figure 4 shows relative emotion densities as well as its corresponding tag cloud generated using the affect lexicon.

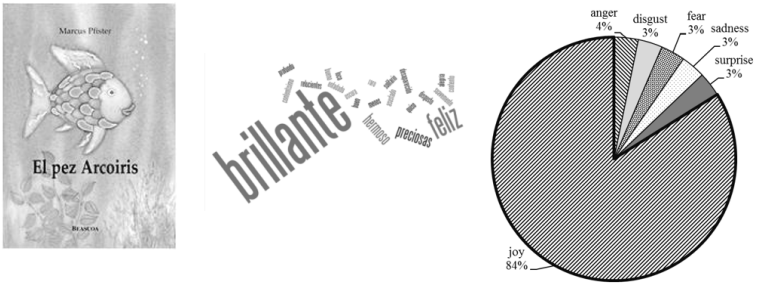


Figure 4. *Results for "El pez arcoiris", a text categorized in the Fable literary form*

In a further analysis phase, the 15 literary forms were clustered according to their size in words to aid the subsequent expert evaluation. Two homogeneous clusters were defined, comprising 54% and 46% of the words in the corpus, respectively. These clusters, as well as the literary forms and their respective sizes, are shown in Table 3.

Table 3. *Clustering of the literary forms*

	Literary form	**Number of documents**	**Number of words**	**Number of paragraphs**
Cluster 1	Story	55	24737	3435
	Poem	137	9637	1321
	Dramatic work	11	6536	909
	Legend	21	4339	333
	Myth	16	4193	322
	Article	28	39302	914
	Novel fragment	5	2424	333
	Fable	24	1735	13
	News item	5	1171	105
	Biography	7	5721	133
	News report	3	1546	235
	Interview	5	5996	461
	Song	17	891	58
	Other	28	5103	392
	Subtotal	362	74875	8964
Cluster 2	Reading	163	62641	9212
	Subtotal	163	62641	9212
	Total	525	137516	18176

Figure 5 presents relative emotion density results per cluster, where it can be seen that all six emotions are present in both clusters with uniform relative densities.

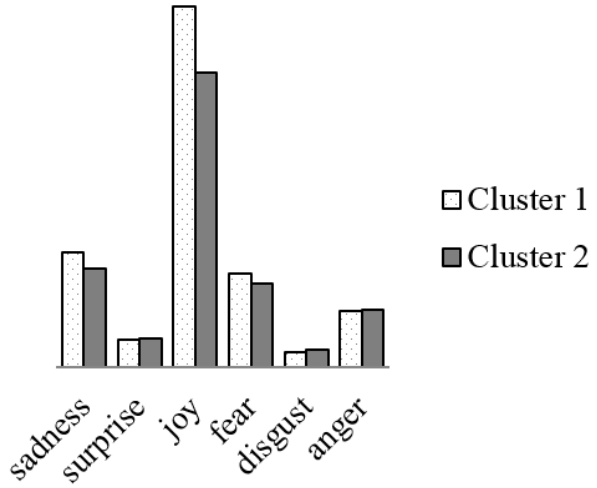
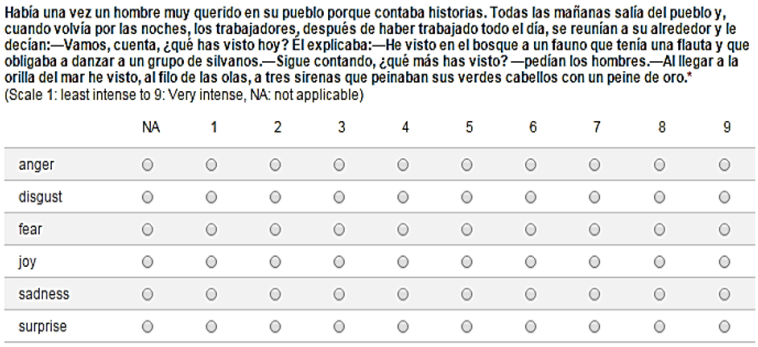


Figure 5. *Relative emotion density for each literary form cluster*

6. EVALUATION STAGE

The results of automatic classification were validated by gathering opinions and value judgments from experts in the field. To this end, it created 24 different surveys including 11 different paragraphs, for a total sample of 264 paragraphs. Each survey was designed so as to determine the emotional intensity of each paragraph regarding each of the six basic emotion classes. Assessment was done using the "Self-Assessment Manikin" (SAM) scale, containing nine categories ranging from least intense to very intense.



Había una vez un hombre muy querido en su pueblo porque contaba historias. Todas las mañanas salía del pueblo y, cuando volvía por las noches, los trabajadores, después de haber trabajado todo el día, se reunían a su alrededor y le decían:—Vamos, cuenta, ¿qué has visto hoy? Él explicaba:—He visto en el bosque a un fauno que tenía una flauta y que obligaba a danzar a un grupo de silvanos.—Sigue contando, ¿qué más has visto? —pedían los hombres.—Al llegar a la orilla del mar he visto, al filo de las olas, a tres sirenas que peinaban sus verdes cabellos con un peine de oro.*
(Scale 1: least intense to 9: Very intense, NA: not applicable)

	NA	1	2	3	4	5	6	7	8	9
anger	○	○	○	○	○	○	○	○	○	○
disgust	○	○	○	○	○	○	○	○	○	○
fear	○	○	○	○	○	○	○	○	○	○
joy	○	○	○	○	○	○	○	○	○	○
sadness	○	○	○	○	○	○	○	○	○	○
surprise	○	○	○	○	○	○	○	○	○	○

Figure 6. *Example survey used to evaluate a text's emotional intensity*

As shown in Figure 6, for each emotion class, the expert must mark the emotional intensity felt for the given text. Stratified proportional sampling was used to calculate the number of paragraphs to include in each survey category in accordance to population characteristics. Each kind of survey included paragraph samples from both clusters (48% of paragraphs from cluster 1 and 52% of paragraphs from cluster 2). Each one of the 24 kinds of surveys was answered by two experts. Given the complexity of the evaluation instrument, no inter-agreement analysis was performed at this research stage. Each expert evaluated all six emotion classes, some emotion classes share the polarity or generate similar emotions, additionally, the surveys included a 1 to 9 intensity scale. These characteristics also make the probability of concordance be very low.

The WordNetAffect graph structure classifies emotions using four different polarity levels: positive emotion, negative emotion, neutral emotion and ambiguous emotion. In this study we consider only two polarities: joy and surprise are considered positive emotions, while anger, disgust, sadness and fear are considered negative emotions.

We also take the following four premises into account:

Premise 1: We consider the survey results to be the truth.

Premise 2: If the survey and the automatic classifier match the emotion, we consider it a True Positive (TP) or True Negative (TN), correspondingly.
Premise 3: If the survey and the automatic classifier do not match the emotion but match the polarity, we also consider it a TP or TN, correspondingly.
Premise 4: If the survey and the automatic classifier do not match neither the emotion nor the polarity, we consider it a FP or FN, correspondingly.

Then, we discarded all cases for which the automatic classifier did not detect any emotions, and those cases for which automatic classification was ambiguous (same number of occurrences for 2 or more emotion classes).

Finally, we construct the analysis matrix for the 6 basic emotion classes shown in Table 4.

Table 4. *Analysis matrix for the 6 emotion classes*

	NEGATIVE				POSITIVE	
Automatic classifier / Manual classifier	angry	fear	disgust	sadness	surprise	joy
angry	TN	TN	TN	TN	FP	FP
fear	TN	TN	TN	TN	FP	FP
disgust	TN	TN	TN	TN	FP	FP
sadness	TN	TN	TN	TN	FP	FP
surprise	FN	FN	FN	FN	TP	TP
joy	FN	FN	FN	FN	TP	TP

The performance of the automatic classifier can be studied by calculating the metrics described in equations (2) to (6). These metrics differ from the classic information retrieval metrics: whereas the classic metrics are calculated using all retrieved and relevant documents into account, these modified metrics consider

only a representative sample of the paragraphs, classified according to their literary form.

$$ACCURACY' = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

$$PRECISION' = \frac{TP}{TP + FP} \quad (3)$$

$$RECALL' = \frac{TP}{TP+FN} \quad (4)$$

$$F - MEASURE' = \frac{2 * PRECISION' * RECALL'}{RECALL' + PRECISION'} \quad (5)$$

$$Fall - Out(FRP) = \frac{FP}{FP + TN} \quad (6)$$

Metrics results for the automatic emotion classifier are presented in Table 5.

Table 5. *Metrics results*

Metric	Value
Accuracy	0,63190184
Precision	0,69105691
Recall	0,79439252
F-measure	0,93043478
Fall-out **(**FRP)	0,67857142

7. RESULTS AND DISCUSSION

Comparisons between the automatic classifier's and survey results show an overall above-average conformity to the correct values (accuracy = 63%), above-average predictive capacity

(precision = 69%) and good classifier sensitivity (recall = 80% and f-measure = 93%). The True Positive rate was 80%, while the True Negative rate was only 32%. This indicates that the automatic classifier has an acceptable prediction rate for the detection of positive emotions (recall = 80%), but fails to predict negative emotions (VPR = 32%), thus giving a high false alarm rate (FPR = 68%). This in turn may be explained by the high number of negative sentences in the texts, such as "I will never be happy again", "The children were unable to show their love," etc. It must also be noted that the percentage of positive emotions is always higher than the percentage of negative emotions for all literary forms except songs, which can be considered as having neutral polarity. Finally, we must remark on the fact that of the 137516 words in the corpus, only 8250 of them were emotion-related, of which only 1821 were unique words.

8. Conclusions and Future Work

This study shows that the predominant emotion in Chilean school texts is happiness (68%). Whether intended or not, this was the expected result. Less expected was that the second and third most common emotions are sadness (16%) and fear (12%). The current lexicon's restrictions notwithstanding, we can state that is possible to predict emotion in Spanish texts using lexicon-based affect analysis techniques. However, the classifier's regular predictive capacity and conformity with the true values may be explained by considering that only 6% of the words in the Spanish-language texts belong to an emotion class. Also, the lexicon used in this work was created for the English language and it may well be that its translation into Spanish did not always capture correctly every phrase sense. One of the lessons learned through this work is that the construction of a lexicon for affect analysis in Spanish requires, rather than a translation, an interpretation of the relevant phrases and their synonyms so as to represent their real sense and their real usage in the target language. Likewise, we did not consider colloquial Chilean expressions in this work. In this regard, this work is only the first step in our research, and we are working on including negative

sentence handling and the use of amplifiers and simplifiers. We also will work on local ironies, so as to better understand phenomena related to subjectivity. Finally, given that many primary school level texts are usually short in length and use a limited vocabulary (only 8250 of the 137516 words can be found in the affective lexicon resource used), these results validate our approach and encourage us to pursue further research in this area.

REFERENCES

1. Grefenstette, G., Qu, Y., Shanahan, J. G. & Evans, D. A. 2004. Coupling niche browsers and affect analysis for an opinion mining application. *Proceedings of Recherche D'Information Assistée par Ordinateur (RAIO)*.
2. Ekman, P. 1993. Facial expression and emotion. *American Psychologist*, 48/4, 384-392.
3. Plutchik, R. & Kellerman, H. 1980. A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, Research, and Experience: Vol. 1. Theories of Emotion* (pp. 3-33). New York: Academic Press.
4. Strapparava, C. & Mihalcea, R. 2008. Learning to identify emotions in text. In proceedings of the *2008 ACM Symposium on Applied Computing* (pp. 1556-1560).
5. Balahur, A.; Mihalcea, R. & Montoyo, A. 2014. Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. *Computer Speech & Language*, 28/1, 1-6.
6. Neviarouskaya, A., Prendiger, H. & Ishizuka, M. 2010. Recognition of affect, judgment, and appreciation in text. In proceedings of the *23rd International Conference on Computational Linguistics (COLING'10)* (pp. 806-814), Aug.
7. Izard, C. E. 1977. *Human Emotions*, New York: Plenum Press.
8. Mohammad, S. 2011. From *Once Upon a Time* to *Happily Ever After*: Tracking emotions in novels and fairy tales. In proceedings of the *ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)* (pp. 105-114), Portland, OR, Jun.
9. Ptaszynski, M., Rzepka, R., Araki, K. & Momouchi, Y. 2014. Automatically annotating a five-billion-word corpus of Japanese

blogs for sentiment and affect analysis. *Computer Speech & Language*, 28/1, 38-55.
10. Sugimoto F. & Masahide, Y. 2006. A method for classifying emotion of text based on emotional dictionaries for emotional reading. In proceedings of the *2nd International Symposium on Communications, Control and Signal Processing*, Marrakech, Morocco, Mar.
11. Ptaszynski, M., Dokoshi, H., Oyama, S., Rzepka, R., Kurihara, M., Araki, K. & Momouchi, Y. 2013. Affect analysis in context of characters in narratives. *Expert Systems with Applications*, 40/1, 168-176.
12. Soto, A., Cabrero, C., Menta, A. & Al, E. 2014. Estudio comparativo sobre el empleo de diccionarios en el análisis de sentimientos para textos cortos. *XVII Congreso Español sobre Tecnologías y Lógica Fuzzy (ESTYLF'14)*, Zaragoza, España, Feb.
13. Haddi, E., Liu, X. & Shi, Y. 2013. The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17, 26-32.
14. Hassan-Montero, Y. 2010. Usabilidad de los tag-clouds: Estudio mediante eye-tracking. *SCIRE: Representación y Organización del Conocimiento*, 16/1, 15-33.
15. Strapparava, C. & Valitutti, A. 2004. WordNet-Affect: An affective extension of WordNet. In *4th International Conference on Language Resources and Evaluation (LREC'04)* (pp. 1083-1086).

ACKNOWLEDGEMENTS

This paper is the result of the work of the SOMOS (SOftware - MOdelling - Science) research group, which is funded by the Dirección de Investigación and Facultad de Ciencias Empresariales of the Universidad del Bío-Bío, Chile. This work was partially supported by Organic. Lingua, CIP-ICT-PSP.2010.6.2 - Multilingual online services, project reference: 270999.

CLAUDIA MARTÍNEZ
DEPTO. INGENIERÍA INFORMÁTICA,
UNIVERSIDAD CATÓLICA DE LA SANTÍSIMA CONCEPCIÓN,
UCSC ALONSO DE RIBERA 2850, CONCEPCIÓN, CHILE.
E-MAIL: <CMARTINEZ@UCSC.CL>

JORGE FERNÁNDEZ
DEPTO. INGENIERÍA INFORMÁTICA,
UNIVERSIDAD CATÓLICA DE LA SANTÍSIMA CONCEPCIÓN,
UCSC ALONSO DE RIBERA 2850, CONCEPCIÓN, CHILE.
E-MAIL: <JAFERNANDEZ@ING.UCSC.CL>

ALEJANDRA SEGURA
DEPTO. SISTEMAS DE INFORMACIÓN,
UNIVERSIDAD DEL BÍO-BÍO, UBB
AVDA. COLLAO 1202, CONCEPCIÓN, CHILE.
E-MAIL: <ASEGURA@UBIOBIO.CL>

CHRISTIAN VIDAL-CASTRO
DEPTO. SISTEMAS DE INFORMACIÓN,
UNIVERSIDAD DEL BÍO-BÍO, UBB
AVDA. COLLAO 1202, CONCEPCIÓN, CHILE.
E-MAIL: <CHVIDAL@UBIOBIO.CL>

CLEMENTE RUBIO-MANZANO
DEPTO. SISTEMAS DE INFORMACIÓN,
UNIVERSIDAD DEL BÍO-BÍO, UBB
AVDA. COLLAO 1202, CONCEPCIÓN, CHILE.
E-MAIL: <CLRUBIO@UBIOBIO.CL>