THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

*PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,*
*Volume 12, Number 2/2011, pp. 143–150*

# CONNOTATIVE MEANING MARKUP FOR WORDNET

Dan TUFIŞ, Dan ŞTEFĂNESCU

Institute for Artificial Intelligence of the Romanian Academy, 13 Septembrie no. 13, 050711, Bucharest
E-mail: tufis@racai.ro, danstef@racai.ro

This article describes an annotation of the synonymy sets in Princeton WordNet2.0, in line with the principles of Osgood's "Semantic Differential" theory. According to this theory, connotative meaning of most adjectives can be rated on a scale, the ends of which are antonymic adjectives. Such a pair of antonymic adjectives is called a factor. The method generalizes previous approaches to apply for all categories of content words (not only adjectives) and takes into account word sense distinctions. In addition to the WordNet structure, the method incorporates knowledge from the SUMO/MILO ontology. The information attached to the synsets generalizes the usual subjectivity markup (positive, negative, and objective) according to a user-based multi-factorial differential semantics model.

*Key words*: Annotation; Connotatative meaning; Differential semantics; Ontology; Synset; Wordnet.

## 1. INTRODUCTION

The recent developments of WordNet Affect [14] and SentiWordNet [3], as well as the hot topic of subjectivity analysis [1, 6, 10, 11, 12, 15. 16], to name just a few relevant papers, all try to remedy the lack of explicit information regarding the sentiment load of the words recorded in a semantic dictionary such as WordNet [4]. However, this type of research is not new as more than 50 years ago the pioneering work of Osgood, Suci, and Tannenbaum [9] on the theory of semantic differentiation gave strong evidence that connotative meanings could be outlined and measured by using a semantic differential technique. According to this theory, the words of a lexical stock can be qualitatively and quantitatively differentiated along the scale defined by an antonymic pair of words. Osgood and his colleagues asked many subjects to rate the meaning of a word, phrase, or text on different scales defined in terms of pairs of bipolar adjectives (good-bad, active-passive, strong-weak, optimistic-pessimistic, beautiful-ugly, etc.). Each pair of bipolar adjectives is a factor in the semantic differential technique. A very interesting fact discovered by Osgood and his colleagues was that most of the variance in the text affecting judgment was explained by only three major factors: the evaluative factor (e.g., good-bad), the potency factor (e.g., strong-weak), and the activity factor (e.g., active-passive). Of these, according to [9], the most discriminative is the evaluative one.

Based on semantic differential model, Kamps and Marx [5] developed algorithmic methods to assess the connotative meaning of adjectives in Princeton WordNet (v1.7). They illustrated their method beginning with the evaluative factor (good-bad). They found 5410 adjectives related to "good" and "bad", then applied the same method with the next two best discriminative factors identified by Osgood and his colleagues: the potency factor (strong-weak) and the activity factor (active-passive). The set of adjectives related to the bipolar adjectives from each of the factors mentioned above represented the same cluster of 5410 adjectives. Depending on the selected factor, various facets of connotative meanings come under scrutiny. The inspiring work of Kamps and Marx still has some major limitations:

1. Although the adjectives make up the major class of the subjectivity bearing words, the other open class categories have significant potential for expressing subjective meanings.
2. The majority of researchers working on subjectivity agree that the subjectivity load of a given word is dependent on the senses of the respective word; yet, in Kamps and Marx's model (KMM, henceforth) the sense distinctions are ignored, making it impossible to assign different scores to different senses of the word in case.

The implementation and the model presented in this article address the two limitations.

## 2. COMPUTATIONAL MODEL FOR THE LEXICAL FACTORIAL MARKUP

Let us begin with some definitions, slightly modified, from KMM. We will progressively introduce new definitions to serve our extended approach.

**Definition 1**. Two words $w_\alpha$ and $w_\beta$ are *related* if there exists a sequence of words ($w_\alpha\ w_1\ w_2 \ldots w_\mathrm{i} \ldots w_\beta$) so that each pair of adjacent words in the sequence belong to the same synset. If the length of such a sequence is $n + 1$ one says that $w_\alpha$ and $w_\beta$ are *n-related.*

For example, the words "good" and "proper" are 2-related since the sequence (good right proper) observes the above definition:

Two words may not be related at all or may be related by many different sequences, of various lengths. In the latter case, one would be interested in their minimal path-length.

**Definition 2.** Let MPL($w_i$, $w_j$) be the partial function:

$$\mathrm{MPL}(w_i, w_j) = \begin{cases} n & \text{if n is the smallest number such that } w_i \text{ and } w_j \text{ are } n\text{-related} \\ \mathit{undefined} & \text{otherwise.} \end{cases} \tag{1}$$

MPL($w_i$, $w_j$) has the following properties:

(1) $\mathrm{MPL}(w_i, w_j) = 0$ iff $w_i = w_j$

(2) $\mathrm{MPL}(w_i, w_j) = \mathrm{MPL}(w_j, w_i)$

(3) $\mathrm{MPL}(w_i, w_j) + \mathrm{MPL}(w_j, w_k) \geq \mathrm{MPL}(w_i, w_k)$.

and, thus, MPL is a distance measure that can be used as a metric for the semantic relatedness of two words. Taking the example from [5], "good" and "bad" are 4-related (good, sound, heavy, big, bad)[1], because **good** and **sound** belong to the synset *01130226-a*, **sound** and **heavy** belong to the synset *00663845-a*, **heavy** and **big** belong to the synset *02316892-a* and, finally, **big** and **bad** belong to the synset *01459996-a*

Observing the properties of the MPL partial function, one can quantify the relatedness of an arbitrary word *w* to one or the other word of a bipolar pair. To this end, KMM introduced another function TRI:

**Definition 3**. Let TRI ($w_i$, $w_\alpha$, $w_\beta$), with $w_\alpha \neq w_\beta$ be:

$$\mathrm{TRI}(w_i, w_\alpha, w_\beta) = \begin{cases} \dfrac{\mathrm{MPL}(w_i, w_\alpha) - \mathrm{MPL}(w_i, w_\beta)}{\mathrm{MPL}(w_\alpha, w_\beta)}, & \text{when all MPLs are defined} \\ \mathit{undefined}, & \text{otherwise.} \end{cases} \tag{2}$$

When defined, **TRI(*$w_i$*, *$w_\alpha$*, *$w_\beta$*)** is a real number in the interval [–1, 1]. The words $\mathrm{w}_\alpha$ and $\mathrm{w}_\beta$ are the reference words – the bipolar words of a factor, while $w_i$ is the word of interest for which TRI is computed. If one takes the negative values returned by the partial function **TRI (*$w_i$*, *$w_\alpha$*, *$w_\beta$*)** as an indication of $w_i$ being more similar to $w_\alpha$ than to $w_\beta$ and the positive values as an indication of $w_i$ being more similar to $w_\beta$ than to $w_\alpha$, then a zero value could be interpreted as $w_i$ being neutrally related with respect to $w_\alpha$ and $w_\beta$. This is different from being unrelated. Therefore, if **α-β** specifies the bipolar words (the factor) used for the computation of relatedness of $w_i$, one could define a proper function **$\mathrm{TRI}^*_{\alpha\text{-}\beta}$ (*$w_i$*)** as follows, with the value 2 representing unrelatedness of $w_i$ with respect to the *α-β* factor.

$$\mathrm{TRI}^*_{\alpha-\beta}(w_i) = \begin{cases} \mathrm{TRI}(w_i, \alpha, \beta), & \text{iff TRI}(w_i, \alpha, \beta) \text{ is defined} \\ 2, & \text{otherwise.} \end{cases} \tag{3}$$

For the major factors identified by [9], namely the evaluative factor (good-bad), the potency factor (strong-weak), and the activity factor (active-passive), one obtains the following scoring functions:

$\mathrm{EVA}(w_i) = \mathrm{TRI}^*_{\text{good-bad}}(w_i)$

$\mathrm{POT}(w_i) = \mathrm{TRI}^*_{\text{strong-weak}}(w_i)$

$\mathrm{ACT}(w_i) = \mathrm{TRI}^*_{\text{active-passive}}(w_i)$.

[1] This is not the only path sequence between "good" and "bad."

Since the KMM defines a factor as a pair of words related by the antonymic relationship (disregarding the senses of the two words), we generalize the notion of a factor to a pair of synsets. In the following, we will use the colon notation to specify the sense number of a literal that licenses the synonymy relation within a synset. Synonymy is a lexical relation that holds not between a pair of words but between specific senses of those words. That is, the notation {literal$_1$:$n_1$ literal$_2$:$n_2$ … literal$_k$:$n_k$} will mean that the meaning given by the sense number $n_1$ of the literal$_1$, the meaning given by sense number $n_2$ of the literal$_2$ and so on are all pair-wise synonymous. The term *literal* is used to denote the dictionary entry form of a word (lemma).

The antonymy is also a lexical relation that holds between specific senses of a pair of words. The synonyms of the antonymic words, taken pairwise, definitely express a semantic opposition. Take for instance the antonymic pair <rise:1 fall:2>. These two words belong to the synsets {rise:1, lift:4, arise:5, move up:1, go up:1, come up:6, uprise:6} and {descend:1, fall:2, go down:1, come down:1}. The pair <rise:1 fall:2> is explicitly encoded as antonymic (i.e., there is an antonymic relationship between the respective word senses). There is, however, a conceptual opposition between the synsets to which the two words belong, that is between any pair of the Cartesian product: {rise:1, lift:4, arise:5, move up:1, go up:1, come up:6, uprise:6}⊗{descend:1, fall:2, go down:1, come down:1}. This semantic opposition is exploited in our model of synset factorial annotation. We denote the semantic opposition of two synsets $S_\alpha$, $S_\beta$, making an S-factor by writing that $S_\alpha \approx \neg S_\beta$ or $S_\beta \approx \neg S_\alpha$.

**Definition 4.** An ***S*-factor** is a pair of synsets ($S_\alpha$, $S_\beta$) for which there exist $w_i^\alpha : s_i^\alpha \in S_\alpha$ and $w_i^\beta : s_i^\beta \in S_\beta$ so that $w_i^\alpha : s_i^\alpha$ and $w_i^\beta : s_i^\beta$ are antonyms and $MPL\left(w_i^\partial, w_i^\beta\right)$ is defined. $S_\alpha$ and $S_\beta$ have opposite meanings, but only $w_i^\alpha : s_i^\alpha$ and $w_i^\beta : s_i^\beta$ are antonyms. For these situations, we consider that $MPL\left(S_\alpha, S_\beta\right) = MPL\left(w_i^\partial, w_i^\beta\right)$. We should mention that not every pair of synsets with opposite meanings represents a factor since it is not the case that MPL is always defined.

Given a factor $w_i^\alpha - w_i^\beta$ and the corresponding *S*-factor $S_\alpha$-$S_\beta$, each word *w* in WordNet that can be reached on a path from $w_i^\alpha$ to $w_i^\beta$ is given a score number, which is a function of the distances from *w* to $w_i^\alpha$ and to $w_i^\beta$. The set of these words defines the coverage of the $w_i^\alpha$ - $w_i^\beta$ factor – COV($w_i^\alpha$ - $w_i^\beta$). The set of all synsets containing the words in COV($w_i^\alpha$ - $w_i^\beta$) defines the semantic coverage of the corresponding *S*-factor – SCOV($S_\alpha$-$S_\beta$).

The experiments show that the coverage of the vast majority of the factors, corresponding to the same POS category, is the same. From now on, we will use *U* to designate this common coverage. Table 1 gives coverage figures for each of the POS categories in the PWN 2.0. For adjectives, the size of the U coverage (literals) has a similar value to the one reported by [5] (5410). The difference might be explained by the fact that we use a different version of the Princeton WordNet.

*Table 1*

POS categories coverage for PWN 2.0

| Word Class | Factors | *U* Coverage (literals) | Maximal Semantic Coverage (synsets) |
|---|---|---|---|
| Adjectives | 335 | 5,307 (24.68%) | 5,291 (28.50%) |
| Adverbs | 335 | 1,943 (41.69%) | 1,571 (42.87%) |
| Nouns | 85 | 11,109 (9.59%) | 11,007 (13.81%) |
| Verbs | 254 | 6,467 (57.19%) | 8,589 (64.58%) |

The SCOV($S_\alpha$-$S_\beta$) as defined above, purposely ignored the word senses in order to allow for a comparison with Kamps and Marx's findings. However, for our objectives, we further introduce the notions of *semantic type of a synset, typed S-factor,* and *scoped synset with respect to a typed S-factor*, which represent major deviations from KMM. Before that, we need to introduce the mapping between the WordNet synsets and the SUMO/MILO concepts. The Suggested Upper Merged Ontology (SUMO), Mid Level Ontology (MILO) [7] and its domain ontologies (http://www.ontologyportal.org/), form the largest formal public ontology in existence today, containing roughly 20,000 terms and 70,000 axioms (when all SUMO, MILO, and domain ontologies are combined). It is owned by the IEEE, but it is freely downloadable. One of the major attractions of this ontology is that it has been mapped to the WordNet lexicon [8]. Using this mapping, synsets are labeled with a SUMO/MILO concept which we will refer to below as the synset's *semantic type*. We were especially interested in the hierarchical structure of the ontology our extension of the KMM relies on.

Another useful mapping for the WordNet synsets is the DOMAINS taxonomy [2]. The DOMAINS structuring uses Dewey Decimal Classification codes to classify the 115425 PWN synsets into 168 distinct classes (http://wndomains.itc.it/). For uniformity, we will also refer to a DOMAINS class attached to a WordNet synset as the synset's *semantic type*. Depending on the intended granularity of the annotation, one could use either SUMO/MILO or DOMAINS semantic types [2].

**Definition 6.** An S-factor $S_\alpha$-$S_\beta$ is said to be a *typed S-factor* if the types of the synsets $S_\alpha$ and $S_\beta$ have a common ancestor. If this ancestor is the lowest common ancestor, it is called the 0-*semantic type* of the *S*-factor. The direct parent of the *n-semantic type* of an S-factor is called the *n*+1-*semantic type* of the *S*-factor.

**Definition 7.** A synset $S_i$ with the type L is *n-scoped* relative to a typed *S*-factor $S_\alpha$-$S_\beta$ if L is a node in a sub-tree of the SUMO/MILO or DOMAINS hierarchy having as root the *n*-semantic type of the *S*-factor $S_\alpha$-$S_\beta$. We say that ***n*** defines the **level of the scope coverage of the factor** $S_\alpha$-$S_\beta$ and that every synset in this coverage is ***n*-scoped**.

We use the notation $SCOV_n(S_\alpha\text{-}S_\beta)$ for the scope coverage of level *n* of an S-factor $S_\alpha$-$S_\beta$. If the root of the tree has the semantic type γ, we will use also use the notation $\mathrm{SCOV}_n(S_\alpha\text{-}S_\beta)_\gamma$ or simply[3] $\mathrm{SCOV}(S_\alpha\text{-}S_\beta)_\gamma$. Put into other words, $\mathrm{SCOV}(S_\alpha\text{-}S_\beta)_\gamma$ is the set of synsets the semantic types of which are subsumed by γ.

The original KMM definition of *relatedness* between words remains the same when the value of the scope coverage level is increased so as to reach the top of the ontology. When *n* is maximized, we obtain the maximum coverage in which any S-factor can describe any sense for the words in *U*. An identical definition may be formulated by using, instead of SUMO/MILO or DOMAINS labels, the WordNet hierarchy.

For the previous example with the words "good" and "proper," their $\mathrm{SCOV}_0$ root has the semantic type the *NormativeAttribute* concept since the semantic type of the synset 0161119-a (good:14 right:13 ripe:3) is *SubjectiveAssessmentAttribute*, the semantic type of the synset 00140845a (right:6 proper:3 suitable:3) is *NormativeAttribute* and *SubjectiveAssessmentAttribute* ***ISA*** *NormativeAttribute*.

The introduction of typed S-factors is necessary in order to counteract the effects of the way in which *relatedness* is defined. The MPL (see Definition 2) frequently links semantically unrelated synsets. To eliminate this inconvenience, we restrict the MPL computation only for words belonging to synsets that are n-scoped relative to the chosen S-factor.

For a certain synset, the level of scope coverage of an S-factor $S_\alpha$-$S_\beta$ should have a direct influence on the score assigned to it. Each synset in $\mathrm{SCOV}_n(S_\alpha\text{-}S_\beta)$ is characterized by the S-factor, in a way which can be quantified by a TRI$^*$-like score (Eq. 3). The synsets in $\mathrm{SCOV}_0$ $(S_\alpha\text{-}S_\beta)$ are best characterized, meaning that their scores for the $S_\alpha$-$S_\beta$ factor are highest. For the synsets in $\mathrm{SCOV}_n(S_\alpha\text{-}S_\beta)$ that cannot be found in $\mathrm{SCOV}_{n-1}(S_\alpha\text{-}S_\beta)$, the scores are smaller and we say that the characterization of these synsets in terms of $S_\alpha$-$S_\beta$ factor is weaker. This means that a higher value of *n* should imply an increased neutrality of the synset or its literals with respect to the S-factor. Our model captures this through a slight modification of the TRI function in Eq. 2, where $w_\alpha$ and $w_\beta$ are the antonyms belonging to $S_\alpha$ and $S_\beta$ respectively, and $w_i$ is a literal of a synset $S_j$ in $\mathrm{SCOV}_n(S_\alpha\text{-}S_\beta)$ but not in $\mathrm{SCOV}_{n-1}(S_\alpha\text{-}S_\beta)$:

$$\mathrm{TRI}^{+}(w_i,S_\alpha,S_\beta)=\frac{MPL(w_i,w_\alpha)-MPL(w_i,w_\beta)}{MPL(w_\alpha,w_\beta)+n}. \quad (4)$$

Since we imposed the requirement that $S_j$ be in $\mathrm{SCOV}_n(S_\alpha\text{-}S_\beta)$, $\mathrm{TRI}^{+}(w_i,S_\alpha,S_\beta)$ is defined for all literals in $S_j$, thus for any $w_i \in S_j$ the value of $\mathrm{TRI}^{+}(w_i,S_\alpha,S_\beta)$ is in the [–1,1] interval. The scores computed for the synsets in $\mathrm{SCOV}_n(S_\alpha\text{-}S_\beta)$ remained unchanged in $\mathrm{SCOV}_{n+k}(S_\alpha\text{-}S_\beta)$. for any $k \geq 0$.

[2] Our actual implementation of the presented approach uses only the SUMO/MILO semantic types.

[3] The level of the scope coverage for a given factor and the semantic type of the root of the associated sub-tree are, obviously, dependent on one another.

The above modification of the TRI function insures that the score of a synset gets closer to zero (neutrality) with the increase of *n*. Of course, this push towards neutrality with the increase of *n* can be defined and implemented in various ways, depending on how fast one wants it to happen.

**Definition 8.** Let $S_\alpha$-$S_\beta$ be an S-factor and $S_i$ a synset in SCOV$_n$($S_\alpha$-$S_\beta$); TRIS ($S_\text{i}$, $S_\alpha$, $S_\beta$) is defined as the average of the TRI$^+$ values associated with the literals forming the synset $S_i$.

$$\mathrm{TRIS}(S_i, S_\alpha, S_\beta) = \frac{\sum_{j=1}^{m} \mathrm{TRI}^+(w_j, S_\alpha, S_\beta)}{m}. \tag{5}$$

The semantic typing of an *S*-factor and the definition of the corresponding semantic coverage can be achieved at various granularities, depending on available classifications of synsets in the backing wordnet. As mentioned before, for Princeton WordNet, there are two distinct synset annotations available which are relevant to the purpose of our research: domains labels (animals, biology, geography, plants, psychological features, etc.) and SUMO/MILO concepts (Animal, Plant, FieldOfStudy, BiologicalProcess, etc.).

A typed *S*-factor is represented by indexing the S-factor with its type (of a desired granularity), as in the examples below:

({unfairness:2...}<-> { fairness:1...})$_{\text{NormativeAttribute}}$ ; ({discomfort:1...} <–> {comfort:1...})$_{\text{StateOfMind}}$
({distrust:2...} <–> {trust:3...})$_{\text{TraitAttribute}}$ ; ({decrease:2... }<->{increase:3...}) $_{\text{QuantityChange}}$
({bad:1...}<->{good:1...}) $_{\text{SubjectiveAssessmentAttribute}}$ ; ({inactive:2...}<->{active:1}) $_{\text{BiologicalAttribute}}$

In the following, if not otherwise specified, by “S-factors” we mean typed S-factors. Unless there is ambiguity, the type of a typed S-factor will be omitted.

## 2. COMPUTING THE S-FACTORS

The aim of our research was to associate each synset S$_\text{k}$ of WordNet with a vector <$F_1$, $F_2$… $F_n$>, where $F_i$ is a pair *(score; level)* with *score* and *level* representing the value of the $i^{\text{th}}$ *S*-factor and, respectively, the minimal *S*-factor coverage level in which $S_k$ was found. For instance, let us assume that the first two S-factors in the description of nominal synsets are:

({*mercilessness*:2 *unmercifulness:1*}<->{*mercifulness*:2 *mercy:2*})$_{\text{+ SubjectiveAssessmentAttribute}}$
({*wildness*:*3*}<->{ *tameness*:2 *domestication:2*})$_{\text{InternalAtttribute}}$

then for the synset {*beast:2, wolf:5, savage:2, brute:1, wildcat:2*}$_{\text{Human}}$ the vector is: <(–0.2272 ; 2) (–0.25 ; 2) . . .>.

The values signify that the synset {*beast:2 ...*}$_{\text{Human}}$ is 2-scoped with respect to each of the two S-factors (meaning that it occurred in the coverages of level 2 of the two S-factors but not in coverages of smaller level), and its meaning is significantly closer to the meaning of mercilessness (–0.2272), and the meaning of wildness (–0.25).

In our experiments, in order to ensure the same sets of synsets for all factors of a given part-of-speech we set the level of the semantic coverages to 7 (corresponding to the U-coverages). For each of the typed *S*-factors $S_\alpha$-$S_\beta$ and for each synset $S_i$ in their respective semantic coverage SCOV< $S_\alpha$, $S_\beta$ >$_\gamma$ we computed the $\mathrm{TRIS}\left(S_i, S_\alpha, S_\beta\right)$ score. Each synset from the coverage of each POS category was associated with a vector of scores, as described above. Since the number of S-factors depends on the POS category (noun, verb, adjectives, and adverbs) the lengths of each of the four type vectors is different. For instance, each noun synset (in the noun coverage) is associated with an 85-cell vector. The cell values in a synset vector have very different values, showing that factors have different discriminative power for a given word sense. Because we considered U coverages, all S-factors are relevant and the cells in any synset vector are filled with pairs *(score; level*).

For the noun part of the WordNet 2.0 lexical ontology, we identified 85 typed S-factors, all of them covering the same set of 11,037 noun literals (9.62%) with their senses clustered into 10,874 synsets (13.64%).

For the verb part of the PWN 2.0, we identified 246 typed S-factors, all of them covering the same set of 6,443 verb literals (56.98%) with their senses encoded into 8,516 synsets (63.04%).

For the adjective part of the PWN 2.0 lexical ontology, we identified 332 typed S-factors, all of them covering the same set of 5,299 literals (24.72%) with their senses encoded into 5,241 synsets (28.23%). The same factors were used for the adverbs derived from adjectives. In this way, a total of 1,933 adverbs (41.48%) clustered into 1,536 synsets (41.92%) were successfully annotated. These results are summarized in Table 2, which is a revised version of Table 1. Intuitively, the numbers in Table 2 should be identical with those in Table 1. The differences appear to be owed to the incompleteness or inconsistencies of the SUMO/MILO annotations in PWN. Some synsets are not mapped onto a SUMO/MILO concept or, sometimes, the taxonomical relation between pairs of synsets in WordNet is opposite to taxonomic relation between their types in SUMO/MILO.

*Table 2*

Multifactorial annotation of the PWN2.0

| Word Class | Typed S-Factors | S-Factors Coverage (literals) | S-Factors Coverage (synsets) |
|---|---|---|---|
| Adjectives | 332 | 5,299 (24.72%) | 5,241 (28.23%) |
| Adverbs | 332 | 1,933 (41.48%) | 1,536 (41.92%) |
| Nouns | 85 | 11,037 (9.62%) | 10,874 (13.64%) |
| Verbs | 246 | 6,443 (56.98%) | 8,516 (63.04%) |

In case the user restricted the coverages to lower levels, the original maximal semantic coverages split into smaller subsets for which several S-factors become irrelevant. The cell values corresponding to these factors are filled in with a conventional value outside the interval [–1, 1]. Thus, we have defined the following annotation situations:

1. A synset of a certain POS is not in the POS maximal semantic coverage. This case signifies that the synset cannot be characterized in terms of the differential semantics methodology and we conventionally say that such a synset is "objective" (insensitive to any *S*-factor). Since this situation would require a factor vector with each cell having the same value (outside the [–1, 1] interval) and as such a vector would be completely uninformative, we decided to leave the "objective" synsets unannotated. As one can deduce from Table 2, the majority of the synsets in PWN2.0 are in this category (almost 90,000).
2. Any synset of a certain POS in the POS coverage will have an associated factor vector. There are 26,167 such synsets. The $i^{\text{th}}$ cell of such a vector will correspond to the $i^{\text{th}}$ *S*-factor $S_\alpha$-$S_\beta$. We may have the following sub-cases:
   (a) All cell scores are in the [–1,1] interval, and in this case all factors are relevant, that is, from any word in the synset one could construct a path to either of the words forming a factor, irrespective of the factor itself. A negative score in the $i^{\text{th}}$ cell of the factor vector signifies that the current synset is more semantically related to $S_\alpha$ than to $S_\beta$, while a positive score in the $i^{\text{th}}$ cell of the factor vector signifies that the synset is more semantically related to $S_\beta$ than to $S_\alpha$. A zero score in the $i^{\text{th}}$ cell of the factor vector signifies that the synset is neutral with respect to the $<S_\alpha, S_\beta>$ factor.

   (b) Several cell scores are not in the interval [–1, 1], say $FV[i_1]=FV[i_2] \ldots =FV[i_k] = 2$. This signifies that the factors corresponding to those cells ($<S_{\alpha1}, S_{\beta1}>$, $<S_{\alpha2}, S_{\beta2}>$, …, $<S_{\alpha3}, S_{\beta3}>$) are irrelevant for the respective synset and that the current synset is not included in the scope of the above-mentioned factors, owing to the selected scope level of the coverage. We say that the synset is "objective" with respect to the irrelevant factors.

We developed an application that allows text analysts to choose the S-factors they would like to work with. The interface allows the user to both select/deselect factors and to switch the order of the poles in any given factor. Once the user decides on the relevant S-factors for his/her application and domain, the synsets are marked up as required according to the selected S-factors. This version of the WordNet can be saved and used as needed in the planned application.

Let us exemplify our approach for the noun and verb synsets. Let us suppose that we would like to differentiate the noun synsets according to three factors:
({*discomfort*:1…}<->{*comfort*:1…})$_{\text{StateOfMind}}$; ({*pain*:2…}<->{*pleasure*:1…})$_{\text{EmotionalState}}$; ({*distrust*:2…}<->{*trust*:3…})$_{\text{TraitAttribute}}$

and the verb synsets according to three other factors:

({*get worse*:1…}<->{*get well*:1…})$_{\text{OrganismProcess}}$; ({*suffer*:1… }<->{*enjoy*:4…})$_{\text{Abstract}}$; ({*disbelieve*:1…}<->{*believe*:1 …})$_{\text{Entity}}$

A sentence like "His **lies** will be **dealt** with in the **court** and his **immorality** will be **proved**." would have the following annotations:

His **lies:1** <*comfort*:-0.07 *pleasure*:-0.16 *trust*:-0.07> will be **dealt:2** <*get well*:0.33 *enjoy*:0 *believe*:0.62> within the **court:1** <*comfort*:-0.06 *pleasure*:-0.05 *trust*:-0.06> and his **immorality:2** <*comfort*:-0.15 *pleasure*:-0.05 *trust*:-0.07> will be **proved:3** <*get well*:0.11 *enjoy*:-0.16 *believe*:0.25>.

Grosser analysis suggests that we have a subjectively loaded sentence that expresses lack of *comfort* (i.e., discomfort (average score: –0.09), lack of *pleasure* (i.e., pain (average score: –0.08), lack of *trust* (i.e., distrust (average score –0.06), *getting well* (average score: 0.22), not *enjoying* (i.e., suffering (average score: –0.08) and *believing* (average score: 0.43). If one decides to describe everything in terms of the good-bad dichotomy, this sentence conveys a rather negative connotation: discomfort, pain, distrust, and suffering are definitely bad things. On the other hand, the "bad" things are taken care of and so the S-factors *getting well* and *believing* have positive values as induced by the verbs *deal* and *prove*.

If one compares this with the <P,N,O> markup in SentiWordNet [3] one would obtain for the same nouns the annotations:

His **lies:1** <P:0 N:0 O:1> will be **dealt**:2 <P:0.125 N:0 O:0.875> within the **court:1** <P:0 N:0 O:1>
and his **immorality:2** <P:0.75 N:0 O:0.25> will be **proved:3** <P:0 N:0 O:1>.

In terms of the <P, N, O> triad, one would eventually obtain an almost objective statement (average score 0.825) with a significant load of positivism (average score 0.175) and no negativity at all. One should note that unlike the markup in SentiWordNet where the three values of the subjectivity annotation sum to one, in our approach this is not true, as the S-factors are considered independent.

## 3. EXTENDING THE MAXIMUM SEMANTIC COVERAGES

Although the maximum semantic coverage of the S-factors for the adjectives contains more than 28% of the PWN2.0 adjectival synsets, many adjectives with connotative potential are not in this coverage (for instance *predictable;* yet, its antonym *unpredictable* **is** in the *U* coverage of adjectives). This happens because the definition of the *relatedness* (Definition 1) implicitly assumes the existence of synonyms for one or more senses of a given word. Therefore from monosemous words in monosense synsets a path towards other synsets cannot be constructed anymore. Because of this, there are isolated "bubbles" of *related* synsets that are not connected with synsets in maximum semantic coverage. In order to assign values to at least a part of these synsets, we experimented with various strategies out of which the one described herein was considered the easiest to implement and, to some extent motivated, from a conceptual point of view.

The approach we have followed is similar for all the synsets which are not in the maximal semantic *coverages (M)*, but the algorithms for extending these coverages slightly differ depending on the part of speech we are considering. The basic idea is to transfer the vectors from synsets in $M$ to those in $\overline{M}$ provided that they have "similar meanings". We say that $S_i^{POS} \in M_{POS}$ and $S_j^{POS} \in \overline{M}_{POS}$ have "similar meanings" if $SUMO/MILO(S_i^{POS}) = SUMO/MILO(S_j^{POS})$ and $S_i^{POS}$ and $S_j^{POS}$ are directly linked by a WordNet relation of a certain type. For adjectival synsets we consider the relations *similar_to* or *also_see*, for verbal synsets we consider the relations *hyponym*, *also_see* or *subevent*, whereas for the nominal synsets we take into account only the *hyponymy* relation. Consequently, we increased the S-factor coverage to the values presented in Table 3.

*Table 3*

The extended coverage of the S-factors

| Word Class | S-Factors Extended Coverage (literals) | S-Factors Extended Coverage (synsets) |
|---|---|---|
| Adjectives | 8,576 (40.00%) | 7,634 (41.12%) |
| Adverbs | 2,393 (51.35%) | 1,995 (54.44%) |
| Nouns | 27,027 (23.57%) | 22,190 (27.84%) |
| Verbs | 8,831 (78.10%) | 10,717 (79.33%) |

## 4. CONCLUSIONS

We revised and improved our proposed method for lexical annotation of the synsets of a wordnet [13], which generalizes the SentiWordNet subjectivity markup according to a user-based multi-criteria differential semantics model. We discussed the method for annotating the synsets in PWN2.0, irrespective of their part of speech. We anticipate that these annotations can be imported to other language wordnets, provided they are aligned with PWN2.0.The annotation system does not depend on the language of the wordnet, but requires its alignment with the Princeton WordNet 2.0, from which the SUMO/MILO and DOMAINS markup can be automatically imported.

## ACKNOWLEDGEMENT

The work reported here is funded by the STAR project, financed by the Ministry of Education, Research and Innovation (UEFISCDI) under the grant no. 742/19.01.2009.

## REFERENCES

1. ANDREEVSKAIA, A., BERGLER, S., *Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses*, Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006), pages 209–216, Trento, Italy, 2006.
2. BENTIVOGLI L., FORNER, P., MAGNINI, B., PIANTA, E., *Revising WordNet domains hierarchy: Semantics, coverage, and balancing*, Proceedings of COLING 2004 Workshop on "Multilingual Linguistic Resources", Geneva, 2004, pp.101–108.
3. ESULI, A., SEBASTIANI, F., *SENTIWORDNET: A publicly available lexical resource for opinion mining*. Proceedings of the 5th Conference on Language Resources and Evaluation LREC-06, Genoa, 2006, pp. 417–422.
4. FELLBAUM, C., *WordNet: An Electronic Lexical Database*, Academic Press, Cambridge, MA, 1998.
5. KAMPS J. and MARX, M., *Words with attitude*, Proceedings of the 1st International WordNet Conference, Mysore, India, 2002, pp. 332–341.
6. MIHALCEA R., BANEA, C., WIEBE J., Learning *multilingual subjective language via cross-lingual projections*, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, Prague, pp. 976–983.
7. NILES I., PEASE, A., Towards a standard upper ontology, In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, 2001, Ogunquit, Maine, pp. 2–9.
8. NILES I., PEASE, A., *Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology*, Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), 2003, Las Vegas pp. 23–26.
9. OSGOOD, E. C., SUCI, G. TANNENBAUM, P., *The measurement of meaning*, University of Illinois Press, Urbana IL, 1957.
10. PANG, B. LEE, L. *Opinion mining and sentiment analysis*, Foundations and Trends in Information Retrieval, **2**, *1–2*, 1–135, 2008.
11. POLANYI, L., ZAENEN, A., *Contextual valence shifters*, Shanahan, Y. Qu and J. Wiebe (eds.), *Computing Attitude and Affect in Text: Theory and Applications,* The Information Retrieval Series, Vol. **20**, Springer Verlag, Dordrecht, Netherlands, 2006 pp. 1–10.
12. RILOFF, E., WIEBE, J., WILSON, T., *Learning subjective nouns using extraction pattern bootstrapping*. Proceedings of the Seventh Conference on Natural Language Learning (CONLL-2003), 2003, Edmonton, Canada, pp. 25–32.
13. TUFIŞ D., ŞTEFĂNESCU D., *A Differential Semantics Approach to the Annotation of the Synsets in WordNet*, Proceedings of LREC 2010, Malta, 2010.
14. VALITUTTI, A., C. STRAPPARAVA, C., STOCK, O.*, Developing affective lexical resources*, Psychology Journal, **2**, *1*, pp. 61–83, 2004.
15. WIEBE, J., *Learning subjective adjective from corpora*, Proceedings of the 17th National Conference on Artificial Intelligence (AAAI'2000), Austin, Texas, 2000 pp. 735–740.
16. WIEBE, J., WILSON, T., BRUCE, R., BELL, M., MARTIN, M. *Learning subjective language*, Computational Linguistics, **30**, *3*, pp. 277–308, 2004.

*Received January 15, 2011*