Managing Multiword Expressions in a Lexicon-Based Sentiment Analysis System for Spanish

Antonio Moreno-Ortiz and **Chantal Pérez-Hernández** and **M. Ángeles Del-Olmo**
Facultad de Letras
Universidad de Málaga
29071 Málaga. Spain
{amo, mph, mariadelolmo@uma.es}

Abstract

This paper describes our approach to managing multiword expressions in Sentitext, a linguistically-motivated, lexicon-based Sentiment Analysis (SA) system for Spanish whose performance is largely determined by its coverage of MWEs. We defend the view that multiword constructions play a fundamental role in lexical Sentiment Analysis, in at least three ways. First, a significant proportion conveys semantic orientation; second, being units of meaning, their relative weight to the calculated overall sentiment rating of texts needs to be accounted for as such, rather than the number of component lexical units; and, third, many MWEs contain individual words that carry a given polarity, which may or may not be that of the phrase as a whole. As a result, successful lexicon-based SA calls for appropriate management of MWEs.[1]

1 Introduction

In recent years, *sentiment analysis* or *opinion mining* has become an increasingly relevant sub-field within natural language processing that deals with the computational treatment of opinion and subjectivity in texts. The fact that emotions and opinions condition how humans communicate and motivate their actions explains why the study of evaluative language has attracted a great deal of attention from a wide range of disciplines (Pang and Lee, 2008).

With the advent of the Web 2.0 and the widespread use of social networks, it is easier than ever before to gain access to vast amounts of sentiment-laden texts. User reviews are particularly interesting for companies as a tool for product improvement. Different opinions and trends in political or social issues can be identified, to the extent that many companies have decided to add sentiment analysis tools to their social media measurement and monitoring tools with a view to improving their business.

With regard to MWEs, their relevance to Natural Language Processing in general, and to Sentiment Analysis in particular, can hardly be overstated since they constitute a significant proportion of the lexicon of any natural language. It is estimated that the number of MWEs in the lexicon of a native speaker has the same order of magnitude as the number of single words (Jackendoff, 1997) and even these ratios are probably underestimated when considering domain-specific language, in which the specialized vocabulary and terminology are composed mostly by MWEs. As Erman and Warren (2000: 29) point out, the fact that half of spoken and written language comes in preconstructed multiword combinations makes it impossible to consider them as marginal phenomena. Further, a large number of such expressions

[1] This work is funded by the Spanish Ministry of Science and Innovation (Lingmotif Project FFI2011-25893).

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 1–10,
Atlanta, Georgia, 13-14 June 2013.

express emotions and opinions on the part of the speaker, so it follows that any lexicon-based approach to sentiment analysis somehow needs to account for multiword constructions.

2 Sentiment Analysis in perspective

Sentiment Analysis approaches mainly fall into one of two categories, which are usually referred to as the lexicon-based approach and the machine-learning approach. The latter is undoubtedly more popular for many reasons, an important one being a faster bootstrapping process, but also reasonably good performance (Pang and Lee, 2005; Aue and Gamon, 2005). In fact, machine learning techniques, in any of their flavors, have proven extremely useful, not only in the field of sentiment analysis, but in text mining and information retrieval applications in general, as well as a wide range of data-intensive computational tasks. However, their obvious disadvantage in terms of functionality is their limited applicability to subject domains other than the one they were designed for. Although interesting research has been done aimed at extending domain applicability (Aue and Gamon, 2005), such efforts have shown limited success. An important variable for these approaches is the amount of labeled text available for training the classifier, although they perform well in terms of recall even with relatively small training sets (Andreevskaia and Bergler, 2007).

In contrast, lexicon-based approaches rely on dictionaries where lexical items have been assigned either *polarity* or *valence*, which has been extracted either automatically from other dictionaries, or, more uncommonly, manually. Although the terms *polarity* and *valence* are sometimes used interchangeably in the literature, especially by those authors developing binary text classifiers, we restrict the usage of the former to non-graded, binary assignment, i.e., positive / negative, whereas the latter is used to refer to a rating on an *n*-point semantic orientation scale. The works by Hatzivassiloglou and Wiebe (2000), and Turney (2002) are perhaps classical examples of such an approach. The most salient work in this category is Taboada et al. (2011), whose dictionaries were created manually and use an adaptation of Polanyi and Zaenen's (2006) concept of Contextual Valence Shifters to produce a system for measuring the semantic orientation of texts, which they call SO-CAL(culator). This is exactly the approach we used in our Sentitext system for Spanish (Moreno-Ortiz et al., 2010).

Hybrid, i.e., semi-supervised, approaches have also been employed, as in Goldberg and Zhu (2006), where both labeled and unlabeled data are used. Extraction of lexical cues for semantic orientation (i.e., polarity) is usually performed semi-automatically, for example by Mutual Information scores obtained from adjectives or adverbs, which are the most obvious word classes to convey subjective meaning. To a lesser extent, nouns (e.g. Riloff et al., 2003) and verbs (e.g. Riloff and Wiebe, 2003) have also been used to identify semantic orientation. It is worth noting at this point that no mention has been made thus far of MWE's. The reason is simply that they have by and large been ignored, probably due to the increased complexity that dealing with them involves.

Sentiment Analysis approaches can also be classified according to output granularity. Most systems fall in the *Thumbs up or Thumbs Down* approach, i.e., producing a simple positive or negative rating. Turney's (2002) work, from which the designation derives, is no doubt the most representative. A further attempt can be made to produce not just a binary classification of documents, but a numerical rating on a scale. The rating inference problem was first posed by Pang and Lee (2005), and the approach is usually referred to as *Seeing Stars* in reference to that work, where they compared different variants of the original SVM binary classification scheme aimed at supporting *n*-ary classification. Gupta et al. (2010) further elaborated on the multi-scale issue by tackling multi-aspect, i.e., pinpointing the evaluation of multiple aspects of the object being reviewed, a feature we regard as essential for high-quality, fine-grained sentiment analysis, but one that requires very precise topic identification capabilities.

2.1 Sentiment Analysis for Spanish

Nor surprisingly, work within the field of Sentiment Analysis for Spanish is, by far, scarcer than for English. Besides, most studies focus on specific domains, typically movie reviews.

Cruz et al. (2008) developed a document classification system for Spanish similar to Turney's (2002), i.e. unsupervised, though they also tested a supervised classifier that yielded better results. In

both cases, they used a corpus of movie reviews taken from the Spanish Muchocine website. Boldrini et al. (2009) carried out a preliminary study in which they used machine learning techniques to mine opinions in blogs. They created a corpus for Spanish using their Emotiblog system, and discussed the difficulties they encountered while annotating it. Balahur et al. (2009) also presented a method of emotion classification for Spanish, this time using a database of culturally dependent emotion triggers. Finally, Brooke et al. (2009) adapted a lexicon-based sentiment analysis system for English (Taboada et al., 2011) to Spanish by automatically translating the core lexicons and adapting other resources in various ways. They also provide an interesting evaluation that compares the performance of both the original (English) and translated (Spanish) systems using both machine learning methods (specifically, SVM) and their own lexicon-based semantic orientation calculation algorithm, SO-CAL, mentioned above. They found that their own weighting algorithm, which is based on the same premises as our system, achieved better accuracy for both languages, but the accuracy for Spanish was well below that for English.

Our system, Sentitext (Moreno-Ortiz et al., 2010; 2011), is very similar to Brooke et al.'s (2009) in design: it is also lexicon-based and it makes use of a similar calculation method for semantic orientation. It differs in that the lexical knowledge has been acquired semi-automatically and then manually revised from the ground up over a long period of time, with a strong commitment to both coverage and quality. It makes no use of user-provided, explicit ratings that supervised systems typically rely on for the training process, and it produces an index of semantic orientation based on weighing positive against negative text segments, which is then transformed into a ten-point scale and a five-star rating system.

Yet another way in which our system differs from most other systems, including Taboada et al.'s (2011), is in the relevance given to multiword expressions vis-à-vis individual words.

3 Sentitext: a SA system for Spanish

Sentitext is a web-based, client-server application written in C++ (main code) and Python (server). The only third-party component in the system is Freeling (Atserias et al., 2006; Padró, 2011), a powerful, multi-language NLP suite of tools, which we use for basic morphosyntactic analysis. Currently, only one client application is available, developed in Adobe Flex,[2] which takes an input text and returns the results of the analysis in several numerical and graphical ways, including visual representations of the text segments that were identified as sentiment-laden. For storage, we rely on a relational database (MySQL), where lexical information is stored.

Given that it is a linguistically-motivated sentiment analysis system, special attention is paid to the representation and management of the lexical resources that Sentitext uses for its analysis. The underlying design principle is to isolate lexical knowledge from processing as much as possible, so that the processors can use the data directly from the database. The idea behind this design is that all lexical sources can be edited at any time by any member of the team, which is facilitated by a PHP interface specifically developed to this end. We believe this approach is optimal for lexicon-based systems, since it allows improvements to be easily incorporated simply by updating the database by means of a user-friendly interface.

3.1 Data sources

Sentitext relies on three major sources: the individual word dictionary (*words*), the multiword expressions dictionary (*mwords*), and the context rules set (*crules*), which is our implementation of Contextual Valence Shifters (Polanyi and Zaenen, 2006).

The individual word dictionary currently contains over 9,400 items, all of which are labeled for valence. The acquisition process for this dictionary was inspired by the bootstrapping method recurrently found in the literature (e.g., Riloff and Wiebe, 2003, Aue and Gamon, 2005). We adapted this methodology in the following way: first, we established a set of 22 antonymic pairs of words to be used as seed words, which we fed to the Spanish version of the OpenOffice thesaurus in order to track its contents for sentiment-carrying words. However, rather than doing this automatically, we built an interactive tool that presented a user with consecutive rounds of candidate words to be added to the dictionary, thus providing the means to

[2] This application can be accessed and tested online at http://tecnolengua.uma.es/sentitext

block wrong polarity assignments, caused mainly by polysemy, that would propagate to subsequent sets of synonymous words. The resulting dictionary was thoroughly revised manually and actual valences were added by lexicographers using the GDB tool. In Section 4, we elaborate on this process of manual valence assignment in relation to the MWEs dictionary, which does not differ from the one used in the word dictionary. Lexical items in both dictionaries in our database were assigned one of the following valences: -2, -1, 0, 1, 2. However, since the word dictionary contains only sentiment-carrying items, no 0-valence word is present.

The SA system most similar to ours (Taboada et al., 2011) uses a scale from -5 to +5, which makes sense for a number of graded sets of near synonyms such as those given as examples by the authors (p. 273). In our opinion, however, as more values are allowed, it becomes increasingly difficult to decide on a specific one while maintaining a reasonable degree of objectivity and agreement among different (human) acquirers, especially when there is no obvious graded set of related words, which is very often the case. In fact, our initial intention was to use a -5 to 5 scale, but this idea was abandoned, as the difficulty for assigning such fine-grained valences became apparent in actual practice on a large scale dictionary.

This does not imply that valence values for actual words and MWEs in context are limited to these. In a lexicon-based SA system that computes a sentiment rating based on weighing positive against negative text segments there should be a way to distinguish not only between, for example, the adjectives “good” and “bad”, but also deal with the semantics of qualifiers, as in “very good”, and “extremely good”. This is where context rules come into play.

3.2 Context rules

It is important to understand the way our context rules work in order to appreciate how closely they interact with the other lexical data sources, especially the multiword dictionary. Simply accounting for negative and positive words and phrases found in a text would not be enough. There are two ways in which their valence can be modified by the immediately surrounding context: the valence can change in degree (intensification or downtoning), or it may be inverted altogether. Negation is the simplest case of valence inversion.

The idea of Contextual Valence Shifters (CVS) was first introduced by Polanyi and Zaenen (2006), and implemented for English by Andreevskaia and Bergler (2007) in their CLaC System, and by Taboada et al. (2011) in their Semantic Orientation CALculator (SO-CAL). To our knowledge, apart from Brooke et al.’s (2009) adaptation of the SO-CAL system, Sentitext is the only sentiment analysis system to implement CVS for Spanish *natively*.

Our CVS system is implemented in what we call Context Rules, which are expressed as the following data structure:

1. Unit Form: Freeling-compliant morpho-syntactic definition of the item being modified (e.g.: "AQ" for qualifying adjectives).
2. Unit Sign: polarity of the item being modified (e.g. "+").
3. CVS Definition: modifier definition (e.g.: *very*,“very”).
4. CVS Position: position of the modifier (e.g. "L" for left).
5. CVS Span: maximum number of words where the modifier can be found in the modified item.
6. Result: valence result of the modification. This result can be expressed as either an operator or a set valence. An operators is one of the following
   - INV (valence/polarity INVersion)
   - INT*n* (valence INTensification of *n*)
   - DOW*n* (valence DOWntoning of *n*).

The *n* argument in the last two operators is the degree by which the operator is to be applied. The result can also be a set valence, in which case it looks like any valence expressed in the dictionaries.

This system allows us to describe fairly elaborate context rules; for instance, having multiword modifiers such as those in (1) and (2) below. A context rule for type (1) constructions would cause the polarity of the negative adjective to be inverted, whereas a rule for type (2) constructions would intensify the valence of the negative adjective.

(1) *no tener nada de* (be not at all) + negative adjective:
“Ese no tiene nada de tonto/estúpido/...”
(“He’s not at all dumb/stupid/…”)

(2) *(ser) un completo* (be a complete) + negative adjective:
"Es un completo idiota" ("He's a complete idiot")

The implementation of this kind of context rules gives us greater flexibility than simply having a repository of MWEs. Without context rules, it would be very difficult to represent (and successfully process for SA) these types of MWEs, where part of them is defined by the existence of a given semantic prosody that triggers a certain polarity (e.g., adjectives denoting a negative quality).

3.3 Computing Sentiment

Sentitext returns a number of metrics in the form of an XML file which is then used to generate the reports and graphical representations of the data. The crucial information is a *Global Sentiment Value* (GSV), which is a numerical score (on a 0-10 scale) for the sentiment of the input text. Other data include the total number of words, total number of lexical words (i.e., content, non-grammatical words), number of neutral words, etc.

To arrive at the global value, a number of scores are computed. The most important is what we call *Affect Intensity*, which modulates the GSV to reflect the percentage of sentiment-conveying words that the text contains. Before we explain how this score is obtained, it is worth stressing the fact that we do not count words (whether positive, negative, or neutral): we count identified text segments that correspond to lexical units (i.e., meaning units from a lexical perspective). A segment is one of the following:

1. A single word or MWE as found in the text (or rather, its lemmatized form), either neutral or otherwise. MWEs are not marked in any special way in Sentitext's output, except for the fact that the individual words it is composed of appear in the lemmatized form in which they are stored in the database.
2. A single word or MWE identified as a sentiment-conveying lexical item, whose valence has been modified by a context rule, either by inversion or by intensification.

As we mentioned before, items in our dictionaries are marked for valence with values in the range -2 to 2. Intensification context rules can add up to three marks, for maximum score of 5 (negative or positive) for any given segment.

The simplest way of computing a global value for sentiment would be to add negative values on the one hand and positive values on the other, and then establish it by simple subtraction. However, as others have noted (e.g., Taboada et al., 2011), things are rather more complicated than that. Our Affect Intensity measure is an attempt to capture the effect that different proportions of sentiment-carrying segments have in a text. We define the Affect Intensity simply as the percentage of sentiment-carrying segments. Affect Intensity is not used directly in computing the global value for the text, however: we first adjust the upper and lower limits (initially -5 and 5). The adjusted limit or *Upper Bound* equals the initial limit unless the Affect Intensity is greater than 25 (i.e., over 25% of the text's lexical items are sentiment-carrying). Obviously, this figure is arbitrary, and has been arrived at simply by trial and error. The Upper Bound is obtained by dividing the Affect Intensity by 5 (since there are 5 possible negative and positive valence values).

A further variable needs some explaining. Our approach to computing the GSV is similar to Polanyi and Zaenen's (2006) original method, in which equal weight is given to positive and negative segments, but it differs in that we place more weight on extreme values. This is motivated by the fact that it is relatively uncommon to come across such values (e.g. "extremely wonderful"), so when they do appear, it is a clear marker of positive sentiment. Other implementations of Contextual Valence Shifters (Taboada et al., 2011) have put more weight only on negative segments when modified by valence shifters (up to 50% more weight), operating under the so-called "positive bias" assumption (Kennedy and Inkpen, 2006), i.e., negative words and expressions appear more rarely than positive ones, and therefore have a stronger cognitive impact, which should be reflected in the final sentiment score.

In our implementation, equal weight is placed on positive and negative values. However, we do not simply assign more weight to both extremes of the scale (-5 and 5), we place more weight increasingly to each value by multiplying them by different factors, from -12.5 to 12.5 in 2.5 increments[3].

[3] Our rating scale is based on a 0-10 scale, i.e., a 11-point scale, which is the most familiar for Spanish users, commonly used for grading. Sentitext outputs its rating using such a scale, and then this is converted to 5-star rating system.

What we aim to achieve with these increments is to give more weight to extreme values. For example, a text segment that has been assigned a valence of +4, which warrants a 10 factor, would end up having twice as much weight as two +2 segments (5 factor): 10x4x1=40; 5x2x2=20. The reason for this is that such extreme values are rarely found and, when they are, they invariably signal strong opinion.

The resulting method for obtaining the Global Sentiment Value for a text is expressed by Equation 1 below,

$$GSV = \frac{(\sum_{i=1}^{5} 2.5i \cdot i \cdot N_i + \sum_{i=1}^{5} 2.5i \cdot i \cdot P_i) \cdot UB}{5 \cdot (LS - NS)} \quad (1)$$

where N_i is the number of each of the negative valences found, and P_i is the equivalent for positive values. The sum of both sets is then multiplied by the Upper Bound (*UB*). *LS* is the number of lexical segments and *NS* is the number of neutral ones. Although not expressed in the equation, the number of possible scale points (5) needs to be added to the resulting score, which, as mentioned before, is on a 0-10 scale.

This formula was arrived at by trial and error and heuristics, starting from the simple addition and weighing of positive and negative valences. We found that accounting for the proportion of neutral-to-polarity segments was clearly necessary, because otherwise a fully neutral text with a few polarity segments would be analyzed as highly positive or negative, which is usually not the case. Similarly, opinion texts commonly show a number of mild opinion expressions, but if extreme values are found, they largely determines the overall opinion of the text.

Although we think that the positive bias path is worth exploring, we have not to date made comparisons with our current method. In the following section we describe previous performance tests of our system and mention some other ways in which it could be improved.

3.4 Performance

Sentitext was designed, from the beginning, with domain independence in mind. However, our first formal evaluation of the system (Moreno-Ortiz et al., 2010) was performed using a set of user reviews from the Spanish Tripadvisor website. The results of our experiment showed that good performance on a domain-specific corpus implied even better performance on general language texts.

Table 1 below shows a tendency toward low recall of negative segments, which we think may be caused by the "positive bias" effect mentioned in the previous section. In any event, these figures are more than reasonable for a sentiment analysis system.

Dataset	Precision	Recall
Global segments	0,848	0,616
Positive segments	0,838	0,669
Negative segments	0,864	0,525

Table 1: Precision and recall results in global, positive and negative segment valences.

A second evaluation (Moreno-Ortiz et al., 2011) was carried out using a greater variety of types of user reviews: movies, books and music, consumer goods, and electronics. We also introduced new features, such as a slightly modified system for calculating the GSV (modified Affect Intensity threshold) and conversion of the 0-10 score to a 5-point star-rating system. Introducing the star-rating system posed interesting questions, such as defining what is a miss and what is a hit, when comparing Sentitext's results to human ratings. Performance results were consistent with the previous evaluation, and confirmed a tendency to obtain better results for reviews of non-content objects (i.e. not books and movies), such as electronics.

A recent evaluation (Moreno-Ortiz and Pérez-Hernández, 2013) has been carried out using a large set of Twitter messages. This work was developed for the TASS workshop (Villena-Roman et al., 2013), where a double challenge was proposed by the organizers that consisted of classifying over 60,000 tweets according to their polarity in 3 levels + none and 5 levels + none, respectively. This time performance was significantly poorer, which we attribute to both the nature of the texts, and the imposed distinction between neutral and no polarity, which we find irrelevant[4]. It has served,

[4] In this scheme, no polarity means that no lexical segments carrying polarity were found, whereas neutral means that positive and negative text segments cancel each other out. Our Affect Intensity measure could easily be used for this, but such a distinction is not really useful for most applications, and usually not taken into account in the literature.

however, as proof that our GSV calculation needs to be modified in order to account for extremely short texts.

4 MWEs in Sentitext

Our criteria for the lexical representation of MWEs were largely determined by our choice of tools for basic morphosyntactic analysis, i.e., tokenization, part-of-speech tagging, and lemmatization. Freeling has the advantage of offering a very flexible MWE recognition engine.

An important advantage of using Freeling is that, being open source, the lexical resources it uses for its analysis are installed in the system in the form of text files, which allows for relatively easy editing. This is particularly useful for the acquisition of MWEs, since, although Freeling includes only a reduced set of common phrases, it is fairly straightforward to update the text file that contains them.

As for the criteria we have employed for the inclusion of an item in our database, we follow Baldwin and Kim's (2010) loose definition of *MWEhood* and typology of idiomaticity. They distinguish between lexical, semantic, pragmatic, and statistic idiomaticity, where MWEs may display one or more of those types. Some of them are idiomatic at more than one level, whereas others at one (statistical idiomaticity, in the case of collocations, for example).

4.1 Annotation schema

As of February 2013, the Sentitext MWE lexicon contains over 19,000 entries, most of which are, as expected, noun phrases. The full distribution according to syntactic category is shown in Table 2 below.

MWE Category	Number	Proportion
Noun Phrases	10,421	55%
Verb Phrases	4,768	25%
Adverbial Phrases	2,255	12%
Interjections[5]	781	4%
Adjectival Phrases	436	2%
Prepositional phrases	237	1%
Conjunctions	122	1%

Table 2: Distribution of MWE categories in the Sentitext lexicon.

[5] Interjections include idioms and other set phrases that have the form of a full sentence.

Freeling uses the EAGLES tagset recommendations for morphosyntactic annotation of corpora (EAGLES, 1996), which have consistently proved their viability in the past. The EAGLES recommendations do not impose a particular representation scheme for MWEs, and Freeling takes a simple compositional approach in which MWEs are sequences of categorized individual words.

Each morphological tag is composed of different data fields, depending on which morphosyntactic category it belongs to; some categories, like interjections, have just one field, while others have up to seven fields (e.g., verb phrases), some of which may be instantiated at runtime. For example, the morphologically invariable MWE *gafas de sol* ("sunglasses") is represented as

(3) gafas_de_sol,gafas_de_sol,NCMS000

where the tag "NCMS000" specifies that it is: N = noun, C = common, M = masculine, S = singular. Whereas in (4) below (*oso polar*, "polar bear"), the MWE is defined as a noun phrase composed of two lemmas that can be instantiated to any valid words form at runtime.

(4) <oso>_<polar>,oso_polar,$1:NC

4.2 Acquisition and valence assignment

Our *mwords* dictionary was obtained mainly from dictionaries and corpora, and the initial collection was subsequently enhanced during the extensive application testing process. We regard our acquisition of lexical items as an ongoing effort.

Prior to tagging our initial set of MWEs for Freeling, a review process was carried out to ensure that they adhered to certain varietal and statistical criteria. Castilian Spanish was taken as the standard, and very rarely are other varieties accounted for.

The most time-consuming task was obviously identifying and marking up the components of the MWEs that can be inflected. This was a lengthy process, and the results had to be checked exhaustively, since a mistake could result in an MWE not being identified in any of its forms. This was performed manually, but aided by an interface that provided a set of templates with the most commonly used morphological structures, also reducing the possibility of typing mistakes. Next we added the morphological tags, a semiautomatic process that employed RE pattern matching and then a manual check.

Valence assignment was a manual process in which lists of MWEs were rotated among team members, all native speakers of Spanish with training in Linguistics, to keep personal bias to a minimum, and hard cases were checked against corpora and decisions made on actual usage.[6] Agreement was usually high, since ambiguity and polysemy in MWEs is lower than that of individual words, especially in terms of polarity.

As mentioned in section 3.1 above, the valences assigned to the items in our database can range from -2 to 2. However, the results obtained from Sentitext's analyses can exceed these limits after the application of context rules. For example, the MWE *loco de atar* ("mad as a hatter") has a valence of -2. If we analyze the phrase *completamente loco de atar* with Sentitext, the analyzer will recognize the adjective phrase *loco de atar*, as well as the premodifying adverb *completamente*, which intensifies its valence by 2; this will result in a score of -4 for the entire phrase.

It is worth mentioning that MWEs do not require specific context rules –since their tags are the same as those used for individual words (AQ in this example), the rule that states that the adverb *completamente* to the right of an adjective intensifies its valence by 2 applies to both adjectives and MWEs tagged as such. This, which is a consequence of Freeling's annotation scheme, simplifies the acquisition and maintenance of context rules.

4.3 The role of MWEs in GSV calculation

As Table 3 shows, more than half of the MWEs in our lexicon are neutral, but this does not mean that they have no effect on the overall emotional content of texts. Neutral MWEs can be modified by words or other MWEs through the application of context rules in such a way that their polarity and/or valence is altered.

MWE Polarity	Number	Proportion
Neutral	10,823	56%
Negative	5,578	30%
Positive	2,586	14%

Table 3: Distribution of MWEs polarity in the Sentitext lexicon

For comparison's sake, our single words lexicon contains 9,404 words, all of them polarity-carrying, of which 6,907 (73%) are negative and 2,497 (27%) are positive. This is very similar to the distribution of sentiment-laden MWEs, with negative items being much more frequent than positive ones.

It is also important to note that, even when MWEs are neutral, their identification is necessary to produce the right number of lexical segments, which is taken into account in obtaining the GSV for the text.

There is yet another crucial way in which failing to identify a MWE will interfere with calculation of our GSV: if a sentiment-carrying word is part of a MWE, and that MWE is not accounted for by the *mwords* dictionary, the individual word (whose valence may or may not be correct or relevant) will be incorrectly tagged for valence.

This is particularly true of non-compositional MWEs, where the valence of the MWE cannot be deduced or calculated from the valences of the individual words that it comprises. By maintaining the MWE in the database, we eliminate the problem of having Sentitext identify parts of a MWE as individual words.

For example, the word "honor" tends to have a positive polarity, but it is also a word that frequently appears in neutral, negative and positive MWEs:

- Positive: *palabra de honor* (word of honor)
- Neutral: *dama de honor* (bridesmaid).
- Negative: *delito contra el honor* (offense against honor).

Examples of neutral individual words that appear in polarity-carrying MWEs are the following:[7]

- *darse a la bebida* (take to drink) [-2]
- *números rojos* (in the red) [-2]
- *alzamiento de bienes* (concealment of assets) [-2]
- *apaga y vámonos* (it can't be helped) [-2]
- quedarse a cuadros (be astonished) [-2]
- haber química (get on well) [2]
- ir como la seda (go smoothly) [2]

[6] The corpora used were the COE (*Corpus de Opinión del Español*), a collection of product reviews and opinion texts, compiled by our research team, and the *Corpus del Español*, a 100 million words reference corpus compiled by Mark Davies freely available for research purposes at http://www.corpusdelespanol.org.

[7] The number in square brackets marks the valence that the MWE has in our lexicon.

In all these cases no individual word that is part of the MWEs shows any polarity whatsoever, while the MWEs themselves clearly do.

It is also common to find cases in which polarity-carrying individual words are part of MWEs that have the opposite polarity:

- *amor egoísta* (selfish love) [-2]: *amor* has valence [2] as an individual word.
- ¡a buenas horas, mangas verdes! (about time, too!) [-1]: *bueno* has valence [1].
- (querer) con locura (madly in love) [2]: *locur*a" has valence [-2].
- *libre de obstáculos* (free of obstacles) [2]: *obstáculo* has valence [-1].
- morir de gusto (die of pleasure) [2]: *morir* has valence [-2].

In all these cases, not being able to account for the MWEs, would have even a stronger negative effect on the overall result.

5 Conclusion

We have shown several significant ways in which MWEs contribute to the semantic orientation of the text as a whole.

First, MWEs show a much higher proportion of polarity items (44% in our lexicon) than single lexical items do. The distribution of polarity MWEs is also very relevant. Negative MWEs make up for more than double of positive ones (30% vs. 14%), which means that the higher the proportion of MWEs there are in a text, the more likely it is for it to be negative overall.

Second, the number of lexical units they contain would alter the global calculation of semantic orientation. And, finally, the polarity of those lexical items, if computed individually, often interferes with that of the MWE as a unit. Of particular importance is the case of non-compositional MWEs, where the valence of the MWE cannot be deduced or calculated from the valences of the individual words that it comprises. This is not only a question of neutral words acquiring a certain polarity when they appear in a MWE: as we have shown, some words may also reverse their polarity from positive to negative or the other way around.

As a result, we believe that proper management and extensive coverage of MWEs in lexicon-based Sentiment Analysis systems is critical to successfully analyzing input texts.

References

Andreevskaia, A. and S. Bergler. 2007. CLaC and CLaC-NB: knowledge-based and corpus-based approaches to sentiment tagging. *Proceedings of the 4th International Workshop on Semantic Evaluation* (pp. 117–120). ACL, Prague, Czech Republic.

Atserias, J., B. Casas, E. Cornelles, M. González, L. Padró, and M. Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of the 5th ELREC International Conference*. ELRA, Genoa.

Aue, A. and M. Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. *Proceedings of RANLP 2005*. Borovets, Bulgaria.

Balahur, A., Z. Kozareva, and A. Montoyo. 2009. Determining the polarity and source of opinions expressed in political debates. *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 468–480). Springer-Verlag, Berlin, Heidelberg.

Baldwin, T. and S. Kim. 2010. Multiword expressions. *Handbook of Natural Language Processing, 2nd edition.* N. Indurkhya and F. J. Damerau (eds.) (pp. 267–292). CRC Press, Boca Raton.

Boldrini, E., A. Balahur, P. Martínez-Barco, and A. Montoyo. 2009. EmotiBlog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. *Proceedings of the 2009 International Conference on Data Mining* (pp. 491–497). CSREA Press, Las Vegas, USA.

Brooke, J., M. Tofiloski, and M. Taboada. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. *Proceedings of RANLP 2009,* (pp. 50–54). Borovets, Bulgaria.

Cruz, F., J.A. Troyano, F. Enriquez, and J. Ortega. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del Lenguaje Natural*, 41: 73–80.

EAGLES. 1996. Recommendations for the Morphosyntactic Annotation of Corpora (EAG--TCWG--MAC/R).

Erman, B. and B. Warren. 2000. The Idiom Principle and the Open Choice Principle. *Text*, *20*(1): 29–62.

Goldberg, A. B. and X. Zhu. (2006). Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. *Proceedings of the 1st Workshop on Graph Based Methods for NLP* (pp. 45–52). ACL, Stroudsburg, PA, USA.

Gupta, N., G. Di Fabbrizio, and P. Haffner. 2010. Capturing the stars: predicting ratings for service and product reviews. *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search* (pp. 36–43). ACL, Stroudsburg, PA, USA.

Hatzivassiloglou, V. and J. Wiebe. 2000. Effects of adjective orientation and gradability on sentence sub-

jectivity. *18th International Conference on Computational Linguistics* (pp. 299–305). ACL.

Jackendoff, R. 1997. *The Architecture of the Language Faculty*. MIT, Massachusetts.

Kennedy, A. and D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, *22*(2): 110–125.

Moreno-Ortiz, A., F. Pineda, and R. Hidalgo. 2010. Análisis de valoraciones de usuario de hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio. *Procesamiento de Lenguaje Natural*, 45: 31–39.

Moreno-Ortiz, A., C. Pérez, and R. Hidalgo. 2011. Domain-neutral, linguistically-motivated Sentiment Analysis: a performance evaluation. *Actas del XXVII Congreso de la SEPLN* (pp. 847–856). Huelva, Spain.

Moreno-Ortiz, A. and C. Pérez-Hernández. (2013). Lexicon-based Sentiment Analysis of Twitter messages in Spanish. *Procesamiento de Lenguaje Natural*, 50: 93–100.

Padró, L. 2011. Analizadores multilingües en FreeLing. *Linguamatica*, *3*(2): 13–20.

Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in NLP - Volume 10* (pp. 79–86).

Pang, B. and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of ACL 2005* (pp. 115–124). ACL.

Pang, B. and L. Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retreival*, *2*(1-2): 1–135.

Polanyi, L. A. and Zaenen. 2006. Contextual valence shifters. *Computing Attitude and Affect in Text: Theory and Applications* (pp. 1–10). Springer, Dordrecht.

Riloff, E. J. and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. *Proceedings of the 2003 Conference on Empirical Methods in NLP* (pp. 105–112). ACL, Stroudsburg, PA, USA.

Riloff, E., J. Wiebe, and T. Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4* (pp. 25–32). ACL, Stroudsburg, PA, USA.

Taboada, M., J. Brooks, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for Sentiment Analysis. *Computational Linguistics*, *37*(2): 267–307.

Turney, P. D. 2002. Thumbs up or Thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the ACL* (pp. 417–424). ACL, Philadelphia, USA.

Villena-Román, J., J. García, C. Moreno, L. Ferrer, S. Lana, J. González, and A. Westerski. (2013). TASS-Workshop on sentiment analysis at SEPLN. *Procesamiento de Lenguaje Natural*, 50: 37-44.