Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 621 – 629

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

The Distribution of Affective Words in a Corpus of Newspaper Articles

David Brett, Antonio Pinna*

Dipartimento di Scienze Umanistiche e Sociali, Università degli Studi di Sassari, Via Roma 151, Sassari 07100, Italy

Abstract

This paper explores the possibility of automatically measuring and comparing affectiveness features in texts from different domains and sub-domains. More specifically our main research question concerns the distribution of affective words within the various subgenres of a newspaper corpus along three affective parameters: Valence (V), pleasantness, Arousal (A), the intensity of the emotion, and Dominance (D), the degree of control exerted by the perceiver over the stimulus. The study is largely based on work by Warriner et al. (in press), who recently divulged a study reporting affective ratings for approximately 14,000 English lemmas in terms of V, A and D. 100,000 token samples of newspaper language from 10 subsections of the Guardian newspaper were analyzed for the presence of these lemmas and the average V, A and D values for each subsection were calculated. Crime and Travel were seen to be those with the most atypical values.

Keywords: Affectiveness; emotion; valence; arousal; dominance; newspaper language;

1. Introduction

In the psycholinguistic tradition of the investigation of human emotions, scholars in the field of behavioral research have studied emotional reactions to verbal stimuli by means of lists of words rated by evaluators along the three components elaborated by Osgood et al. (1957) in their theory of emotions. These components are known as Valence (V), the degree of (un)pleasantness of the emotion evoked by the word (from miserable to happy); Arousal (A), the intensity of the emotion triggered by the word (from calm to intense); and Dominance (D), the degree to

* Corresponding author. Tel.: +39-079-229612; fax: +39-079-228211.
E-mail address: dedalo@uniss.it

doi:10.1016/j.sbspro.2013.10.690

which the word induces some emotion that makes the addressee feel vulnerable or dominant. In studies dealing with English, the most commonly used list of words rated for their affective values has been Bradley and Lang's (1999) ANEW (Affective Norms for English Words) that includes approximately 1,000 words rated for the three components. Among the various studies for which ANEW was used, an important and recent one is Leveau et al.'s (2012) work that employs various norms elaborated for different European languages to establish a metanorm that enables the assessment of texts along the affective dimensions of V and A. The value of the metanorm for the automatic evaluation of emotions in texts has been demonstrated by significant correlations between human judgment and computer assessment with respect to both components. Despite its evident merits, Leveau et al.'s (2012) metanorm is still based on a relatively limited number of words, including just over 6,000 words for the V dimension and over 4,000 words for the A component. A more extensive list was compiled by Warriner et al. (in press) to include approximately 14,000 lemmas rated on a 9-point scale along the three dimensions. This new list comprises nouns (63.5%), adjectives (22.5%) and verbs (12.6%), with just over 1% of words belonging to other parts of speech. Such a large list of affective words constitutes an important asset for psycholinguistic research, but it may also be an invaluable resource for corpus studies of the use of emotionally-loaded words in real communicative contexts.

Nomenclature

V	Valence
A	Arousal
D	Dominance
PoS	Part-of-Speech

2. Aim

The aim of the present paper is to apply the theoretical framework outlined above to a corpus-based investigation of journalistic prose, focusing on the distribution of affective words in a corpus of online newspaper articles. We are particularly interested in exploring whether Warriner et al.'s (in press) list can be employed to highlight notable differences in the emotional potential of the various sections of a newspaper corpus. Should this be the case, we may proceed to examine how the lemmas of the sections showing extreme values plot in terms of the three variables in the two combinations: V v. A and V v. D. A final query concerns whether particular part-of-speech categories bear more of the emotional charge than others.

3. Materials and methods

The corpus taken into examination was the Guardian corpus, collected by the authors at the University of Sassari, Italy. This is a 1M token corpus composed of texts downloaded from 10 different sections of the online version of the well-known British newspaper. The sections are Travel, (UK) Crime, Football, Banking, Politics, Education, Obituaries, Technology, World News and Films (details provided in Appendix A).

The method of study was two-fold: we initially extracted average values for the three affective scales in each of the 10 sub-sections of the corpus with the aim of verifying whether substantial differences on one or more of the scales could be observed. The second stage consisted of examining in greater detail the two sub-sections which were seen to contrast the most.

3.1. Analysis of the affective ratings of the ten sub-sections

The first stage involved comparing a wordlist of each section with the list of affective ratings provided by Warriner et al. (in press). When an item from the wordlist was also found on the affective ratings list the item was added to a new list with five columns: item -> freq. -> V -> A -> D. This operation was carried out by way of a

tailormade perlscript, written by the authors. The list was then ordered by frequency, and the average affective values were calculated on the basis of the top 200 most frequent items for each sub-section.

3.2. Analysis of the affective ratings of the two outlying sub-sections

Before examining the two outlying sections in greater detail, we lemmatized the wordlists. To be more specific, we tagged the texts using the CLAWS PoS Tagger provided by the University of Lancaster ([http://ucrel.lancs.ac.uk/claws/trial.html]; tagset: C5; output: vertical). We then extracted the types and tokens for each PoS category using another tailormade perl script. We lemmatized all the inflected verbs, and calculated the total frequency for each verb lemma. The 200 most frequent of these were then tallied with Warriner et al.'s list to recuperate their affective ratings. This operation was repeated for adjectives and nouns forms (excluding proper nouns).

The results of these procedures were then plotted on bubble graphs with the two combinations of the affective values for each macro PoS category on the x- and y-axes (V v. A; V v. D), with the size of the bubble representing the frequency of the lemma in the given sub-section.

4. Results and discussion

The results concerning the calculation of the average values for the ten sub-sections in terms of V v. A, and V v. D, can be seen in Figures 1 and 2. The two graphs display a very clear pattern: a central group is formed by such sub-sections as Politics, Football, World, Education and Technology, while in both graphs two outliers are clearly visible. Travel has the highest averages in terms of V and D, and amongst the lowest values in terms of A. On the other hand, Crime is the subsection with the lowest V and D values, and the highest A values but one. The other sub-sections proceed along a rough line which links these two outliers, with the exception of Film which has the highest A values and the second highest V values.

In a certain sense it is hardly surprising that Crime turns out to be that sub-section with the lowest levels of V. All crimes have victims, therefore it is perhaps predictable that D levels would be low. Strong emotions are provoked by many crimes, ergo the high A values. Similarly an oft-noted characteristic of the travel-writing/tourism genre is that of Euphoria. Dann (1996:65) describes how this genre almost invariably presents destinations and the activities to be carried out therein through rose-tinted glasses, glossing over any negative, unpleasant, or even mundane, aspects that may be encountered by the potential traveler/tourist. Furthermore, there is a tendency to emphasize only the most endearing, fascinating and spectacular facets of the proposed trip or destination. This could easily result in high average V values. While the ideal holiday for some may consist of thrill-seeking activities such as bungee-jumping and white water rafting, many see the vacation as a chance to unwind, hence the low A values. Finally, while not all can make important choices in the workplace, the choice of holiday destination is generally up to the holidaymaker him or herself. As a result, it is no wonder that D levels are far higher than those of Crime.

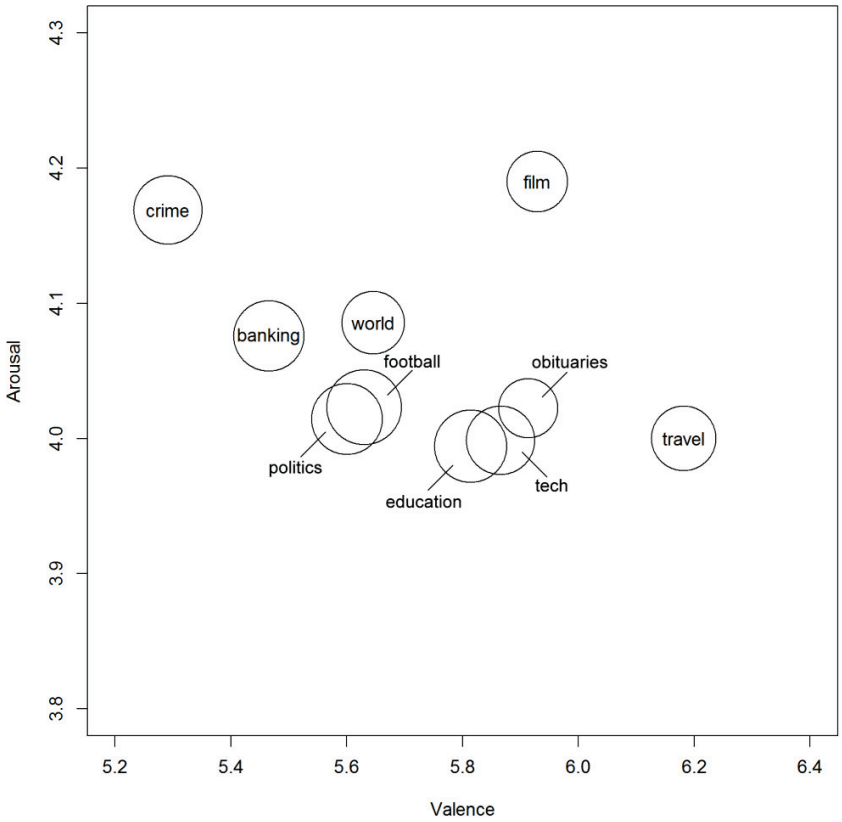

Fig. 1. Average values for V and A of ten sub-sections of Guardian corpus.

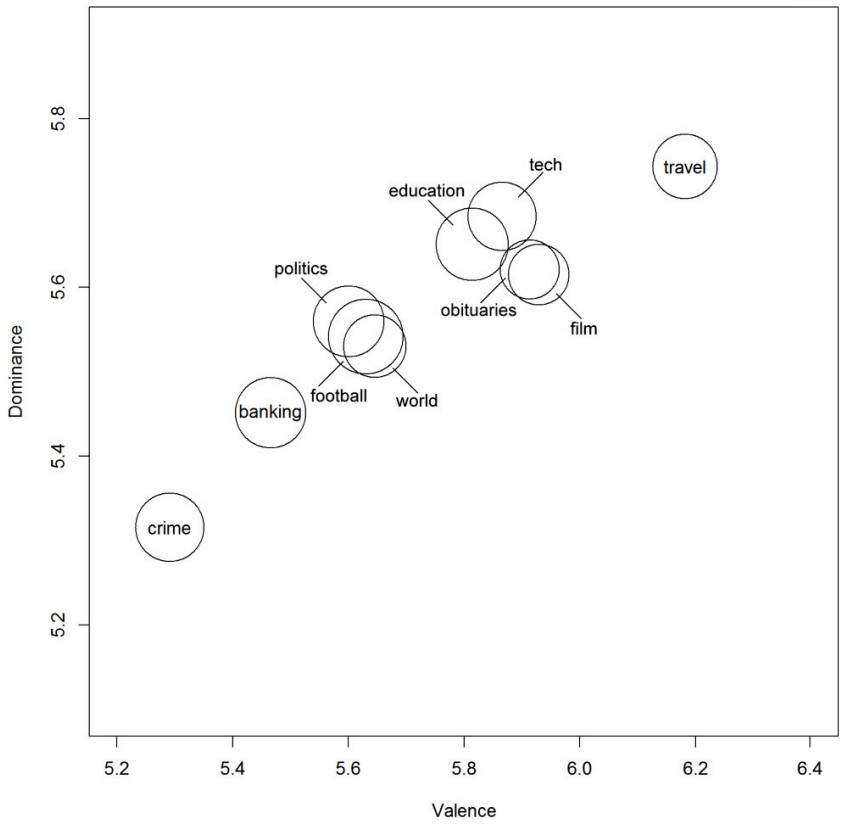

Fig. 2. Average values for V and D of ten sub-sections of Guardian corpus.

We then proceeded to examine in greater detail the two sub-sections showing the greatest differences. The V values plotted against those for A can be seen in Figures 3 to 5, which concern Verb, Adjective and Noun lemmas. In all cases a considerable amount of overlap can be noted, however, areas in which one sub-section is predominant can also be clearly seen. For example, in all three graphs the area covered by Crime is relatively larger than that covered by Travel words, as it includes lemmas, some of which with rather high frequencies, which have very low V values (even <2). Many of these items also have rather high A values; this is the case with die, kill and arrest in the Verb category, fatal, violent and angry in the Adjective category and attack, gang, murder, riot and suspicion in the Noun category. At the other extreme of the V scale, we see a series of lemmas pertaining exclusively to the Travel section. There are two areas characterized by high levels of V, but differentiated by way of their levels of A. In particular there are words with high levels of both V and A; among these we find the nouns adventure, festival, wildlife, skiing and travel, the adjectives free, pretty and excellent, and the verbs create, discover, travel, win and love. On the other hand, the group composed of lemmas with high V and low A includes the verbs sleep, book and play, the adjectives, simple, good and fresh, and the nouns valley, sea and park.

The central area in which terms are present from both sub-sections features a large quantity of what we may call generic terms, which could be common to any domain. These include references to Time and Place, such as week, night, day, town, country and way, as well as the terms time and place themselves.

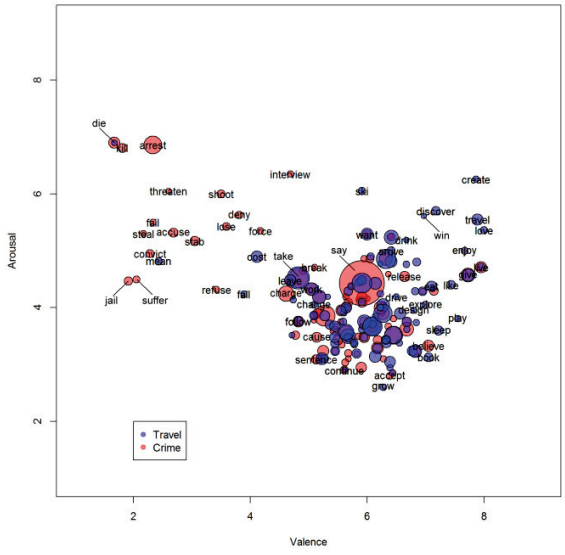

Fig. 3. Verbs - bubble graph of V and A values. Bubble size represents relative frequency
(To aid readability not all items have been labeled).

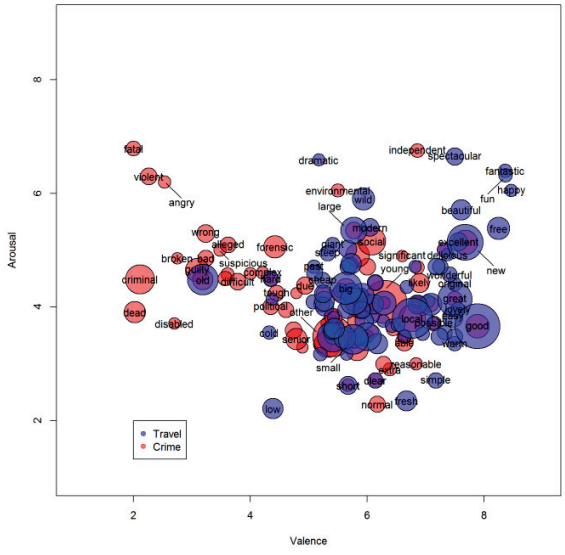

Fig. 4. Adjectives - bubble graph of V and A values. Bubble size represents relative frequency.
(To aid readability not all items have been labeled).

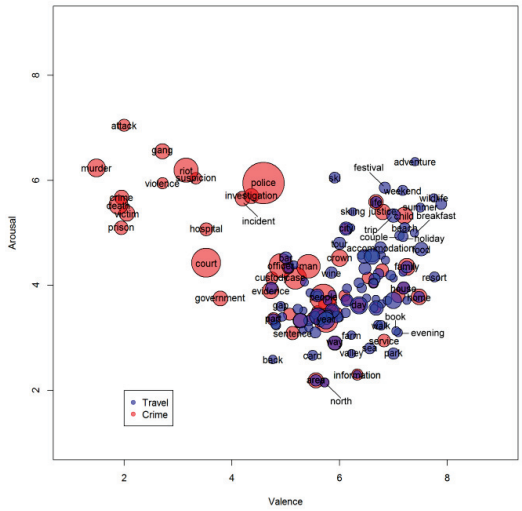

Fig. 5. Nouns - bubble graph of V and A values. Bubble size represents relative frequency.
(To aid readability not all items have been labeled).

The methodology we have adopted is effective in highlighting differences in affectiveness among the various newspaper subsections in our corpus, particularly in relation to Crime and Travel articles. An interesting feature of the graphs in Fig. 3 to 5 is the U-shaped relationship between the dimensions of V and A, a fact which is the result of extreme values of V having more arousing potential. The results show a tendency for news reports treating crime to be particularly effective in stirring intense, unpleasant emotive reactions in its reading public, while travel articles manage to achieve their persuasive power by means of pleasant, emotionally-loaded words covering a rather large (mid to high) spectrum of intensity.

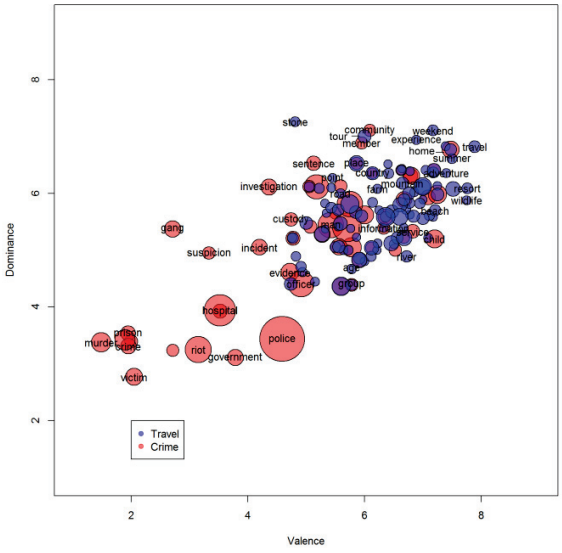

Fig. 6. Nouns - bubble graph of V and D values. Bubble size represents relative frequency.
(To aid readability not all items have been labeled).

We further analyzed the relationship between the V and D components. In this paper we only include Fig. 6 that shows how nouns are mapped with respect to these dimensions, as all graphs, including those plotting adjectives and verbs, display a similar pattern. There is indeed a linear relationship between these components, since words that evoke more pleasant feelings make people also feel more in control. The graph, however, points out that the most unpleasant and submissive responses are likely spurred by a set of crime-related lemmas in Crime reports (e.g. victim, murder, death, crime, prison and violence) characterized by low values of both V and D. At the other end of the scale, we almost exclusively find lemmas from the Travel section of the corpus (e.g. weekend, experience, travel and summer), presenting high values for both dimensions. In between these extremes, there is a considerable amount of overlap, including some generic terms mostly referring to Time and Place and belonging to both sections (e.g. year, week, area and country) with mid values in both variables.

The graphic representation of the results is useful for the identification of the words with similar/contrasting affective values in the sub-sections examined, however, to get an overall idea of the strength of the differences a statistical test was adopted: independent student's t was deemed to be the most suitable for the task at hand. The test was conducted on the results for the 200 most frequent lemmas in the various macro PoS categories. The data fed into the test were weighted for frequency, e.g. the adjective guilty, with a V value of 3.09, was not counted just once, but rather a number of times corresponding to its frequency (46). This weighting for frequency was necessary in order to avoid giving undue importance to items lower in the frequency ranking, as the most frequent items were often ten or more times more frequent than those close to the bottom of the list.

The application of student's t-test for independent samples confirms the clear distinction between the affective values for the most frequent items in the two sub-sections of the newspaper. Also in this case, the noun category is that which emerges as displaying the greatest contrast: the t-values are 57.09, -29.71 and 45.52, for V, A and D, respectively. These are extremely high values, considering that the degrees of freedom are 11,201 and $p<0.005$ is obtained with values at or above 2.576. The results for Verbs also confirm the strength of the contrast, with values of 25.09, -10.2 and 4.72 (df = 9,827), for V, A and D, respectively. The adjective category, while showing very high t values for V and D (26.52 and 18.93; df = 5,205), does not show a statistically significant difference for A (t = 0.13). We may therefore conclude that the differences in affective ratings of the two sub-sections are not only statistically significant, but also that they are strikingly so.

While the results reported above do seem to suggest that analyses using Warriner et al.'s affective word ratings may provide a useful yardstick to gauge emotional differences in texts, a number of caveats must be mentioned. The list of affective ratings does not include indication of part-of-speech, hence it is impossible to judge whether the affective rating for *mug*, for example, relates to the verb or the noun. Similarly, *mean* could have very different affective ratings if seen as a verb or an adjective. Furthermore, the different senses of words are not accounted for, therefore, when viewing the affective ratings for *bar* and *club*, we might presume that the informants were referring to what is probably the most frequent sense, that of PLACE FOR LEISURE ACTIVITY, however, we cannot exclude the possibility that some had the far more negative sense of BLUNT OBJECT in mind. However, such cases are limited, as is their probable impact on the validity of the data reported.

5. Conclusions

The methodology we have adopted appears to be effective in highlighting differences in affectiveness amongst the various newspaper subsections in our corpus, particularly in relation to Crime and Travel articles. Each of these two sections has been shown to have its own specific emotive profile. Semantic areas that characterize the Crime section and give it its distinctive affective flavor, as is the case with crime-related (e.g. *kill*, *murder*, *violent*) and institution-related words (e.g. *prison*, *court*, *police*), have lower values of V and D, and higher values of A than semantic areas specific to the Travel section. The latter, on the other hand, is distinguished by words with higher values of V and D that are however spread along the A scale from low to high, as is the case of the semantic sets including the ‘outdoor place’ words (e.g. *valley*, *park*, *lake*), the ‘excellence’ words (e.g. *excellent*, *perfect*, *great*), and the ‘feel-good’ words (e.g. *fun*, *free*, *love*).

The potential of this methodology may also be profitably applied to other kinds of linguistic analysis, such as the perception of ethnic groups or gender differences, shifts in affectiveness over time, and variations in emotive language in different types of newspaper. The extremely clear distinction revealed by the statistical analysis suggests

that the methodology may indeed be feasible in the search for the above differences, which are likely to be more subtle. A final point concerns part-of-speech categories. Our results show the Noun category to be that which is most revealing for affectiveness. Should further analysis confirm this, the finding would allow researchers to concentrate mainly on these parts-of-speech, rather than adjectives and verbs, which in our study appear to be less emotionally charged.

Appendix A. The composition of the Guardian corpus

	Sub-section	N.texts	Tokens	Types	N.authors
1	TRAVEL	96	103489	14209	62
2	CRIME	194	101505	9068	48
3	FOOTBALL	114	102249	9173	23
4	BANKING	160	101629	9115	38
5	POLITICS	132	101201	8794	37
6	EDUCATION	143	101916	9460	42
7	OBITUARIES	187	102094	13536	66
8	TECH	170	108581	11733	37
9	WORLD	173	101460	11809	78
10	FILMS	295	100599	16118	24

References

Bradley, M.,M. & Lang, P. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida. Available at http://www.uvm.edu/~pdodds/files/papers/others/1999/bradley1999a.pdf (last visited on 19 January 2013)

Dann, G. (1996), *The Language of Tourism. A Sociolinguistic Perspective.*CAB International

Leveau, N., Jhean-Larose, S., Denhière, G. & Nguyen, B. (2012). Validating an interlingual metanorm for emotional analysis of texts. *Behavior Research Methods*, *44*, 1007-1014.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

Osgood, C.E., Suci, G.J. & Tannenbaum, P. (1957) *The measurement of meaning*, University of Illinois Press

Warriner, A.B., Kuperman, V., & Brysbaert, M. (in press). Norms of V, A, and D for 13,915 English lemmas. *Behavior Research Methods*. Available at http://crr.ugent.be/archives/1003 (last visited on 19 January 2013)