STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXII**, Number 2, 2017
DOI: 10.24193/subbi.2017.2.04

ROEMOLEX - A ROMANIAN EMOTION LEXICON

ANAMARIA BRICIU AND MIHAIELA LUPEA

Abstract. In Natural Language Processing tasks, semantical and lexical resources are of paramount importance for efficient implementations of solutions. However, availability of tools for any other language than English is fairly limited, therefore leaving the field open to improvements and new developments. This paper presents the second version of RoEmoLex, a lexicon containing approximately eleven thousand Romanian words tagged with a series of emotions and two valences. We describe the steps followed in the improvement process of the first version of the resource: the addition of new terms, and the extension of emotional concepts to finer-grained tags.

1. Introduction

Semantical and lexical resources are widely used in natural language processing tasks, either as supports of knowledge-based methods, or as aids for hybrid techniques involving machine learning. Hand-crafted lexicons, exhaustive thesauri or semantic tools (e.g. FrameNet) add an element of language understanding to traditional statistical models, generally improving performance in tasks like sentiment analysis, word sense disambiguation or document summarization.

However, there are relatively few resources designed for languages other than English, which makes it difficult to implement viable solutions for the analysis of non-English content. As far as the Romanian language is concerned, the existence of RoWordNet [15], a large semantic network that encompasses information in the form of synsets (i.e. groups of full synonyms) linked by lexical-semantic relations, marks the central point of Romanian language processing research due to its size, completeness, and information richness. Some

Received by the editors: November 2, 2017.

2010 *Mathematics Subject Classification.* 68T50, 91F20.

1998 *CR Categories and Descriptors.* I.2.7. [**Computing Methodologies**]: Artificial Intelligence - *Natural Language Processing*; H.3.1. [**Information Systems**]: Information Storage and Retrieval - *Content Analysis and Indexing*.

Key words and phrases. natural language processing, linguistic processing, emotion analysis, lexicon.

aligned corpora for machine translation, as well as a number of small-scale emotion lexicons have also been developed, which is encouraging, but leaves the field open to exploration and expansion. This was our motivation for taking a translated version of the lexicon proposed by Mohammad and Turney in [7], processing the existing data and bringing a series of modifications and additions to it in order to create a dependable Romanian emotion lexicon.

This work discusses the continuation of the work presented in [6], where a series of processing and structuring steps applied to the original, automatically translated lexicon were described. Enhancements to this processed first version in the form of new term additions brought the lexicon to its current form, consisting of approximately 11000 words tagged with affect information, specifically Plutchik's [9] eight basic emotions (*Anger, Anticipation, Surprise, Joy, Trust, Fear, Sadness, Disgust*) and polarity tags (*Positivity, Negativity*), and a series of derived emotions formed through the combination of two basic emotions.

The paper is structured as follows: in Section 2, related work is presented, while in Section 3 we describe RoEmoLex, following the development process from the initial revision of the English translation to this new round of enhancements. Section 4 is reserved for conclusions and proposals for future work.

2. RELATED WORK

As relevant to the current article, there are two directions of research that need to be discussed: the available lexical and semantic resources for Romanian text, and the particularities of existent affect lexicons.

First, to address the former issue, RoWordNet [3], [15] must be introduced. This is the Romanian version of the Princeton WordNet, and it is a lexical database which allows the modeling of linguistic knowledge with regards to aspects of the real world. There is sentiment information in form of three polarity scores: *positive, objective, negative*, domain data (every term is associated with a conceptual type or domain - see SUMO, Section 3.4.), and a series of relationships that structure the data in an intuitive manner (e.g. *is a*, or *part of* relations between lexical units). Beyond RoWordNet, there are a few other resources that offer the possibility of modular integration in a more complex system (e.g dexonline[1]), but none with the applicability and richness in information that RoWordNet offers.

Secondly, in terms of resources for affect detection and recognition in text, for English, a widely used tool is LIWC [8], [13], a text analysis application with a comprehensive dictionary in which terms are marked as representative

[1]http://pydex.lemonsoftware.eu/

of a series of categories and subcategories, including Affective Processes. Other resources are DepecheMood [10] and ANEW [2]. The first, DepecheMood, is a lexicon created in an entirely automated manner by using affective annotation provided implicitly by readers of articles on a news site. The second resource, ANEW, provides a set of normative emotional ratings for a series of words. It is important to note that, in contrast to our lexicon, in the case of ANEW, emotion is viewed as a space in three dimensions [5], pleasure, arousal and dominance, rather than as represented by fixed atomic units in the form of basic emotions from which other emotions are formed.

More relevant to the current work are EmoLex [7], which will be described in detail in Section 3.1, as a starting point of our work, and the multilingual WordNet-Affect project described in [1]. The latter represents an aligned English-Russian-Romanian affect lexicon, considering six emotions (*Anger, Disgust, Fear, Joy, Sadness, Surprise*) and the corresponding synsets present in WordNet-Affect. It is a carefully crafted resource, with valuable emotional information that we integrated in RoEmoLex. Details about this process can be found in Section 3.3.

3. ROEMOLEX – ROMANIAN EMOTION LEXICON

The starting point of this research is a lexicon proposed by Mohammad and Turney in 2010, the NRC Word-Emotion Association Lexicon or EmoLex [7], consisting of a series of English words and their associations with Plutchik's eight emotions and two polarity tags. We used the automatically translated Romanian version of this lexicon, with the proposed improvements consisting in removing non-emotive words (i.e. words with an overall emotion and polarity score equal to 0), eliminating duplicate lines, and adding RoWordNet information for each word. Finally, we added new terms to the lexicon, aiming to build a more comprehensive resource.

3.1. **Original lexicon.** The NRC Word-Emotion Association Lexicon [7], or EmoLex, was created by crowdsourcing the task through Amazon's Mechanical Turk. Hundreds of online annotators were given a series of words and asked which of eight emotions (*Anger, Anticipation, Surprise, Joy, Trust, Fear, Sadness, Disgust*) and two valences (*Positivity, Negativity*) the term expresses, and to what extent it does (*none, weak, moderate, strong*). Then, based on the annotators' feedback, the authors considered the number of assignments into each intensity category for a word-emotion pair. If there were more appearances of *no*, or *weak* intensity, it was considered that the word did not evoke the emotion in question. Alternatively, if the majority of responses said *moderate* or *strong* intensity, the word was tagged with the specified emotion. The resulting lexicon contained 2000 words tagged with eight emotions and

sentiment valences, but was later expanded to 14182 words and 25000 senses for English.

Through automatic translation, EmoLex has been made available for over twenty languages, including Romanian. English words are converted using Google Translate, which maintains the emotional correspondence to a relatively high extent [14], but, as is to be expected, also introduces noisy data, for which a series of post-processing steps are required.

3.2. **Post-processing of translated data.** Initially, the Romanian version of the lexicon consisted of 8581 lines (i.e. words). However, many of these were null lines, i.e. terms expressing no sentiment valence or emotion. These were *non-evocative* terms, which we considered to be of little use in further emotion analysis applications. Therefore, we discarded them along with any duplicate entries. Note that the notion of duplicate entry is defined as the same term having two entries in the lexicon with all polarity and emotion tags equal. A series of identical terms have remained in the lexicon, as they have different scores, and account for polysemantic words.

After these steps, we were left with 3989 unique lines, which we aligned with RoWordNet by means of assigning each term to the corresponding synset (where possible). Using this synset information, we introduced part of speech, positive, negative and objective SentiWordNet scores, and SUMO categories for each term. More details regarding these processing steps, and about the first version of RoEmoLex, respectively, can be found in [6].

In the current paper, we introduce a new structural addition in the form of affective annotation for primary, secondary and tertiary emotions, as theorized by Plutchik in the form of dyads.

In his general psychoevolutionary theory of emotion, Plutchik identified eight basic, biologically primitive emotions: *Anger*, *Anticipation*, *Disgust*, *Fear*, *Joy*, *Sadness*, *Surprise*, *Trust* and formulated ten postulates to characterize his theory. Of these, of relevance to the current work is the sixth, which states that beyond these eight, all other emotions are mixed or derivative states; that is, they occur as combinations, mixtures, or compounds of the primary emotions [9]. Considering this postulate, we included tags for a series of secondary emotions defined by the author as results of meaningful combinations. Utilizing the Wheel Of Emotions shown in Figure 1 as an expressive visual representation of Plutchik's theory, the following combinations of two basic emotions can be identified:

- **Primary dyads** (emotions combined are one petal apart): *Optimism* (Anticipation and Joy), *Love* (Joy and Trust), *Submission* (Trust and Fear), *Awe* (Fear and Surprise), *Disapproval* (Surprise

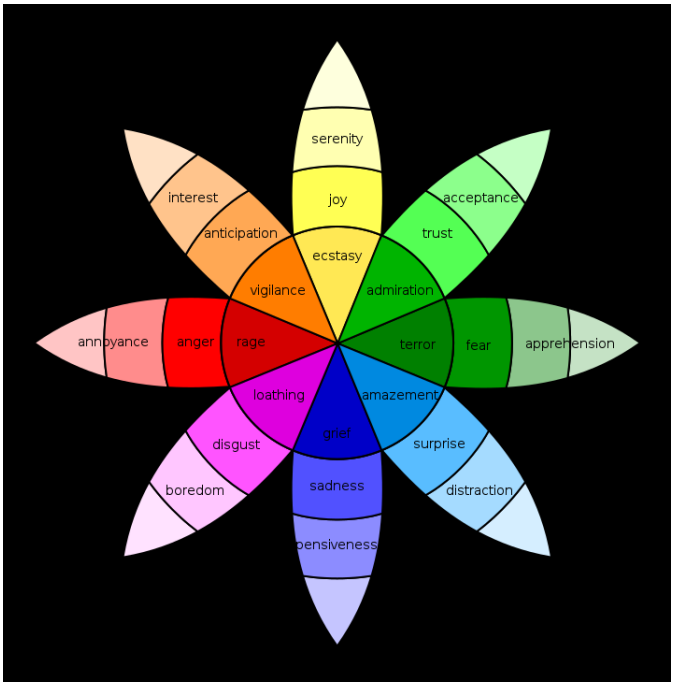

FIGURE 1. Plutchik's Wheel Of Emotions

and Sadness), *Remorse* (Sadness and Disgust), *Contempt* (Disgust and Anger), *Aggressiveness* (Anger and Anticipation)

- **Secondary dyads** (emotions combined are two petals apart): *Hope* (Anticipation and Trust), *Guilt* (Joy and Fear), *Curiosity* (Trust and Surprise), *Despair* (Fear and Sadness), *Disbelief* (Surprise and Disgust), *Envy* (Sadness and Anger), *Cynicism* (Disgust and Anticipation), *Pride* (Anger and Joy)
- **Tertiary diads** (emotions combined are three petals apart): *Anxiety* (Anticipation and Fear), *Delight* (Joy and Surprise), *Sentimentality* (Trust and Sadness), *Shame* (Fear and Disgust), *Outrage* (Surprise and Anger), *Pessimism* (Sadness and Anticipation), *Morbidness* (Disgust and Joy), *Dominance* (Anger and Trust)

Therefore, for every word having both affective tags from a pair, we considered the term to express the emotion resulted from their combination, and tagged them in our lexicon accordingly. Table 1 shows a series of representative words for a few categories.

Emotion	Terms
Optimism	credință/**faith** căsătorie/**marriage**
Aggressiveness	ghilotină/**guillotine** duel/**duel**
Hope	a aspira (la)/**to aspire** determinat/**determined** obiectiv/**objective**
Despair	demoralizat/**demoralized** chin/**anguish** cimitir/**cemetery**
Anxiety	anxietate/**anxiety** diagnostic/**diagnostic** vigilență/**vigilence**
Pessimism	condamnare/**condemnation** îngrozitor/**awful** moarte/**death**

TABLE 1. Examples of words associated with derived emotions

3.3. **Lexicon enrichment.** A final step in the process of improving RoEmoLex was the addition of new words. The expansion of the lexicon was done in two steps: the addition of synonyms of existent entries from RoWordNet, and integration of the data present in the similar resource *Russian-Romanian WordNet-Affect*.

3.3.1. *Lexicon enrichment through addition of RoWordNet synonyms.* The first proposed enrichment method was the addition of synonyms of existent entries. We considered these new terms as having the same emotional and polarity content as the original terms, and we thus assigned them the same emotion and valence tags as the original words. This was done with the help of RoWordNet, taking as *synonyms* all words in the same synset with the entry present in RoEmoLex. After this step, we obtained an additional 6455 entries.

Despite the fact that the lexicon almost tripled in size, the part of speech hierarchy was preserved, as can be seen in Table 2, with nouns and adjectives still accounting for the majority, but with the percent of verbs and idioms each rising about 4 percent. This is due to the fact that verbs represented the terms with most synonyms per entry (e.g. a vorbi - **to talk**: a discuta - **to discuss**, a grăi - **to speak**, a conversa - **to converse**, a dialoga; a sublinia - **to emphasize**: a accentua - **to accentuate**, a reliefa, a puncta), and quite a

Part of Speech	**% of lexicon in initial lexicon**	**% of lexicon after 1st round of enrichment**	**% of lexicon after 2nd round of enrichment**
	3989 words	10444 words	11051
Noun	57.48% — 2293 words	53.27% — 5563 words	51.841%— 5729 words
Adjective	23.41% — 933 words	20.22% — 2112 words	20.76% — 2295 words
Verb	12.86% — 513 words	16.42% — 1715 words	16.40% — 1813 words
Idiom	3.50% — 140 words	7.02% — 734 words	7.94% — 878 words
Adverb	2.65% — 106 words	3.03% — 316 words	3.00%— 332 words
Interjection	0.1% — 4 words	0.04% — 4 words	0.03%— 4 words

TABLE 2. Part of Speech Distribution

few of the synonyms added were, in fact, phrases and idioms (e.g.start - **start**: semnal de start - **start signal**) as opposed to terms with the same part of speech as the original word.

3.3.2. *Lexicon enrichment through integration of data from Romanian-Russian WordNet-Affect.* Another way to enrich the lexicon was the addition of terms from the Russian and Romanian WordNet-Affect [1]. This resource consists of six lists of representative synsets organized according to Ekman's [4] six emotions: *Anger, Disgust, Fear, Joy, Sadness, Surprise.* The motivation in integrating the Russian-Romanian WordNet-Affect data in our lexicon was the improvement of RoEmoLex through the addition of carefully annotated data with the starting point in WordNet-Affect [12], a dependable resource for affective computing.

	# of synsets	**# of Romanian words**
Anger	117	330
Disgust	17	60
Fear	80	248
Joy	209	641
Sadness	98	364
Surprise	27	87

TABLE 3. Russian-Romanian WordNet-Affect Data

The development of the Russian-Romanian WordNet-Affect started from a set of English synsets annotated for each of the six emotions and made available for the SemEval-2007 task "Affective Text" [11]. The authors then proceeded with automatic translation of the synsets from English, manually correcting any inconsistencies and checking the data, and furthermore adding any relevant synonyms of the generated Romanian terms. Statistics concerning

the final form of the Russian-Romanian WordNet-Affect can be viewed in Table 3.

Emotion	**% of common words**
Anger	46.86%
Disgust	39.53%
Fear	42.85%
Joy	53.125%
Sadness	38.01%
Surprise	60.93%

TABLE 4. Percent of common data between RoEmoLex and Russian-Romanian WordNet-Affect

We compared this data with RoEmoLex content, hoping to discover sets of new words to add to our lexicon. As it can be seen in Table 4, the percent of common data varies little among emotions, which points to a balanced understanding of emotion in RoEmoLex. *Surprise* is the emotion with the most common words, while *Sadness* is the one with the fewest. It is interesting to note that a significant portion of the terms that were not in RoEmoLex, irrespective of emotion, were idioms and expressions (e.g. lipsit de veselie - **devoid of joy**, din nefericire - **unfortunately**, cu părere de rău - **with regret** for *Sadness*, lua pe neașteptate - **to catch someone off guard**, pune în încurcătură - **to discomfit** for Surprise). Another reason for the relatively high discrepancy is that the Russian-Romanian WordNet-Affect accounts for multiple spellings of a word. For example, in the multilingual resource, *a mâhni* - ***to dishearten*** can be found in both this form and *a mîhni*, a more infrequent spelling, typically encountered in older texts, while in RoEmoLex only the first form is present. We chose to include such terms in the lexicon, but, where existent, the emotional and polarity content of the term with the more common spelling are duplicated for the word with the less frequent form.

As for the degree of annotation agreement between the two resources, Table 5 shows that *Disgust*, *Fear* and *Sadness* are the emotions most consistently tagged, with more than 70% of the common words having the same emotion tag in both resources. For terms with conflicting tags, we did a manual verification and validation of the annotation in our lexicon.

Going back to Table 2, it can be seen that this second round of additions brought an increase in the number of adjectives and idioms in the RoEmoLex, with noun and verb percentages decreasing slightly. This is owed to the data structure of the Russian-Romanian WordNet-Affect, with a large amount of

Emotion	% of words with matching tags
Anger	69.29%
Disgust	82.35%
Fear	76%
Joy	56.37%
Sadness	77.17%
Surprise	48.71%

TABLE 5. Matching tags in Russian-Romanian WordNet-Affect and RoEmoLex

adjective synsets present, and the manual inclusion of translations (inimă grea - **heavy heart**, îndoială de sine - **self-doubt**).

In total, we acquired 607 new lexicon entries, with *Anger, Disgust, Fear, Joy, Surprise* and *Sadness* tags from the Russian-Romanian WordNet-Affect lexicon, and *Positivity, Negativity, Anticipation* and *Trust* manually added. This put the final number of terms in RoEmoLex at just over 11000.

3.4. **Lexicon samples.** In this section, we include a series of representative RoEmoLex entries for a better understanding of the structure and data format within the lexicon. We note that the presented table limits itself to introducing the tags for the eight basic emotions for reasons of space, and does not introduce the derived emotions presented in Section 3.2. For examples of terms corresponding to these derived tags, please refer to Table 1.

In Table 6, the **P** and **N** table headers refer to the *Positivity* and *Negativity* tag, respectively. **POS** stands for *Part of Speech* information. **SUMO** is an abbreviation for Suggested Upper Merged Ontology, and is a field mapped from RoWordNet designed to offer a semantic context to the term in question.

4. CONCLUSIONS AND FURTHER WORK

In this paper, we presented the development process of a new version of RoEmoLex, namely the enhancements we proposed in order to create a comprehensive emotion analysis resource for Romanian texts. We described the original lexicon, EmoLex, from which RoEmoLex was translated, briefly presenting the initial round of data processing. Finally, we outlined the mechanics of two rounds of lexicon enrichment, presenting a series of statistics and some sample entries to illustrate aspects of the current form of the lexicon.

We state that RoEmoLex can constitute a viable starting point for emotion analysis in Romanian texts, but note that there is still work that can be done in terms of improving the resource. For example, the inclusion of fuzzy logics

Word	**POS**	**P**	**N**	*Anger*	*Anticipation*	*Disgust*	*Fear*	*Joy*	*Sadness*	*Surprise*	*Trust*	**SUMO**
ascuns **concealed**	Adjective	0	1	0	1	0	1	0	0	1	0	Covering
ascuns **hidden**	Adjective	0	1	0	0	0	0	0	0	0	0	Subjective Assessment Attribute
demonic **demonic**	Adjective	0	1	1	0	1	1	0	1	0	0	Subjective Assessment Attribute
vacanță **holiday**	Noun	1	0	0	1	0	0	1	0	0	0	Vacationing
autenticitate **authenticity**	Noun	1	0	0	0	0	0	0	0	0	1	True
nedreptate **injustice**	Noun	0	1	1	0	1	0	0	1	0	0	Normative Attribute
bătaie de joc **mockery**	Idiom	0	1	1	0	1	0	0	0	0	0	Expressing
liniște sufletească **hearts-ease**	Idiom	1	0	0	0	0	0	1	0	0	0	EmotionalState
abia **barely**	Adverb	0	1	0	0	0	0	0	1	0	0	Subjective Assessment Attribute
vai **oh!**	Interjection	0	1	0	0	1	1	0	1	0	0	-
a îndrăzni **to dare**	Verb	1	0	0	0	0	0	0	0	0	1	Subjective Assessment Attribute
a sugruma **to strangle**	Verb	0	1	1	0	1	1	0	1	0	0	ViolentContest

TABLE 6. RoEmoLex - Sample entries

for evaluating the degree of membership of a word to all the emotion classes would be an approach that modeled human understanding of emotion and its expression better, and it is a track for further work that we will investigate.

The importance of continuous refinement and improvement of such a resource lies in the many interesting applications that can be developed provided a dependable emotion analysis module, from the simple study of emotional content and its evolution in a given text to integrating the module into more complex systems (e.g. customized interaction in tutoring systems, mental health monitoring applications).

References

[1] Victoria Bobicev, Victoria Maxim, Tatiana Prodan, Natalia Burciu, and Victoria Angheluş. Emotions in words: Developing a multilingual wordnet-affect. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 375–384. Springer, 2010.

[2] Margaret M. Bradley, Peter J. Lang, Margaret M. Bradley, and Peter J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings, 1999.

[3] Ştefan Daniel Dumitrescu. Rowordnetlib - the first api for the romanian wordnet. *Proceedings of the Romanian Academy, Series A*, 16(1):87–94, 2015.

[4] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[5] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*, chapter 18. Lexicons for Sentiment and Affect Extraction.

[6] Mihaiela Lupea and Anamaria Briciu. Formal concept analysis of a romanian emotion lexicon. In *Proceedings of the 2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 111–118, 2017.

[7] Saif M Mohammad and Peter D Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34. Association for Computational Linguistics, 2010.

[8] James W Pennebaker, Roger J Booth, and Martha E Francis. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 2007.

[9] Robert Plutchik. A general psychoevolutionary theory of emotion. *Emotion: Theory, Research, and Experience*, 1:3–33, 1980.

[10] Jacopo Staiano and Marco Guerini. Depeche mood: a lexicon for emotion analysis from crowd annotated news. pages 427–433, 2014.

[11] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 1556–1560, New York, NY, USA, 2008.

[12] Carlo Strapparava and Alessandro Valitutti. Wordnet-affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086, 2004.

[13] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[14] Vaibhav Tripathi, Aditya Joshi, and Pushpak Bhattacharyya. Emotion analysis from text: A survey. *Center for Indian Language Technology Surveys*, 2016.
[15] Dan Tufiș, Eduard Barbu, Verginica Barbu Mititelu, Radu Ion, and Luigi Bozianu. The romanian wordnet. *Romanian Journal of Information Science and Technology*, 7(1-2):107–124, 01 2004.

Faculty of Mathematics and Computer Science, Babeş–Bolyai University, Cluj-Napoca, Romania
E-mail address: baic1326@scs.ubbcluj.ro
E-mail address: lupea@cs.ubbcluj.ro