Emotions in words: developing a multilingual WordNet-Affect

Victoria Bobicev, Victoria Maxim, Tatiana Prodan,
Natalia Burciu, Victoria Angheluş

Technical University of Moldova,
168, Stefan cel Mare bd., Chisinau, Republic of Moldova
vika@rol.md, maxivica@yahoo.com, tatiana.ursulenco@gmail.com,
natusicb@yahoo.com, lazu_vic@yahoo.com

Abstract. In this paper we describe the process of Russian and Romanian WordNet-Affect creation. WordNet-Affect is a lexical resource created on the basis of the Princeton WordNet which contains information about the emotions that the words convey. It is organized in six basic emotions: *anger, disgust, fear, joy, sadness, surprise.*
We translated the WordNet-Affect synsets into Russian and Romanian and created an aligned English – Romanian – Russian lexical resource. The resource is freely available for research purposes.

Key words: sentiment analysis, lexical representation of affects, multilingual lexical resources

1 Introduction

Currently, the researchers in the field of the natural language processing drew their attention to the fact that texts contain not only objective information but also the emotional attitude of the author towards this information.

These days, the booming growth of Web 2.0 technologies allows every user to participate actively in web content creation (blogs, social networks, chats). The volumes of texts with emotionally-rich content grow in geometrical progression. This makes the subjective analysis of texts especially topical.

So far, the sentiment analysis and studies of the word affect were concentrated on English. The example is the SemEval-2007 task of Affective Text classification [6]. Most lexical resources have been created for English, as well. For example, SentiWordNet is a lexical resource for opinion mining which assigns to each synset of the WordNet three sentiment scores: positiveness, negativity, objectivity [11].

Recently, most of the Internet use growth was supported by non-native English speakers: starting 2000, for non-English speaking regions, the growth has surpassed 3,000% to compare with 342 % of the over-all growth.[1]

[1] http://www.internetworldstats.com/stats.htm

Consequently, the amount of the text data written in languages other than English rapidly grows [3]. This raise increases the demand for automatic text analysis tools and linguistic resources for languages other than English. The tool development has progressed for Western European (French, German) and Asian (Japanese, Chinese, Arabic) [4].

At the moment, resources for Eastern European languages are not easily available. In order to fill this gap, we developed a linguistic resource, starting from WordNet-Affect, through the translation in Russian and Romanian languages, editing of the translated synsets and aligning them to English source.

2 WordNet-Affect

WordNet-Affect[2] is a well-used lexical resource which contains information about the emotions that the words convey. Compared with the complete WordNet, WordNet-Affect is a small lexical resource but valuable for its affective annotation.

WordNet-Affect [7] was created starting from WordNet DOMAINS [12]. WordNet-Affect produces an additional hierarchy of the *affective domain labels*, independent from the domain hierarchy, wherewith the synsets that represent affective concepts are annotated. The "affective words" are considered to be words that have "emotional connotation" [13]. There are words that not only describe directly some emotions (for example, *joy*, *sad* or *scare*) but also are related to emotions like words describing *mental state*s, *physical* or *bodily states*, *personality traits*, *behaviours*, *attitudes*, and *feelings* (such as *pleasure* or *pain*).

The collection of the WordNet-Affect synsets used in our work was provided as a resource for the SemEval-2007 "Affective Text". This task was focused on text annotation by affective tags [6]. There is not the whole WordNet-Affect but a part of it being more fine-grain re-annotated using six emotional category labels: *joy*, *fear*, *anger*, *sadness*, *disgust*, *surprise* [8]. This choice of the six emotions comes from psychological research into human non-verbally expressed emotions [5].

```
a#01943022 awed awestruck awestricken in_awe_of
```

Fig. 1. A synset of WordNet-Affect.

The data is described in Table 1. The whole data is provided in six files named by the six emotions. Each file contains a list of synsets; one synset per one line. An example of a synset is shown in figure 1.

The first letter in the line indicates the part of speech; it is followed by the number of the synset and then all synset words are listed. The representation was simple and easy for further processing. There were a large number of word combinations, collocations and idioms. One of them can be seen in the example. These parts of synsets presented a problem during translation.

[2] For research purposes, WordNet-Affect is available upon request at http://wndomains.itc.it

Table 1. Data sets of affective words.

Classes	#synsets	%synsets	#words	%words
anger	128	21.0	318	20.7
disgust	20	3.3	72	4.7
fear	83	13.5	208	13.5
joy	228	37.2	539	35.1
sadness	124	20.3	309	20.1
surprise	29	4.7	90	5.9
Total	612	100.0	1536	100.0

3 Development of Romanian and Russian WordNet

Romanian WordNet has been created by the Alexandru Ioan Cuza University of Iaşi during European project BalkaNet [9]. After the BalkaNet project ended, the Research Institute for Artificial Intelligence, at the Romanian Academy continued to update the Romanian WordNet and currently it contains 33151 noun synsets, 8929 verb synsets, 851 adjective synsets and 834 adverb synsets [10]. It can be accessed through the online MultiWordNet[3] interface where WordNets for several languages are aligned to Princeton WordNet.

First of all in our work we checked WordNet-Affect synsets using online interface of MultiWordNet. We just copied to our set all the synsets which already are in the Romanian WordNet and did not process these synset further. As result, 166 synsets were found in the Romanian WordNet, the majority of them being available for nouns and verbs. The adjectives and adverbs are less represented. The statistics of the already existing Romanian synsets is presented in table 2.

Table 2. Data sets of the already existing Romanian WordNet synsets.

Classes	# synsets in WordNet-Affect	# synsets from Romanian WordNet	% synsets from Romanian WordNet
anger	116	35	30.1
disgust	17	7	41.1
fear	76	25	32.8
joy	210	63	30.0
sadness	97	24	24.7
surprise	26	12	46.1
Total	542	166	30.6

There is completely different situation for Russian. Several attempts were taken to develop the Russian WordNet. RussNet is a project of computer thesaurus of Russian vocabulary [1]. An alternative project of Russian version of WordNet is Russian WordNet [2]. Both projects are non-commercial. Two commercial projects aimed to develop WordNets in Russian: RuThes is informational thesaurus used in UIS

[3] http://multiwordnet.itc.it/english/home.php

RUSSIA[4] and Russian WordNet project by the Novosoft company group[5]. Unfortunately, little information is available and even less freely available resources.

4 Development of Romanian and Russian WordNet-Affect

In order to create the two data sets, we applied the three-step approach: (1) automatic translation; (2) removing irrelevant translations; (3) generating Romanian and Russian synsets**.**

4.1 Automatic Translation

The translation was done automatically using bilingual dictionaries. We used Electronic Romanian-English Dictionary ROMEN from PRIMASOFT[6]. It consists of English-Romanian, Romanian-English, English-Russian and Russian-English parts, each containing more than 200 000 entries. In our work we used only the parts with English as a source language. There were a number of word combinations, collocations and idioms in the dictionaries which we have used in target languages. For the automatic translation, the dictionary was organized in a list of source words followed by the target translations. An example of the dictionary entry is presented in figure 2.

```
Joy
Dicţionar general:
noun: bucurie; confort; fericire; plăcere; tihnă;
veselie; voioşie;
verb: a bucura; a înveseli;
```

Fig. 2. An example of dictionary entry**.**

At this stage, our goal was to obtain as many affective words as possible for the analysis. For this purpose we translated every word in the WordNet-Affect synsets. We decided to exclude from the English synsets all the word combinations, collocations and idioms as they could not be translated automatically. The figure 3 presents an example of the translated synset obtained after this step. As it is seen in the example, for the Romanian translation, we also obtained word combinations which were in the dictionary: "`cuprins de veneraţie`", "`cuprins de teamă`".

Some synset elements were not translated. These can be divided into four groups. (1) Word combinations, collocations and idioms which we intentionally removed from English synsets before the translation. (2) Spelling variations of the same word;

4 http://www.cir.ru

5 http://research-and-development.novosoft-us.com

6 http://www.primasoft.biz/romen_eng.php

for example, “jubilance”, “jubilancy” – the first word was translated, the second one was not found in the dictionary. (3) Words which were formed using suffixes like “ness”, “less”, “ful” (for example “heartlessness”); these are unlikely to appear in dictionaries as well as adverbs formed using suffix “ly”. (4) Words which were not translated because of the limitedness of our dictionary. While WordNet can reasonably be mentioned as one of the largest English dictionary, our bilingual dictionary is fairly modest. Table 3 shows the percentage of words which were not translated. Average percentage of not translated words was 21%.

Table 3. Number and percentage of not translated words.

Classes	# of English words	# of translated words	# of not translated words	% of not translated words
anger	318	248	70	22.0
disgust	72	60	13	18.0
fear	208	162	47	22.5
joy	539	420	119	22.0
sadness	309	246	63	20,5
surprise	90	72	18	21.0
Total	**1536**	**1208**	**330**	**21.0**

The second group of words did not present a problem but the first, third and the fourth ones had to be translated manually. It was done during the third step.

```
01943022 a:
awed =     speriat
awestruck =
           cuprins de veneraţie
           cuprins de teamă
awestricken = înspăimântat
```

Fig. 3. An example of English synset translation.

4.2 Removing Irrelevant Translations

Many words in English synsets had several meanings. It was obvious that the automatic translation provided all possible translations for all their senses. We were interested in only one translation which was relevant to the synset meaning. The relevant translation was selected manually. We removed all translations whose meaning was not related to the emotion. For example, the word “taste” in the synset with the meaning “preference” had several meanings but only the last one in the list of possible translations was related to the synset common meaning. The example is demonstrated in figure 4. Thereby, we removed all translations except the last one.

As we translated every word separately, we obtained a lot of duplicates which had to be removed as well. We also watched over the part-of-speech correspondence. In many cases, it was rather difficult, especially for the already mentioned nouns formed using suffixes, for example, “plaintiveness” or “uncheerfulness”.

```
05573914 n:
 preference =
          preferință
 penchant =
          înclinație
          slabiciune
 predilection =
          predilecție
 taste =
          a avea gust
          a gusta, a cunoaște
          a gusta; a degusta (un aliment)
          degustare
          fărâmă, bucățică, îmbucătură (de)
          gust
          înclinație, preferință
```

Fig. 4. An example of one synset translation**.**

4.3 Generating Romanian and Russian Synsets

All words in the synset represent one concept, one meaning. The aim of the third step was to find the adequate translation of exactly this meaning. At this step, we firstly had to attach English glosses to every synset. It made clearer the meaning of the synsets for translators. After the glosses were added to the synsets, the whole set was given to three translators which worked independently. Their task was twofold: (1) to remove the translations which, from their point of view, were irrelevant to the synset meaning described by the gloss; (2) to add as many relevant synonyms as possible to the Romanian and Russian synsets. Thereby, their task was to verify the equivalence of the English, Romanian and Russian synset meanings. They also had to translate the words which remained without translation from the first step. For translation they mostly used online dictionaries.

Bilingual English-Romanian dictionaries used:

- http://hallo.ro,
- http://dictionar.netflash.ro,
- http://www.ectaco.co.uk/English-Romanian-Dictionary;

Romanian thesaurus: http://dexonline.ro/.

Bilingual Russian dictionaries used:

- http://en.bab.la,
- http://dictionary.babylon.com,
- http://russianlessons.net/dictionary/dictionary.php;

Russian thesauri:

- http://slovo.freecopy.ru/,
- http://slovari.yandex.ru/dict/ushakov.

This step was the most laborious and difficult. Many English synsets have quite similar meaning with some nuances. In some cases, the synsets contained obsolete words, which were not found in the dictionary. As it was mentioned above, we tended to avoid word combinations, collocations and idioms. However, in some cases, the exact sense of the English synset could be represented only by some combination of Romanian or Russian words. In some cases even the English synset was presented by word combination. For example, `n#05591681 stage_fright.` Another example contains a German word: `n#05600844 world-weariness Weltschmerz`. In such cases, we did not obtain the proper translation. In some cases, several English synsets have got the same Romanian or Russian words as translations because we could not reflect the nuances of the source language senses in the target languages.

Referring to the problem with suffixes, for instance, the words "`weepiness`", "`dysphoria`", "`plaintiveness`", "`mournfulness`", "`ruthulness`" can hardly be found in dictionaries either in Romanian or English. In order to solve this problem, we searched the lemmas of the mentioned words in the available dictionaries. In this way, we could find the meaning of the words and, by adding the necessary affixes, the Romanian and Russian equivalents were created. For example, to find the adequate translation for the word "`mournfulness`", we searched in the dictionary the word "`mournful`". The result for Romanian is "`îndoliat`" and for Russian "`траурный`". As the word "`mournfulness`" is a noun, we transformed the obtained adjectives into nouns. Likewise, the Romanian equivalent is "`doliu`" and the Russian one is "`траур`".

However, most difficulties appeared with the alignment of adjectives. For example, for the emotional label "`sadness`", many of adjectival synsets translated in Russian contain the words "`грустный`" and "`печальный`". For different adjectival synsets we obtain quite similar translations as well.

4.4 Inter-Translator Agreement

In our case, we could not use standard metrics for inter-translator agreement as we had the output as a set of synonyms. Therefore the agreement was calculated as follows. If **A** was a set of words selected by the first translator for the synset and **B** was a set of words selected by the second translator for the same synset, inter-annotator agreement **IntAgr** was equal to quotient of number of words in **A** and **B** intersection divided by number of words in **A** and **B** union:

$$\mathbf{IntAgr} = \mathbf{A} \cap \mathbf{B} / \mathbf{A} \cup \mathbf{B} . \quad (1)$$

For example, if one translator formed a synset from three words w_k, w_l and w_m and the second translator formed this synset from four words w_k, w_l, w_m and w_n and the

first three words are the same, then A=(w_k w_l w_m), B=(w_k w_l w_m w_n), $A \cap B$ = (w_k w_l w_m), $\mathbf{A} \cup \mathbf{B}$ = (w_k w_l w_m w_n), number of words in **A** and **B** intersection would be 3, number of words in **A** and **B** union would be 4 and therefore inter-translator agreement would be 3/4 = 0.75.

For example the synset "a#01195320 friendly" was translated by the first translator as "prietenesc prietenos amical", by the second translator as "amical prietenos binevoitor", and by the third as "prietenesc prietenos binevoitor". For the first and the second translators the intersection of translations was two words: "prietenos amical" and translation's union were four words "prietenesc prietenos amical binevoitor". Inter-translator agreement in this case was 2/4=0.5. For the second and third translators the intersection of translations was two words: "prietenos binevoitor" and translation's union were four words "prietenesc prietenos amical binevoitor". Therefore, the agreement is the same: 0.5. For the first and third translators inter-translator agreement is again the same: 0.5. All three translators shared only one word "prietenos" and union of translations consisted from four words. Thus, the agreement was 1/4=0.25.

Table 4 presents the average values of the inter-translator agreement. The three translators are presented as T1, T2 and T3.

Table 4. Inter-translator agreement.

Pair of translators	Inter-translator agreement
Russian data	
T1 – T2	0.57
T2 – T3	0.61
T1 – T3	0.59
All	0.29
Romanian data	
T1 – T2	0.58
T2 – T3	0.57
T1 – T3	0.67
All	0.32

As it is seen in the table, the agreement is low. There were some synsets with agreement equal to one as for example in the synset "a#00863650 euphoriant", all three translators translated it as "euforizant". However, for the majority of the synsets, the translators provided more different translations but not many of these translations were common for all translators. In some translated synsets, there was not any single word shared between all three translators. For example, for the synset "a#00670851 gladdened exhilarated", the three translations were "bucurat înveselit înviorat bine_dsipus", "bucuros vesel voios încântat bine_dispus" and "bucurat voios bucuros înveselit". There was no common word for all three translations.

Thus, we decided to form the synsets from words which were in at least two variants among the three translations. In such way, we formed the final synsets. For example, the synset "a#01195320 friendly" was translated as "prietenesc prietenos amical binevoitor" because all these words appeared at least twice in translations. The synset "a#00670851 gladdened exhilarated" was translated as "bucurat înveselit bine_dsipus bucuros voios".

Table 5 contains data on the final number of words in translations for each of the six WordNet-Affect emotions.

Table 5. Data sets of affective words for Russian and Romanian.

Classes	#synsets	# Russian words	# Romanian words
anger	117	393	330
disgust	17	73	60
fear	80	327	248
joy	209	765	641
sadness	98	437	364
surprise	27	129	87
Total	548	2199	1869

It should be mentioned that in the source WordNet-Affect set there were some duplicated synsets. We removed all these repetitions and the number of synsets in our source is smaller. Besides, there were small differences in WordNet-Affect, MultiWordNet and online version of Wordnet because the MultiWordNet uses version 2.0 of WordNet and online version of WordNet is 3.0. It is seen that, despite of smaller number of synsets, the number of words in Romanian and Russian set is bigger than in English. This is due to our tendency to collect in our resource as many words as possible. We aim to use it in statistical methods of emotion recognition in text.

5 Conclusion and Future Work

This paper describes the process of the Russian and Romanian WordNet-Affect creation. WordNet-Affect is a lexical resource created on the basis of Princeton WordNet, which contains information about the emotions that the words convey. It is organized in six basic emotions: *anger, disgust, fear, joy, sadness, surprise*. WordNet-Affect is a small lexical resource but valuable for its affective annotation.

We translated the WordNet-Affect synsets into Russian and Romanian and, afterwards, created English – Romanian – Russian aligned WordNet-Affect. The resource can be used for the automatic recognition of emotions and affects in text. It is freely available for research purposes at http://lilu.fcim.utm.md.

The resource is still under development. The first version based on WordNet-Affect was released in August 2009; the second one, released in October 2009, is already aligned with the Romanian WordNet. Further, we are going to refine the Russian part and to create 'bag-of-words' resource for immediate use in emotion and

affect recognition tasks. The resource has already been used in [14] and it is only one among many possible uses of the word sets.

References

1. Azarova, I., Mitrofanova, O., Sinopalnikova, A., Yavorskaya, M., Oparin I.: Russnet: Building a lexical database for the russian language. In: Workshop on Wordnet Structures and Standardization and How this affect Wordnet Applications and Evaluation, pp. 60-64. Las Palmas (2002)
2. Balkova V., Suhonogov A., Yablonsky S.A.: Russian WordNet. From UML-notation to Internet/Intranet Database Implementation, In: Second International WordNet Conference, GWC 2004, pp. 31-38. Brno, Czech Republic (2004)
3. Crystal, D.: Language and The Internet. Cambridge University Press (2001)
4. Edmonds, P.: Introduction to Senseval. ELRA Newsletters, 7(3), pp. 337–344 (2002)
5. Ekman, P.: An argument for basic emotions. Cognition and Emotion, vol. 6(3-4), pp. 169–200 (1992)
6. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: ACM Symposium on Applied Computing, pp 1556-1560, Fortaleza, Brazil (2008)
7. Strapparava, C., Valitutti, A.: Wordnet-affect: an affective extension of wordnet. In: 4th International Conference on Language Resources and Evaluation, pp. 1083–1086 (2004)
8. Strapparava, C., Valitutti, A., Stock, O.: The affective weight of the lexicon. In: 5th International Conference on Language Resources and Evaluation (LREC 2006), pp 474–481, Genoa, Italy, (2006)
9. Tufis, D., Mititelu, B., Bozianu, L., Mihaila, C.: Romanian wordnet: New developments and applications. In: 3rd Conference of the Global WordNet Association, pp. 337–344, Korea, (2006)
10. Tufiş, D., Ion, R., Bozianu, L., Ceauşu, A., Ştefănescu, D.: Romanian Wordnet: Current State, New Applications and Prospects. In: 4th Global WordNet Conference, GWC-2008, pp. 441-452, University of Szeged, Hungary (2008)
11. Esuli, A., Sebastiani, F.: SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In: 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 417-422, Genoa, Italy (2006)
12. Magnini, B., Cavaglia G.: Integrating subject field codes into wordnet. In: Second International Conference on Language Resources and Evaluation (LREC 2002), pp. 1413—1418, Athens, Greece (2002)
13. Ortony, A., Clore, G. L., Foss, M. A.: The psychological foundations of the affective lexicon. Journal of Personality and Social Psychology, vol. 53, pp. 751–766, American Psychological Association (1987)
14. Sokolova M., Bobicev V.: Classification of Emotion Words in Russian and Romanian Languages. In: RANLP-2009 conference, Borovets, Bulgaria, pp. 415-419 (2009)