TOWARDS DISTRIBUTIONAL SEMANTICS-BASED CLASSIFICATION OF COLLOCATIONS FOR COLLOCATION DICTIONARIES

Leo Wanner: ICREA and Dept of Information and Communication Technologies, Pompeu Fabra University, Barcelona, Spain (leo.wanner@upf.edu)
Gabriela Ferraro: National ICT Australia, Canberra City, Australia (gabriela.ferraro@nicta.com.au)
Pol Moreno: School of Informatics, University of Edinburgh, Scotland (polmorenoc@gmail.com)

Abstract

Automatic acquisition of raw source material is of great aid for the compilation of dictionaries, and, in particular, of specialized dictionaries such as collocation dictionaries. The extraction of collocations from corpora has been actively worked on since the late eighties. The quality of the state-of-the-art extraction algorithms allows the lexicographers to obtain lists of collocations they can work with. However, mere lists of collocations are not sufficient. In collocation dictionaries, collocations are grouped semantically, which also presupposes a semantic classification of collocations. In this article, a distributional semantics-based model is proposed that classifies collocations with respect to broad semantic categories as encountered in dictionaries. In experiments with Spanish verb-noun and noun-adjective collocations from the lexicographic field of emotion nouns, it is shown that the use of features extracted from the context of collocations is decisive for retrieval of draft entries for collocation dictionaries.

1. Introduction

The term “collocation” as introduced by Firth (1957) and cast into a definition by Halliday (1961) encompasses the statistical distribution of lexical items in context: lexical items that form high probability associations are considered collocations. It is this interpretation that underlies, for instance, the COBUILD English Learner’s Dictionary and that is drawn upon in most works on automatic identification of collocations in corpora. However, in contemporary lexicography and lexicology, an interpretation that stresses the idiosyncratic
nature of collocations prevails. Thus, Benson (1989) states that “collocations should be defined not just as ‘recurrent word combinations’, but as ‘arbitrary recurrent word combinations’”. “Arbitrary”, as opposed to “regular”, means that collocations are unpredictable and language-specific. For instance, in English, one takes a walk, while in French, German and Italian one ‘makes’ it: faire une promenade, einen Spaziergang machen, fare una passeggiata, and in Spanish one ‘gives’ it: dar un paseo. In English, one gives a lecture, in German and Italian one ‘holds’ it: eine Vorlesung halten, tenere una lezione, and in Russian one ‘reads’ it: čitat lekciju. According to Hausmann (1984), Cowie (1994), Mel’čuk (1995) and others, a collocation is a binary idiosyncratic co-occurrence of lexical items, where the occurrence of one of the items (the base) is subject of the free choice of the speaker, while the occurrence of the other item (the collocate) is restricted by the base. Thus, in the case of take [a] walk, walk is the base and take is the collocate, in the case of high speed, speed is the base and high is the collocate, etc. It is this understanding of the term ‘collocation’ that we find reflected in general public collocation dictionaries and that we follow in this article.2

Regardless of the ongoing discussion (as witnessed, e.g., at the by now well-established biannual Electronic Lexicography in the 21st Century (eLex) conferences)3 about the future and adequate form of dictionaries in the age of the World Wide Web, general public collocation dictionaries such as the Oxford Collocations Dictionary (OCD), MacMillan Collocations Dictionary (MCD), Longman Collocations Dictionary (LCD), etc. continue to enjoy a great popularity among language learners. Intelligent electronic aids that facilitate the compilation of such dictionaries are thus still of high demand among lexicographers. A great number of works addresses since at least the late eighties one of the fundamental tasks of collocation dictionary compilation, namely the retrieval of (first of all, English) lists of collocations from which then the lexicographers can select those samples that they consider appropriate for inclusion; see, among others, (Choueka 1988, Church and Hanks 1989, Smadja 1993, Evert and Kermes 2003, Kilgarriff 2006, Evert 2007, Pecina 2008, Bouma 2010, Wible and Tsao 2010), and, in particular, A. Kilgarriff’s Sketch Engine (http://www.sketchengine.co.uk/), which proved to be instrumental for the task. However, a closer look at the dictionaries tells us that mere lists of collocations, possibly grouped with respect to the part of speech of their elements (i.e. verb + noun, adj + noun, verb + adverbial, etc.), are only half of the story. An essential feature of the microstructure of the dictionaries is that the elements that co-occur with the headword are grouped according to semantic criteria, such that learners can easier choose among semantically similar (or even quasi-synonymous) elements to express a given meaning. Consider, for instance, a
fragment of the entry for the noun grant in the OCD, where ‘|’ separates the
element groupings:

- ADJ large, substantial | small | full | ...
- VERB + GRANT apply for | be eligible for, qualify for | get, obtain, receive | award (sb), give (sb), make (sb), offer (sb), provide (sb with) | refuse (sb) | cut...

Large [grant] and substantial [grant] are semantically similar enough to form a group that is contrasted to small [grant]; full means ‘covering the entire expenses’ and is thus rather different.4 Apply for means ‘present formally the desire to obtain a grant’, while both be eligible for [a grant] and qualify for [a grant] can be glossed roughly as ‘fulfill the formal conditions to be considered as grantee’. Get, obtain, and receive [a grant] are also quasi-synonymous and can thus be used alternately to express ‘become recipient of a grant’. And so on.

The big advantage of the semantic grouping of collocates in the entry for a headword, compared to the display of a mere list, is for a learner not only that they can immediately see which collocates substantially differ (e.g. that full [grant] is by far not the same as large [grant]), but also which collocates are similar enough to be used alternatively (as, e.g. award, give, make, offer [a grant], and provide sb with [a grant]; or get, obtain, and receive [a grant], already cited above). Especially less advanced learners are in need of this information.

However, automation of semantic grouping (or classification) of the relations between elements of the extracted collocations has hardly been paid attention to so far in computational lexicography. Only a few proposals address the problem (Moreno et al. 2013, Wanner 2004, Wanner et al. 2006, Chung-Chi et al. 2009, Kolesnikova and Gelbukh 2012). All of them, except Moreno et al. (2013), who also use contextual features, draw exclusively upon WordNet (Fellbaum 1999) or EuroWordNet (Vossen 1998) features of the collocation elements, which limits their application to languages with large coverage WordNets. Furthermore, they classify either with respect to the rather fine-grained semantic typology of lexical functions (LFs) as defined in the Explanatory Combinatorial Lexicology (Mel’čuk 1996) or with respect to generic semantic categories like ‘goodness’, ‘heaviness’, ‘measures’, etc. See, for instance, (Wanner 2004, Wanner et al. 2006, Kolesnikova and Gelbukh 2012) for the first and (Chung-Chi et al. 2009) for the second. Moreno et al. (2013) classify with respect to both LFs and generic semantic categories. But collocation dictionaries do not present collocations in terms of LFs or generic semantic categories; as the entry for grant above shows; they choose a typology that is in between.
In this article, we present a proposal for the semantic classification of the collocates of the headwords that

(i) exploits the notion of distributional semantics (Lenci 2008) in that it uses for classification the distributional features of the context in which the collocations are used in the corpus, and

(ii) draws upon categories that correspond in their level of detail to the categories used in general public collocation dictionaries: 16 broad semantic categories for verbal collocates in verb + noun collocations and 5 categories for adjectival collocates in noun + adjective collocations.

So far, we experimented with Spanish collocations from the lexicographic field of emotion nouns. However, as Wanner (2004)’s experiments show, classification experiments on one lexicographic field can be extrapolated with an acceptable loss of accuracy to the vocabulary in general.

In the next section, we briefly introduce the semantic categories that we use in our collocate classification experiments. In Section 3, we outline the foundations of our classification, and in Section 4 we describe the experiments that we carried out to assess the performance of our proposal and discuss their outcome. Section 5 presents the related work, before Section 6 summarizes the insights that we obtain and sketches the directions of our future work on this topic.

2. Semantic Categories of Collocations

The base and the collocate of a collocation are always linked by a semantic relation. This relation is often of abstract nature, such that it is encountered in a large number of collocations. For instance, the same relation can be said to hold between speech and deliver, suicide and commit, step and take, etc. It is the same in the sense that deliver, commit, and take contribute to their respective base the same semantic features. Lexemes that are by no means (quasi-) synonyms when taken in isolation may become quasi-synonymous when used as collocates of the same base; consider, for instance, deliver, hold, give, and make in co-occurrence with speech, or take, make and do in co-occurrence with step. This observation led Mel’čuk define lexical functions (LFs); see (Mel’čuk 1996): each LF captures one type of collocation relation. As mentioned in Section 1, LFs have been used as underlying typology for automatic collocation classification by Wanner (2004) and Kolesnikova and Gelbukh (2012), among others.

However, for learners, dictionaries that list LF instances in their entries (see the French Dictionnaire explicatif et combinatoire du français contemporain I–IV (Mel’čuk 1984–1999)) are rather difficult to consult and comprehend. To remedy this drawback, Alonso Ramos et al. (2010) and Mel’čuk and
Polguère (2007) use glosses instead of LF names; consider a fragment of the entry for Sp. Miedo ‘fear’:

miedo de individuo X a hecho Y (por individuo Z)

... 

sentir ~ abrigar [(ART) ~], albergar [(ART) ~], conocer [(ART) ~], conocer [(ART) ~], experimentar [(ART) ~], pasar [~],

sentir un ~ intenso acojonarse, cagarse [de ~], cagarse, morirse [de ~]

empezar a sentir ~ cobrar [~ a Y], coger [~ a Y]

continuar sintiendo ~ conservar [ART ~], seguir [con ~]

dear de sentir ~ perder [ART ~]

estar dominado por el ~ sucumbir [a ART ~]

...

in Alonso Ramos’ Diccionario de Colocaciones del Español (DiCE)⁵, and a fragment of the entry for Fr. FACTURE ‘bill’ in Mel’cˇuk and Polguère’s Lexique actif du français (LAF) dictionary:

...[X] faire une F. dresser, établir, faire, rédiger [ART ~] [F.]
faite pour Z [au nom de] N₂, F. a pour montant W être [de N₃], s’élever, se monter [à N₃], [X] transmettre une F. à Z envoyer, présenter, remettre [ART ~ à N₄]...

Similarly, MCD uses glosses, but they cover broader categories; consider a fragment of the entry for the noun IMPACT:

• v+N
  ♦ have an impact create, have, make
  ♦ reduce impact absorb, cushion, lessen, limit, minimize, mitigate, reduce, soften
  ♦ examine impact analyse, consider, examine, explore, investigate, understand
  ♦ measure impact assess, calculate, estimate, evaluate, measure, quantify

...

Unfortunately, to the best of our knowledge, no guidelines are published on how collocates are to be grouped into these categories, or, in other words, how each of these categories is defined; neither is a complete category list accessible. Therefore, in order to obtain a coherent collocate typology that matches the level of detail used in general public collocation dictionaries, we defined ad hoc five generic disjoint semantic categories of sample collocates of adjective + noun
collocations and 16 categories of the collocates of sample verb + noun collocations; consider Tables 1 and 2 for the glosses of the categories and their illustration by examples collected manually.\(^6\) In order to facilitate a better understanding of the categories and signal that they are domain-independent, we also include into Tables 1 and 2 examples that are not from the emotion domain, which is the domain on which we carried out our experiments (see Section 4).

The categories group collocations with respect to semantic and syntactic criteria. In contrast to, for instance, LAF and MCD, which use entry-specific glosses, the glosses of our categories are intended to be entry-independent, as they are, for example, in DiCE, where they are lexicographic field-specific. Thus, the category ‘V1’ in Table 2 is glossed as ‘begin with L’ (where ‘L’ is the headword (the base) or simply ‘begin’ of the entry). Our categories can be also viewed as generalizations over concrete LFs. For example, V1 covers the LFs ‘PreparReall’ and ‘IncepOper1’; ‘V2’, which is glossed as ‘execute’, covers the LFs CausPredPlus, Magn + Oper1, Oper1, ContOper1, Reall, PredA1Manif, Perm1Manif; and so on; see (Mel’čuk 1996) for a detailed introduction to LFs.

Consider the verbal collocation fragment of the entry for Sp.MIEDO ‘fear’ in DiCE, organized in terms of our generalized categories and adapted to the layout of MCD:

- \(v+N\)
  - \(\text{\begin{underline}{\textbf{begin}}\end{underline}}\) cobrar, coger
  - \(\text{\begin{underline}{\textbf{experience}}\end{underline}}\) albergar, conocer, experimentar, pasar, sentir, tener, aumentar, abrigrar, acojonarse, cagarse [de \(\sim\)], morirse [de \(\sim\)], conservar [ART \(\sim\)], seguir [con \(\sim\)], sucumbir [a ART \(\sim\)]
  - \(\text{\begin{underline}{\textbf{manifest}}\end{underline}}\) confesar [ART \(\sim\)], demostrar, manifestar, mostrar
  - \(\text{\begin{underline}{\textbf{get involved into}}\end{underline}}\) llenar [a X de \(\sim\)]
  - \(\text{\begin{underline}{\textbf{cause the existence of}}\end{underline}}\) amedrentar [a X], amilnar [a X], atemorizar [a X], causar [\(\sim\) a X], dar [\(\sim\) a X], despertar [\(\sim\) a/en

\(|\sim\) a/en

<table>
<thead>
<tr>
<th>Category</th>
<th>Semantic Gloss</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>‘intense L’</td>
<td>heavy rain, strong tea</td>
</tr>
<tr>
<td>A2</td>
<td>‘weak L’</td>
<td>light rain, gentle touch</td>
</tr>
<tr>
<td>A3</td>
<td>‘positive L’</td>
<td>valuable aid, convincing argument, lucky winner</td>
</tr>
<tr>
<td>A4</td>
<td>‘negative L’</td>
<td>controversial proposal, forced laugh, flimsy excuse</td>
</tr>
<tr>
<td>A5</td>
<td>‘attributed to s.o.’</td>
<td>visible anger, blind [with] fury, infectious laughter</td>
</tr>
</tbody>
</table>

\(\text{\begin{underline}{\textbf{Table 1:}}\end{underline}}\) Semantic categories of adjective (a) + noun(n) collocations (‘L’ stands for “base”)

\(\text{\begin{underline}{\textbf{Table 2:}}\end{underline}}\) Examples

\(|\sim\) a/en
Table 2: Semantic categories of verb(v) + noun(n) collocations (‘L’ stands for “base”)

<table>
<thead>
<tr>
<th>Category</th>
<th>Semantic Gloss</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>‘begin L’</td>
<td>fall [in] love, pluck up courage, offer friendship</td>
</tr>
<tr>
<td>V2</td>
<td>‘perform’/ ‘carry out’/ ‘experience L’</td>
<td>apply pressure, secure affection, cherish hope, give [a] talk, pronounce [a] discourse</td>
</tr>
<tr>
<td>V3</td>
<td>‘manifest L’</td>
<td>show panic, confess love, admit failure</td>
</tr>
<tr>
<td>V4</td>
<td>‘cause the involvement in L’</td>
<td>instigate rebellion, give love, nominate [for] presidency</td>
</tr>
<tr>
<td>V5</td>
<td>‘cause the existence of L’</td>
<td>give hope, found [an] association, cause [an] accident</td>
</tr>
<tr>
<td>V6</td>
<td>‘cause to be target of L’</td>
<td>provoke hatred, draw attention, trigger compassion</td>
</tr>
<tr>
<td>V7</td>
<td>‘undergo L’</td>
<td>receive [a] hug, undergo [an] operation, suffer [from an] attack</td>
</tr>
<tr>
<td>V8</td>
<td>‘suppress manifestation of L’</td>
<td>mask pain, suppress comment, impede [an] attempt</td>
</tr>
<tr>
<td>V9</td>
<td>‘free oneself of L’</td>
<td>withdraw [from] competition, quit membership, give up [the] role</td>
</tr>
<tr>
<td>V10</td>
<td>‘cease or diminish L’</td>
<td>terminate [a] dispute, break off [a] relationship, weaken [the] grip</td>
</tr>
<tr>
<td>V11</td>
<td>‘increase L or its manifestation’</td>
<td>increase pressure, spread rumor, hone skills</td>
</tr>
<tr>
<td>12</td>
<td>‘L begins to affect so’</td>
<td>love overwhels [s.o.], panic spreads, doubts come up</td>
</tr>
<tr>
<td>V13</td>
<td>‘L involves s.o.’</td>
<td>hint comes [from], invention goes back [to], operation is carried out [by]</td>
</tr>
<tr>
<td>V14</td>
<td>‘L continues’</td>
<td>rumor persists, defense withstands, debate goes on</td>
</tr>
<tr>
<td>V15</td>
<td>‘L ceases or diminishes’</td>
<td>support wanes, protest winds down, deadline expires</td>
</tr>
<tr>
<td>V16</td>
<td>‘L concerns sth’</td>
<td>dispute be over sth, disagreement concern sth, operation be on sth</td>
</tr>
</tbody>
</table>

X], infundir [~ a/enX], inspirar [~ a X], intimidar [a X], meter [~ (en el cuerpo) a X], producir [~ a X], provocar [~ a X], sembrar [(ART) ~ en X]

◆ suppress manifestation of ~ afrontar [ART ~]
◆ free onself of ~ perder [ART ~]
put an end to or diminish ~ sacurdirse [ART ~], superar [ART ~], vencer [ART ~], apaciguar [ART ~], aplacar [ART ~], atenuar [ART ~], debilitar [ART ~], disminuir [ART ~]

increase ~ aumentar

~ begins to affect so surgir [en X], aparecer [en X], entrar [a X], apoderarse [de X], asaltar [a X], invadir [a X], propagarse [en X]

~ affects so reinar [en X] (Func1); atenazar [a X], bloquear [a X], corroer [a X], frenar [a X], paralizar [a X], sacudir [a X], dominar [a X]

~ continues seguir

~ ceases or diminishes apagarse, cesar, desaparecer, desvanecerse, disiparse, atenuarse, debilitarse, disminuir

The goal of our semantic classification of collocations described in Sections 3 and 4 below is to obtain entries of this kind. Note, however, that the categories we use are sample categories that shall merely serve the purpose of illustration of our classification techniques. They can be adapted to the needs of the users or redesigned following more rigorous lexicographic guidelines, without, as we believe, that the techniques need to be modified.

3. Semantic Classification of Collocations

3.1 Basic Considerations

The last decade saw an increasing role of distributional semantics in lexically-oriented computational linguistics research. The basic assumption of distributional semantics (see (Lenci 2008) for a detailed review) is that the sense of a lexical item is determined by the context it occurs in. Thus, in

(1)(a) Join one of the most prestigious and long-established golf clubs in England.

(1)(b) Authorities in Florida said a spring breaker’s father allegedly struck cars with a golf club.

the context tells us that the meaning of CLUB in (1)(a) is different from the meaning of CLUB in (1)(b). The contexts of CLUB in (1)(a) and (1)(c)–(f) below let us furthermore deduce that in all of them CLUB stands for ‘organization’.

(1)(c) This guide will help you understand what it takes to join a club.

(1)(d) The club history is well known, but it’s worth remembering again.

(1)(e) Community clubs are generally open to anyone, while company clubs are often limited to employees of that company.

(1)(f) The prestigious New York Athletic Club, one of the world’s premier private clubs has been founded in 1868.
Intuitively, the context predicts the meaning not only of isolated words, but also of words in co-occurrences, that is, collocations. Consider, for instance, (2)(a)–(d):

(2)(a) Dr. Miller will * a free public lecture.
(2)(b) The GM is set to * a chess lecture and play simultaneous games...
(2)(c) If you are * lectures in a course at any level, don’t rely on notes.
(2)(d) It is extremely difficult to * a good lecture without prior organisation.

‘*’ stands for a word that has been omitted, but whose meaning (namely ‘carry out’) in co-occurrence with LECTURE can be deduced from the context. As speakers of English, we also know that ‘carry out’ in co-occurrence with LECTURE can be lexicalized as deliver, give or hold. In co-occurrence with TALK, we would simply pick give:

(3)(a) In 2005, Mark Zuckerberg * a talk and barely anyone showed up.

That is, if speakers of English are able to deduce from the context the meaning of the co-occurrence even if the collocate is missing, it should be also possible to achieve this automatically using machine learning techniques. In what follows, we present a formal view on the semantic classification of collocations (or, more precisely, of collocates) based on their context.

3.2 A more formal view on the semantic classification of collocations

The task of assigning a semantic category label (such as ‘carry out’ in the examples (2)(a)–(d), and (3)(a) above) to the collocate depending on the context of the collocation can be interpreted as a multi-class supervised learning task, where the context of a number of training samples for each category is captured in terms of features. The “learned” model reflects the composition of the most distinctive features for each category, such that when a new sample is classified, its features are compared with the distinctive features of the categories, to choose the most proximate one. Different supervised learning models, ranging from neural networks over Bayesian statistics to kernel-based algorithms, can be used; see, for instance, (Alpaydin 2010). Among the most popular models in Natural Language Processing applications are the kernel-based Support Vector Machines, SVMs (Vapnik 1998). We also use SVMs in our experiments, comparing them with a Bayesian model.

Assuming that various types of linguistic features should compose the context of a collocation in a given sentence (and contribute thus to the distributional semantics of it), we have considered the following features:
Lexical features: bigrams of all tokens in the sentence + base + collocate + base-collocate bigram;

POS-features: Part of Speech (POS) of the base + POS of the collocate + POS of the tokens in the windows of size 2 to the left and to the right of the base and the collocate + POS-trigrams of the POS of the base and the POS of its immediate left and right context + POS-trigrams of the POS of the collocate and the POS of its immediate left and right context;

Morphological features: gender, number, person of the base + number, person, tense, and mode of the collocate + POS bigrams of the syntactic dependents of the base and the POS of the base + POS bigrams of the POS of the syntactic head of the collocate and the POS of the collocate + POS bigrams of the POS of the collocate and the POS of all its remaining dependents;

Syntactic dependency features: syntactic relation between the collocate and the base + syntactic relation between the collocate and its head + syntactic relations between the collocate and its remaining dependents + syntactic relations between the base and its dependents.

To represent collocation instances, the features are, in general, ordered; each instance is represented by a vector of feature values (empty values, such as, for example, the mode of an adjectival collocate, are marked by a “/C0”). For instance, the collocation Sp. *gran veneración*, lit. ‘great veneration’ is represented by the following vector:7

le escuchaban, escuchaban con, con gran, gran veneración, veneración, gran, gran veneración, noun, adjective, verb, preposition, adjective, verb preposition adjective, determiner, determiner preposition adjective, adjective, feminine, singular, 3rd person, singular, 1st person, /C0, /C0, /C0, noun, verb preposition, noun, noun, adjective, adjective modifier, direct object, adjective modifier, /.

During the classification, each co-occurrence (each collocation instance) is assigned to one of the N classes. The goal of the classifier is to construct a function that, given a new data instance (a new collocation), predicts the class to which the new instance belongs.

4. Experiments

In what follows, we present the setup of the experiments carried out to validate the collocation classification strategy proposed above and discuss its outcome.
Our experiments are based on a corpus of Spanish annotated with LFs. The corpus has been compiled as part of the electronic dictionary of collocations of Spanish, DiCE (Alonso Ramos et al. 2010), http://www.dicesp.com, from the corpus of the Spanish Language Academy. The DiCE corpus is composed of sentences that illustrate the use of LF instances listed in the dictionary; for each LF instance, at least one sample sentence is available. In total, the DiCE corpus contains 30,643 sentences. Since we draw in our experiments only upon a subset of collocations, only the corresponding sample of the corpus is used. The distribution of the collocations in our corpus sample across the classes in Tables 1 and 2 is depicted in Figure 1. The figure shows that the distribution is highly unbalanced. In the case of verb–noun collocations, the majority class contains more than 1,800 instances, while the minority class has less than 50 instances. Similarly, in the case of noun + adjective collocations, the majority class contains about 1,800 instances, and the minority class less than 100. This reflects the varying commonality of collocations of different classes in language use.

We carried out two main experiments on the classification of verbal and adjectival collocates with respect to our typologies with the SVM and Naïve Bayes classifiers from the WEKA machine learning environment (http://www.cs.waikato.ac.nz/ml/weka/). The POS, morphological and syntactic dependency features of the LF instances and their contexts were obtained by Bohnet (2009)’s dependency parsing toolkit. To assess the impact of the semantic features, we ran the SVM classifier on feature sets that did not include any semantic features as well as on feature sets that contained, among others, features from EuroWordNet (synset members of both adjectival and verbal collocates and Top Concept paths of the verbal collocates, for instance, ‘Dynamic’ – ‘Stimulating’ – ‘Mental’ – ‘Experience’ for despertar ‘[to] arouse’ or ‘Dynamic’ – ‘Location’ – ‘Existence’ for disiparse ‘dissipate’); see first column of Table 4.
The experiments had two main objectives. Firstly, to explore whether distributional features extracted from the context of collocations suffice to classify collocations with a high accuracy; that is, whether we can dispense with using external semantic resources of the type of WordNet. Secondly, to assess whether the existing interdependence between the individual features is of relevance or whether it suffices to consider the features as such.

In the first experiment, WEKA’s SVM classifier was used together with the LibSVM implementation. SVM captures all features and their interdependencies. In the second experiment, WEKA’s Naïve Bayes (NB) classifier with its default parametrisation was used. NB treats the features as being independent.

A linear kernel was chosen to generate the SVM models since linear kernels proved to be adequate for text classification tasks, which usually need to cope with a high number of features (Abraham and Thampi 2013). For the multi-class classification with SVM, we used the One-vs-All method, where N different binary classifiers are learned to distinguish between the members of a class from the members of the other classes. When classifying a new instance, the N classifiers are run, and the label of the classifier which outputs the largest value is chosen. To account for the highly unbalanced class distribution, we used the Synthetic Minority Oversampling Technique with Different Costs (Akbani et al. 2004). As baseline for both experiments, we used the majority class.

4.2 Experiment results

To assess the performance of our classifier models, we used a 10-fold cross-validation scheme with 10 repetitions. In other words, for each experiment run, we used 9/10 of the sample material (cf. Figure 1 for its distribution) for training and 1/10 for testing, picking each time a different 1/10. Figures 2 and 3 show the performance of the multi-class classification of verbal + noun and noun + adjective collocations, respectively, when POS, morphological, and syntactic and lexical features were used. For both experiments, precision (the share of correctly classified positive instances among all positive instances in the system output) and recall (the share of correctly identified positive instances among all instances that should have been identified as positive) are used to evaluate the classification with respect to the individual classes.

The precision of the SVM classification in the case of verb-noun collocations is rather high. Only for V9 and V16 it is lower than 0.6, which is likely due to the heterogeneity of the contextual features of their instances and, in particular, the reduced size of the training data (recall that for V16, only a small sample is available). The SVM recall is even more stable, with an oscillation around 0.8 and only two outliers for V7 and, again, V16. The performance of NB is much more erratic. For V6, its precision reaches 0.9, while for V16 de facto no
instances have been recognized at all. Its recall is even poorer, including for classes for which SVM provides a good coverage. Overall, the average F-score (the harmonic means between precision and recall) achieved with SVM for verb-noun collocation classification is 0.80 and for NB 0.60.

In the case of the noun-adjective collocation classification, an analogous picture can be observed. In the average, for SVM both precision and recall are higher for noun-adjective collocation classification than for verb-noun classification. In the case of NB, the performance again varies significantly, for recall more than for precision.

Table 3 shows the performance of the majority class baselines, also in terms of precision and recall.

The exploration to what extent lexical features (i.e. word tokens) and semantic features from EuroWordNet influence the accuracy of the classification with SVM led to the results shown in Table 4.
4.3 Discussion

In general, our experiments demonstrate that an automatic classification of verb-noun and noun-adjective collocations with respect to collocation typologies as used in collocation dictionaries is feasible using only contextual features of the collocations. Or, in other words, that the use of external semantic resources whose existence is ensured only for a small subset of languages is not required. Table 4 shows that the performance is, in fact, best if we use all available features, including the semantic ones. But adding semantic features does not lead to a significant rise of accuracy. The use of lexical features is of more importance (cf. the last line in Table 4, which shows the performance when no lexical features are used). This is especially the case for verb-noun collocations. Our hypothesis is that this is likely to be due to the restricted variation of the contexts of the verb-noun collocations in our corpus sample. For adjective-noun collocations, the impact of lexical features is lower. On the other hand, the importance of semantic features seems to be reduced because of the rather abstract nature of the EuroWordNet features we used. Further experiments would be required to validate this hypothesis.

We also observe that using only features of the collocation elements themselves (‘Simple’ in Table 4) leads to poor results for both verb-noun and noun-adjective collocations. However, for noun-adjective collocations, the context is not as important as for verb-noun collocations (cf. the performance with the ‘Simple’ feature set). We assume that this is because of the limited variety of the adjectival collocate lexemes in our corpus sample (recall that our experiments

Table 4: Average percentage of correctly classified instances with different feature sets using SVM (‘Simple’: POS of the base, POS of the collocate, and syntactic relation between the collocate and the base; ‘Simple + Lemmas’: Simple and all the lemmas in the sentence; ‘Simple + Lemmas + Synt’: all the POS, lemmas and syntactic relations in the sentence; ‘Simple + Lemmas + Sem’: Simple, all the lemmas and semantic features in the sentence; ‘Simple + Synt + Sem’: Simple, all syntactic relations and semantic features in the sentence.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>v + n</th>
<th>a + n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Simple</td>
<td>28.68%</td>
<td>64.29%</td>
</tr>
<tr>
<td>2. Simple + Lemmas</td>
<td>80.29%</td>
<td>87.91%</td>
</tr>
<tr>
<td>3. Simple + Lemmas + Synt</td>
<td>80.63%</td>
<td>93.49%</td>
</tr>
<tr>
<td>4. Simple + Lemmas + Sem</td>
<td>80.71%</td>
<td>94.09%</td>
</tr>
<tr>
<td>5. Simple + Synt + Sem</td>
<td>62.94%</td>
<td>92.01%</td>
</tr>
</tbody>
</table>
are done on an emotion vocabulary corpus). Again, further experiments with other corpora will be needed to verify our assumption.

The significant performance difference between SVM and NB (especially in terms of recall) shows that it is essential to capture the interdependence (or correlation) between the features: as pointed out above, SVM allows us to model this interdependence, while NB assumes that the features are independent.

Table 5 summarizes the most prominent confusions between verbal collocate classes. It shows that, in particular, collocates of V2 and V7 are confused. This is likely due to the same syntactic dependency structure of V2 and V7 collocations (such that the POS and morphological contexts are also likely to be similar).

It can be also observed that collocation instances are most often misclassified as V2. This might be due to the fact that the classification algorithm prefers V2 because it is the majority class (and the probability of misclassification is thus lower than it would be with classes that contain only a few instances). This tendency is likely to be further increased by the One-vs-All method we use. The confusion matrix furthermore reveals that half of the instances from the minority class V16 has been misclassified. This is not surprising, given that V16

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
<th>V8</th>
<th>V9</th>
<th>V10</th>
<th>11</th>
<th>V12</th>
<th>V13</th>
<th>V14</th>
<th>V15</th>
<th>V16</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>112</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V2</td>
<td>1729</td>
<td>13</td>
<td>4</td>
<td>34</td>
<td>1</td>
<td>38</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>18</td>
<td>0</td>
<td>16</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V3</td>
<td>7</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>14</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V4</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>124</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V5</td>
<td>1</td>
<td>38</td>
<td>3</td>
<td>2</td>
<td>396</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V6</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>106</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V7</td>
<td>1</td>
<td>82</td>
<td>11</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>169</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V8</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>295</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V9</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>134</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>V10</td>
<td>1</td>
<td>28</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>595</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>V11</td>
<td>0</td>
<td>36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>181</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>V12</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>252</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>V13</td>
<td>1</td>
<td>32</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>5</td>
<td>273</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V14</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>125</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V15</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>145</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V16</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>
contains only 18 instances (cf. Figure 1) and the learning effect can be only limited. Here, the mayor confusions are with V2 and V13.

In general, phrasal collocates of the type pain~be caused by, anger~come from, etc. are captured worst by both SVM and NB. This is certainly due to their syntactic and semantic complexity and also lower frequency.

5. Related Work

As already pointed out in the Introduction, there is only a limited number of works dealing directly with the semantic classification of collocates. Most similar to our proposal is (Moreno et al. 2013). They also use contextual features to classify collocations. However, while they classify with respect to lexical functions (LFs) or very broad semantic categories (five in total), we classify with respect to categories of the kind as used in general public collocation dictionaries. Wanner et al. (2006, 2005) and Gelbukh and Kolesnikova (2012) classify collocations with respect to LFs only. Instead of distributional semantics, they use exclusively EuroWordNet features of the collocation elements, without considering the context of the collocations in the corpus. Moreno et al. (2013) report an average F-score of 0.8 for LFs and an average F-score of about 0.9 for semantic category classification. Wanner et al. (2006) achieve in the average an F-score of 0.84 when classifying with the Nearest-Neighbor (NN) learning model emotion verb-noun collocations with respect to five common LFs. Wanner et al. (2005) reach with the NN and Tree-Augmented Bayesian Network (TAN) classification techniques that also use semantic features from EuroWordNet for four selected noun-adjective LFs an average precision of 0.67 (NN) respectively 0.66 (TAN) and an average recall of 0.66 (NN) respectively 0.7 (TAN). Given that these are figures achieved on emotion vocabulary, as in our experiments, they are directly comparable to the figures in our experiments, and we can thus conclude that it pays off to draw upon the contexts of collocations. Kolesnikova and Gelbukh (2012) achieve with their rule-based models an average F-score of 0.75 when classifying over the LF typology. Furthermore, in contrast to our experiments, this average does not reflect the classification of collocations with one single model, but, rather, the average of the best scores achieved with one of the twelve models from a series of models they use for each individual LF. The figures in our experiments thus suggest that for generalized verb-noun collocation classification the use of distributional context features of collocations in corpora is more promising than the exclusive use of EuroWordNet features of the collocation elements.

We already argued above that apart from the use of semantic features from external resources for model training, both Wanner et al.’s and Gelbukh & Kolesnikova’s approaches suffer from the fact that they use LFs as such and are thus difficult to port. The annotation of corpora in other languages with LFs would be by far more expensive than their annotation with the type of
semantic categories we use since LFs require much more subtle distinctions, with a higher potential to err and to disagree among the annotators. Therefore, to support the compilation of material for collocation dictionaries from corpora without depending on existing external lexical resources and on LF-annotation, we must be able to group the collocates of a given headword automatically in terms of categories used in general public collocation dictionaries, as suggested in this article.

Chung-Chi et al. (2009)’s proposal differs significantly from ours because, on the one hand, they interpret the notion of ‘collocation’ in the sense of Firth (1957), while our interpretation adheres to that introduced by Hausmann (1984) and others, and, on the other hand, they classify the relations between collocate elements with respect to very coarse-grained semantic categories of the type ‘goodness’, ‘heaviness’, ‘measures’, etc., which appear to be too generic for use in collocation dictionaries. Their assessment performance is based on the evaluation by two human judges, with an average precision of 0.76 and an average recall of 0.69.

6. Conclusions and Future Work

We presented novel work on the semantic classification of collocates in verb + noun and noun + adjective collocations in Spanish. Our work differs from most of the previous works in this area in two respects. Firstly, we use distributional context features of the collocation elements rather than EuroWordNet features of the collocation elements only. In addition, we also show that adding EuroWordNet features does not lead to a significant increase of the performance. This makes our proposal more robust since we do not depend on the availability of external language resources. Secondly, we classify over a generic collocation typology of the kind used for grouping collocations in the entries of general public collocation dictionaries rather than over the typology of lexical functions or a typology of very broad semantic categories. It has been argued that the former are too detailed and the latter too coarse grained for the standard user of such dictionaries. The output of our classification can thus be directly used as material for the compilation of collocation dictionaries. Given the extreme work load related to the production of collocation dictionaries, this is an important aid. In the future, we plan to carry out experiments with material in other languages, which also includes the annotation of corpora with collocation class tags, to test other classification models, and to review with lexicographers the collocation categories we use.

Acknowledgements

The presented work has been partially funded by the Spanish Ministry of Economy and Competitiveness (MINECO) under the contract number FFI2011-30219-C02-02.
Notes

1 Chung-Chi et al. (2009) also use this interpretation in their work on collocation classification.

2 This is not to say that the two main interpretations of the term “collocation”, the distributional and the idiosyncratic one, are disjoint, that is, necessarily lead to a different judgment with respect to the collocation status of a word combination. Just the contrary: two lexical items that form an idiosyncratic co-occurrence are likely to occur together in a corpus with a high value of point-wise mutual information (Church and Hanks 1989).

3 The discussion goes from the question on the adequate macrostructure of the dictionary in the light of the necessity to display the use of a lemma in reference corpora to the question whether dictionaries have a future at all.

4 Following the tradition in lexicography and linguistics, we mark semantic glosses (meanings) by single quotes.

5 The format shown here is different from the DiCE format, which lists one collocation instance per line.

6 Since our goal was to provide a collocation classification procedure rather than an exhaustive collocate category typology, we did not aim to cover all LFs discussed in (Mel’čuk 1996). It is also important to note that our generalization is an ad hoc generalization. A thorough feature study might lead to a different, more homogeneous generalization, which would lead to an even better classification.

7 When not needed, as, for instance, in the case of the Naïve Bayes classifier, the vector representation is dismantled and considered as a bag of feature–feature value pairs.

8 Many thanks to the director of DiCE, Margarita Alonso Ramos, for providing the corpus to us.

9 This number reflects the size of the DiCE corpus at the time of our experiments.

10 This parsing toolkit contains a lemmatizer, tagger and a syntactic dependency parser, which performed best on Spanish in the CoNNL 2009 shared task.

11 The “top concepts” in EuroWordNet are abstract language-independent semantic features that characterize in more detail the semantic field (or “base concept”) labels assigned to the individual lexemes (i.e. disambiguated words).

12 The choice between the One-vs-All and All-vs-All methods for multi-class classification has been mainly computational since One-vs-All is faster and more memory-efficient (Ryan and Aldebaro 2004).

References

A. Dictionaries


B. Other literature


