Affective Lexicon Creation for the Greek Language

Elisavet Palogiannidi[1,2], **Polychronis Koutsakis**[4], **Elias Iosif** [2,3], **Alexandros Potamianos**[2,3]
[1]School of ECE, Technical University of Crete, Chania 73100, Greece
[2]"Athena" Research and Innovation Center, Maroussi 15125, Athens, Greece
[3]School of ECE, National Technical University of Athens, Zografou 15780, Greece
[4]School of Engineering and Information Technology, Murdoch University, Australia
epalogiannidi@isc.tuc.gr, p.koutsakis@murdoch.edu.au, {iosife,potam}@central.ntua.gr

Abstract

Starting from the English affective lexicon ANEW (Bradley and Lang, 1999a) we have created the first Greek affective lexicon. It contains human ratings for the three continuous affective dimensions of valence, arousal and dominance for 1034 words. The Greek affective lexicon is compared with affective lexica in English, Spanish and Portuguese. The lexicon is automatically expanded by selecting a small number of manually annotated words to bootstrap the process of estimating affective ratings of unknown words. We experimented with the parameters of the semantic-affective model in order to investigate their impact to its performance, which reaches 85% binary classification accuracy (positive vs. negative ratings). We share the Greek affective lexicon that consists of 1034 words and the automatically expanded Greek affective lexicon that contains 407K words.

Keywords: affective lexicon, affective ratings, valence, arousal, dominance, semantic similarity, sentiment analysis, emotion recognition

1. Introduction

Emotions are an integral part of human cognition, helping humans make decisions even in cases when uncertainty and ambiguity are introduced by cognitive analysis (Bechara et al., 2000). Hence, emotion detection attracts the interest of researchers from numerous areas such as psychology, linguistics, and engineering.

In order to model emotion as it is experienced by humans, resources that contain emotion information in various modalities become a necessity. (Lang et al., 1999) created a picture stimuli resource, known as International Affective Picture System (IAPS). Each stimulus is annotated on the affective dimensions Valence, Arousal and Dominance (V,A,D). The International Affective Digitized Sounds (IADS) is the corresponding resource with sound stimuli that was created by (Bradley and Lang, 1999b). (Bradley and Lang, 1999a) in their attempt to provide a set of normative emotional ratings for English words. They created an affective lexicon for English, a.k.a. ANEW (Affective Norms for English Words). ANEW contains 1034 words and each word is assigned with one score per affective dimension (V,A,D). Since then, numerous affective lexica in many languages have been created. (Redondo et al., 2007) translated the ANEW into Spanish and (Soares et al., 2012) into European Portuguese. (Montefinese et al., 2014) translated the ANEW into Italian and augmented it with more words. Additionally to ANEW, affective lexica with less or more entries can be found for a number of other languages such as German (Kanske and Kotz, 2010) and Dutch (Moors et al., 2013). Bilingual approaches have been reported as well, e.g., (Eilola and Havelka, 2010) created affective norms from English and Finish nouns. (Gilet et al., 2012) provided an affective lexicon of French attributes and investigated the age influence on the ratings collection process. The words of ANEW were also annotated with respect to discrete (categorical) emotions by (Stevenson et al., 2007).

Our contribution to the multilingual collection of affective lexica, is the creation of the first Greek affective lexicon. The affective lexicon is expanded automatically to cover a Greek vocabulary of 407K words using the affective lexicon expansion method of (Malandrakis et al., 2013). Performance in terms of correlation with human ratings and binary classification accuracy is estimated for the Greek and English language with consistent results.

2. Computational models for Emotion

Affective lexica contain words in a target language that are labeled with respect to affective dimensions. The scores of each affective dimension can be either categorical, i.e., basic emotions such as happiness and sadness, or continuous, i.e., a score in a given range over an emotional dimension.

In the continuous representation, emotion can be sufficiently described by three emotional dimensions: 1) $Valence$ is the subjective feeling of pleasantness or unpleasantness and ranges from highly positive to highly negative, 2) $Arousal$ is the subjective state of feeling activated or deactivated and ranges from calming or soothing to exciting or agitating, 3) $Dominance$ represents the controlling and dominant nature of the emotion[1] (Warriner et al., 2013).

The affective lexica we use follow a specific structure i.e., each entry consists of a word (seed word) and a triplet that indicates the affective scores for each dimension (V,A,D). The disadvantage of manually created lexica is that they have low language coverage, since they contain only a few thousand words. Thus computational methods are used to create or expand an already existing lexicon. (Malandrakis et al., 2013) expands such affective lexica covering a sig-

[1] For instance while both fear and anger are unpleasant emotions, anger is a dominant emotion, while fear is a submissive emotion

nificant fraction of the vocabulary of a language. SentiWordNet and WordNetAffect are examples of large affective resources created through computational models. In the first, (Esuli and Sebastiani, 2006) annotated automatically all WordNet (Miller, 1995) synsets. The latter was developed by (Strapparava and Valitutti, 2004) who represented the affective meanings by selecting and labeling a subset of WordNet synsets. Computational models for affective text usually incorporate small manually created resources (Malandrakis et al., 2011), or larger automatically created resources (Chaumartin, 2007).

3. The Greek affective lexicon

In this section, we describe the Greek affective lexicon creation process. We suggest that the words of an already existing affective lexicon can be transferred to the language of interest. Then, they can be shared to the target language's native speakers in order to collect the affective ratings. This process is depicted in Figure 1.

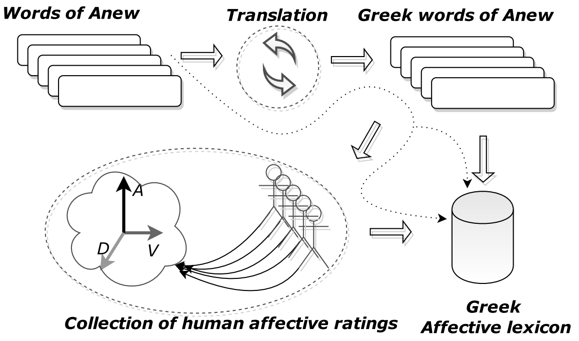

Figure 1: Affective lexicon creation process.

The main steps of this process are: (1) translation of ANEW words from English to Greek and (2) the collection of human affective ratings. The end result is the Greek affective lexicon that consists of 1034 words annotated on (V,A,D).

Translation of the ANEW words to Greek was split into two steps: translation proposals and post-correction. In this way we managed to address the word sense ambiguity issues between the two languages.

Collection of human affective ratings took place in a classroom of 105 engineering students (87 males, 18 females), native Greek speakers, aged from 19 to 30. Each word was rated with respect to (V,A,D) using the 9-point SAM scale (Bradley and Lang, 1994). Each participant was given a sheet with approximately 200 words, a sheet with instructions and the Self Assessment Scale (SAM) pictures [2] for providing their ratings by circling the corresponding image. In order to avoid context influence the word order was randomized. The participants had one hour at their disposal to complete the procedure.

3.1. Affective ratings

Affective annotation is a very subjective task and two (or more) people may provide very different ratings for a word. We assume that if many people repeat the annotation for a specific word, the expected value of the ratings will approximate the true affective score of this word. For this reason each word was rated by 20 participants. Regarding the ratings that were collected, no extreme annotator biases were noticed and hence, none of the annotators was excluded from the process. Fleiss' kappa was applied in order to measure the annotators' agreement for each word. The average Fleiss' kappa over all the words indicates fair agreement for valence ($\kappa = 0.29$) and and less so for dominance ($\kappa = 0.23$) and arousal ($\kappa = 0.20$)[3].

The distributions of the ratings collected for each affective dimension are shown in Figure 2, where dashed lines denote the median values. Comparing the ratings of the three dimensions, we observe the valence distribution is clearly bimodal, which is not as clear for the other two distributions. The concentration is higher on words with high valence. This result was also observed by (Montefinese et al., 2014).

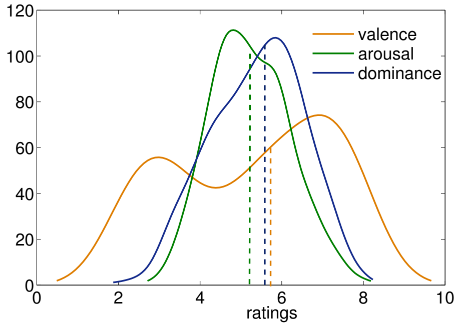

Figure 2: Distribution of the ratings collected for (V,A,D).

In Figure 3, the valence-arousal distribution for the words of the Greek affective lexicon is depicted. Each point corresponds to a word, and some words (translated into English) are shown. The distribution appears to follow the V-shape, well known from literature. The interpretation of V-shape is that negative and positive words generally tend to have high arousal, while words with neutral valence have neutral arousal as well.

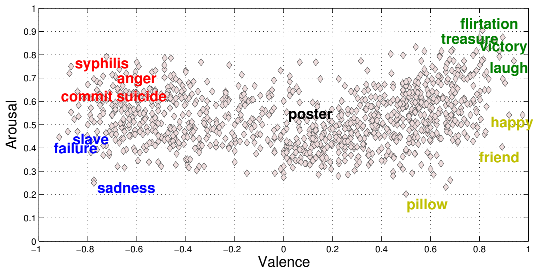

Figure 3: Valence-Arousal distribution of Greek ANEW.

[2] Nine pictures were used to describe each affective dimension as shown in (Bradley and Lang, 1994).

[3] The reported κ values were estimated on 9-point scale. When 9-point scale is transformed to 3-point scale the κ values are higher indicating good agreement for valence ($\kappa = 0.68$) and moderate agreement for arousal ($\kappa = 0.46$) and dominance ($\kappa = 0.49$)

In Figure 4, we show the distributions of the mean values of valence and arousal for the four languages in which ANEW is available. Greek, English (Bradley and Lang, 1999a), Spanish (Redondo et al., 2007) and European Portuguese (Soares et al., 2012) contain the same lexical information, i.e., the same 1034 English words have been translated to the respective languages. The distributions are consistent for the four languages, with the differentiation that the V-shape in Greek is not as clear as in the other languages for words with negative valence.

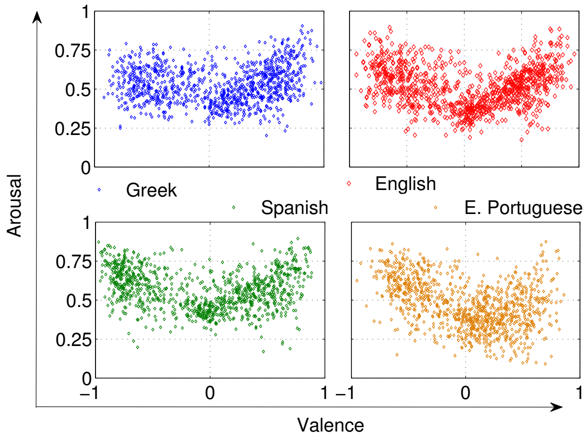

Figure 4: Valence-arousal distributions across languages.

4. Automatic affective lexicon expansion

The model we use for the automatic expansion of the affective lexicon was first proposed by (Malandrakis et al., 2013) and it is an extension of (Turney and Littman, 2002). It requires a small manually annotated affective lexicon for bootstrapping the process of estimating affective ratings of unknown words. The model aims to characterize the affective content of words from the reader perspective estimating scores in the continuous range [-1,1]. In order to start the process, an affective lexicon with human ratings and a semantic model for estimating the semantic similarity between two words are required. N words are selected from the affective lexicon to serve as seed words. The underlying assumption is that the affective rating of an unknown word can be estimated by using: a) the affective ratings of seed words weighted with a semantic similarity metric, that expresses how similar the seed word is to the word of interest, and b) a weight coefficient that aims to capture the importance of the selected seed to the affective estimation. The model is defined as follows:

$$\hat{\upsilon}(w_j) = a_0 + \sum_{i=1}^{N} a_i \upsilon(w_i) S(w_j, w_i), \qquad (1)$$

where w_j is the unknown word, $\hat{\upsilon}(w_j)$ is the affective rating of word w_j, $w_{1..N}$ are the seed words, $\upsilon(w_i)$ and a_i are the affective rating and the weight that corresponds to the seed w_i, respectively and a_0 is the bias. $S(\cdot)$ denotes a metric for computing the semantic similarity between two words. The weights a_i were assigned to each seed because not all seeds are equally salient for the affective ratings estimation of words. They were estimated by using Least Square Estimation (LSE) between the predicted and actual ratings, as described in (Malandrakis et al., 2013). Motivated by (Palogiannidi et al., 2015) we use a regularized LSE, i.e., Ridge Regression (RR) that incorporates a regularization factor that forces the weights to shrink toward zero.

4.1. Semantic similarity features

The $S(\cdot)$ metric used in (1) can be computed within the framework of distributional semantic models that adopt the distributional hypothesis of meaning, i.e., *"similarity of context implies similarity of meaning"*, (Harris, 1954). A contextual window of size $2H+1$ words is centred on the word of interest w_i and lexical features are extracted. For every instance of w_i in the corpus the H words left and right of w_i are extracted, formulating a feature vector x_i. The semantic similarity between two words, w_i and w_j, is computed as the cosine of their feature vectors:

$$Q^H(w_i, w_j) = \frac{x_i . x_j}{||x_i||\ ||x_j||}. \qquad (2)$$

The elements of feature vectors can be weighted according to various schemes based on the corpus frequencies of w_i and w_j (Iosif and Potamianos, 2010). In this work, a binary and a PPMI based weighting scheme are used. According to the binary scheme the vector elements are set either to zero or to one. For the PPMI weighting scheme, vector elements are weighted using the positive point-wise mutual information (PPMI) metric. The point-wise mutual information (PMI) between the word w_i and the n–th feature of its vector x_i, f_i^n, was computed as in (Church and Hanks, 1990):

$$PMI(w_i, f_i^n) = -log \frac{\hat{p}(w_i, f_i^n)}{\hat{p}(w_i)\hat{p}(f_i^n)}, \qquad (3)$$

where $\hat{p}(w_i)$ and $\hat{p}(f_i^n)$ are the occurrence probabilities of w_i and f_i^n , respectively, while the probability of their co-occurrence (within the H window size) is denoted by $\hat{p}(w_i, f_i^n)$. The probabilities were computed according to maximum likelihood estimation using corpus-based word frequencies. PMI is unbounded, yielding scores that lie in the $[-\infty, +\infty]$ interval. PPMI is equivalent to PMI with the modification that the negative scores are set to zero. This is based on the assumption that the contextual features that exhibit negative PMI with the target word (e.g., w_i) do not contribute to the estimation of similarity as much as the features characterized by positive PMI (Bullinaria and Levy, 2007).

In addition to the context-based metrics where second order word co-occurrences are used, the similarity of words can be also estimated by considering their first-order co-occurrence statistics. The underlying assumption is that the co-occurrence of words within a specified context serves as indicator for their semantic relatedness. In this work, we employ a widely-used metric, namely, Google-based semantic relatedness (G) proposed in (Gracia et al., 2006). The word co-occurrence was regarded at the sentential level. A comparison of both types of metrics is presented in (Iosif and Potamianos, 2015) for the task of similarity computation between nouns.

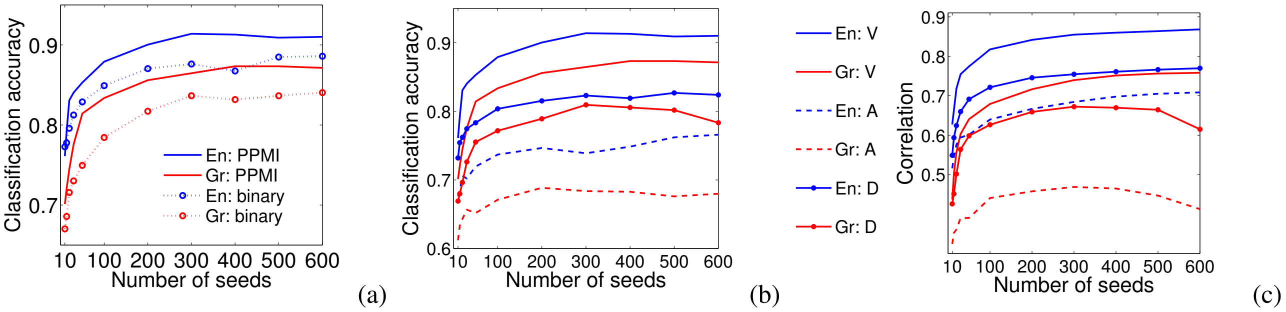

Figure 5: (a) Valence classification accuracy of binary and PPMI weighting schemes for English (En) and Greek (Gr) and V-A-D performance using PPMI weighting scheme for Greek and English: (b) classification accuracy, (c) Pearson correlation.

5. Experiments and evaluation

In order to estimate semantic similarities for a language, a corpus was created using web-harvested data as follows. A lexicon was defined for each language: 135K and 407K entries for English and Greek, respectively. For each lexicon entry a web search query was formulated, while the snippets of 1K top-ranked documents were downloaded and aggregated. Regarding the corpus-based Q^H similarity metric, a narrow window size was used ($H = 1$).

The affective lexicon expansion is evaluated on each language's affective lexicon, applying 10-fold cross validation, using different numbers of seed words. The experimental procedure and seed selection algorithm is as in (Malandrakis et al., 2013). In brief, N seed words are selected from the affective lexicon in order to create a balanced set (the sum of the seeds' affective ratings to be as close to zero as possible), and then a training phase follows in order to learn the weights that correspond to each seed. When ridge regression is used, a tuning step is necessary in order to estimate the appropriate regularization factor. Tuning takes place on held out data and the regularization factor is selected in order to maximize the correlation between the estimated and the human rated affective ratings. The experimental procedure is described in more detail in (Malandrakis et al., 2013; Palogiannidi et al., 2015). Two evaluation metrics were used: two-class (binary) classification accuracy and Pearson correlation coefficient with respect to human ratings for the Greek and the English language.

In Figure 5(a) we show the classification accuracy for valence, using binary and PPMI-based weighting scheme. Both for English and Greek, PPMI clearly outperforms the simpler binary feature vector scheme. In Figure 5(b) we present the classification accuracy and in Figure 5(c) the correlation performance for V-A-D using the PPMI weighting schemes. The performance of the affective model is high, especially for valence and dominance, which is expected, given that valence and dominance are highly correlated (0.86 in Greek and 0.84 in English). The performance achieved for arousal is relatively poor. The results are consistent with respect to the V-A-D dimensions for both languages, although results for Greek are consistently lower than for English for all three dimensions[4]. The affective model appears to be robust and high performing when at least 100 seeds are used reaching the highest performance for 500-600 seeds.

In Figure 6 (a), (b), we show the correlation between the valence estimated and human collected ratings in English and Greek, respectively. Specifically, both context and co-occurrence based semantic similarity metrics were tested with the semantic-affective model. When the co-occurrence-based metric is used we observe that the model lacks robustness, and performance drops especially for large number of seeds. However, when RR is used performance is robust and also the correlation curve is quite close to the one observed when context-based similarities are used. Correlation of Greek valence ratings when co-occurrence based semantic similarities are used in combination with RR is much higher than when using them with LSE. Still, it is much lower than the correlation when context-based similarities are used. Language morphology and characteristics may be responsible for the performance differences between the two languages.

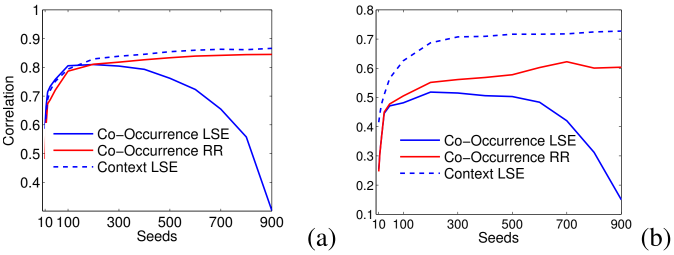

Figure 6: Correlation of automatically estimated and human collected valence ratings for English (a) and Greek (b) using different semantic similarity and weights estimation methods.

6. Shared data

We share two resources. The **Greek affective lexicon**[5] that consists of 1034 Greek words annotated with respect to valence, arousal and dominance. We share the Greek words, the index of the word in ANEW that was translated

[4] This may be attributed to the differences on syntax and morphology that exist between the two languages.

[5] `http://www.telecom.tuc.gr/~epalogiannidi/docs/resources/greek_affective_lexicon.zip`

in Greek and the affective ratings in the range $[-1, 1]$. We also share the automatically estimated **affective lexicon for the Greek language**[6]. This lexicon contains approximately 407K Greek words (selected from the aspell dictionary and the Greek Wikipedia) that are labelled with respect to valence, arousal and dominance. The affective labels have been estimated through the affective model defined by (1) using binary weighting scheme and their range is $[-1, 1]$.

7. Conclusions

We created the first affective lexicon for the Greek language and we collected affective lexica ratings consistent with other languages. The affective lexicon expansion model was evaluated both on English and Greek for the three affective dimensions achieving consistent results. We showed the impact of the semantic similarity metric to the affective ratings estimation with two ways: 1) comparing different types of semantic similarity metrics 2) using different weighting schemes during the semantic model computation. Moreover, we showed that semantic similarity metrics that are not very appropriate for the affective ratings estimation task, can achieve robust performance when they are combined with the appropriate weight learning technique (e.g., co-occurrence-based semantic similarities with ridge regression).

In future work we will study the universality of the presented affective model by experimenting with more languages. Also, we will investigate the creation of lexical resources that consist of larger lexical units, e.g., word pairs.

8. Acknowledgments

This work has been partially funded by the SpeDial project supported by the EU FP7 with grant num. 611396 and the BabyRobot project supported by the EU Horizon 2020 Programme with grant num. 687831. Many thanks to Nikos Malandrakis for sharing his resources and codebase.

9. References

Bechara, A., Damasio, H., and Damasio, A. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, 10(3):295–307.

Bradley, M. and Lang, P. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Bradley, M. and Lang, P. (1999a). Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings, technical report c-1. Technical report, The Center for Research in Psychophysiology, University of Florida.

Bradley, M. and Lang, P. (1999b). The international affective digitized sounds (IADS)[: stimuli, instruction manual and affective ratings]. Technical report, NIMH Center for the Study of Emotion and Attention.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods*, 39(3):510–526.

Chaumartin, F. (2007). Upar7: A knowledge-based system for headline sentiment tagging. In *Proc. of SemEval*, pages 422–425. ACL.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Eilola, T. M. and Havelka, J. (2010). Affective norms for 210 British English and Finnish nouns. *Behavior Research Methods*, 42(1):134–140.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proc. of LREC*, pages 417–422.

Gilet, A., Grühn, D., Studer, J., and Labouvie-Vief, G. (2012). Valence, arousal, and imagery ratings for 835 French attributes by young, middle-aged, and older adults: The French Emotional Evaluation List (FEEL). *European Review of Applied Psychology*, 62(3):173–181.

Gracia, J., Trillo, R., Espinoza, M., and Mena, E. (2006). Querying the web: A multiontology disambiguation method. In *Proc. of International Conference on Web Engineering*, pages 241–248.

Harris, Z. (1954). Distributional Structure. *Word*, 10(23):146–162.

Iosif, E. and Potamianos, A. (2010). Unsupervised Semantic Similarity Computation Between Terms Using Web Documents. *IEEE Transactions on Knowledge and Data Engineering*, 20(11):1637–1647.

Iosif, E. and Potamianos, A. (2015). Similarity computation using semantic networks created from web-harvested data. *Natural Language Engineering*, 21(01):49–79.

Kanske, P. and Kotz, S. (2010). Leipzig affective norms for German: A reliability study. *Behavior research methods*, 42(4):987–991.

Lang, P., Bradley, M., and Cuthbert, B. (1999). International affective picture system (IAPS): Technical manual and affective ratings. Technical report, Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Malandrakis, N., Potamianos, A., Iosif, E., and Narayanan, S. (2011). Kernel Models for Affective Lexicon Creation. In *Proc. of Interspeech*, pages 2977–2980.

Malandrakis, N., Potamianos, A., Iosif, E., and Narayanan, S. (2013). Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392.

Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Montefinese, M., Ambrosini, E., Fairfield, B., and Mammarella, N. (2014). The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior research methods*, 46(3):887–903.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A., De Schryver, M.,

[6] http://www.telecom.tuc.gr/~epalogiannidi/docs/resources/greek_affective_lexicon_automatically_created.zip

De Winne, J., and Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior research methods*, 45(1):169–177.

Palogiannidi, E., Iosif, E., Koutsakis, P., and Potamianos, A. (2015). Valence, Arousal and Dominance Estimation for English, German, Greek, Portuguese and Spanish Lexica using Semantic Models. In *Proc. of Interspeech*, pages 1527–1531.

Redondo, J., Fraga, I., Padron, I., and Comesana, M. (2007). The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, 39(3):600–605.

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., and Frade, C. S. (2012). The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods*, 44(1):256–269.

Stevenson, R., Mikels, J., and James, T. (2007). Characterization of the affective norms for English words by discrete emotional categories. *Behavior Research Methods*, 39(4):1020–1024.

Strapparava, C. and Valitutti, A. (2004). WordNet Affect: an Affective Extension of WordNet. In *LREC*, pages 1083–1086.

Turney, P. and Littman, M. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus, technical report ERC-1094 (NRC 44929). Technical report, National Research Council of Canada.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.