Can we determine the semantics of collocations without using semantics?

Pol Moreno[1], Gabriela Ferraro[2], Leo Wanner[3]

[1]School of Informatics, University of Edinburgh,
Appleton Tower, 11 Crichton Street, Edinburgh EH8 9LE
[2]NICTA (National ICT Australia), London Circuit 7, Canberra City ACT 2601, Australia.
[3]Institució Catalana de Recerca i Estudis Avançats (ICREA) and
Department of Information and Communication Technologies, Pompeu Fabra University,
Roc Boronat, 138, 08018 Barcelona
E-mail: polmorenoc@gmail.com, gabriela.ferraro@nicta.com.au, leo.wanner@upf.edu

Abstract

The extraction of collocations from corpora has been actively worked on since the late eighties. However, so far, an important task of collocation processing, namely the semantic interpretation of the collocate, did not receive much attention, although the semantics of a given word when used as collocate very often varies from the semantics of the same word when used in a free co-occurrence. In this paper, we tackle this problem. Our aim is the automatic semantic disambiguation of collocates, or, more precisely, the classification of collocations with respect to the typology of lexical functions (LFs) introduced in the Explanatory Combinatorial Lexicology. The two main questions underlying our research that seeks a scalable solution independent of any external semantic resources are: (i) how well can we semantically classify collocates without the use of explicit semantic features; and (ii) to what extent can we dispense with explicit lexical information when classifying collocates. To answer these two questions, we carried out machine learning experiments in which we used different training feature sets and LF typologies of different abstraction. So far, we worked on Spanish verb-noun and noun-adjective collocations from the lexicographic field of emotion nouns. However, our approach is, strictly speaking, language-independent.

Keywords: collocations; semantics; lexical functions; classification

1. Introduction

The recognition and extraction of collocations from corpora has been actively worked on since the late eighties (e.g. Choueka, 1988; Church and Hanks, 1989; Smadja, 1993; Evert and Kermes, 2003; Kilgarriff, 2006; Evert, 2007; Pecina, 2008; Bouma, 2010; Wible and Tsao, 2010).[1] However, so far, an important task related to collocation recognition, namely the semantic disambiguation (or classification) of the collocate,[2]

[1] Not all of these works use the term "collocation", but all of them nonetheless extract co-occurrent word combinations.

[2] Here and henceforth, we use the terminology as introduced by Hausmann (1989): the *base* is the semantic head of the collocation and the *collocate* is its dependent. Thus, in the collocation *strong tea*, *tea* is the base and *strong* is the collocate; in *take a rest*, *rest* is the base and *take* is the collocate, etc.

has received only very limited attention by the main stream research in the field. It is important to disambiguate the collocate because the semantics of a given word when used as a collocate very often differs from the semantics of the same word when used in a free co-occurrence. For instance, the meaning of *conduct* in *conduct an investigation* is different from its meaning in *conduct an orchestra* or in *conduct electric current*, and all three differ from its meaning as an isolated lexical item (as in *John conducted himself abominably*). Therefore, it is only when we know the meaning of the collocate in combination with the base that we can understand the meaning of the collocation as a whole and use it appropriately. This is also why in collocation dictionaries the collocates of a lemma are usually grouped according to their meaning and why automatic techniques for semantic classification of collocation collocates should be involved when, e.g., compiling collocation dictionaries from corpora.

In what follows, we tackle the problem of the semantic interpretation (or semantic disambiguation) of collocates. As in Wanner (2004), Wanner et al. (2005; 2006a; 2006b) and Gelbukh and Kolesnikova (2012), we use as reference classification the fine-grained semantic typology of collocations that underlies *lexical functions* (LFs) (e.g. Mel'cuk, 1995). Our goal is also the same: to be able to assign to the collocate of any given collocation in context a semantic class tag from the LF typology. However, unlike these previous works, which use external lexico-semantic resources (namely EuroWordNet; see Vossen, 1998), we aim to explore techniques that do not use any external resources and that are thus more scalable and universal. The two main questions underlying our research are: (i) how well can we semantically classify collocates without the use of explicit semantic features; and (ii) to what extent can we dispense with explicit lexical information when classifying collocates.

So far, we worked on Spanish collocations from the lexicographic field of emotion nouns. The corresponding corpus annotated with LFs has been provided to us by the DICE team of the Universidad de La Coruña (http://www.dicesp.com), Spain. We have chosen Spanish since, to the best of our knowledge, only for Spanish an LF-annotated corpus is available. However, as will become clear from the presentation below, our approach is to a large extent language-independent.

In the next section, we briefly introduce the LF typology. Section 3 outlines the experiments we carried out to assess to what extent the classification of LF instances in the corpus is feasible by exclusively using features encountered in the textual context of these instances. Section 4 comprises a discussion of the outcome of these experiments. Section 5, finally, summarizes the insights we obtain and outlines the directions of our future work on this topic.

2. On the Semantic Collocate Typology

Earlier approaches to collocation extraction from corpora tended to consider any pair

of tokens that shows a significant co-occurrence tendency (a *strong association norm* in terms of Church and Hanks, 1989) to be a collocation, with the consequence that the result lists contained such pairs as *doctor – nurse*, *professor – university*, or *smoker – cigarette*; see, e.g., (Choueka, 1988; Church and Hanks, 1989). While being useful, for instance, for the construction of relational lexica, these pairs do not find their way into collocation dictionaries since they are not, strictly speaking, collocations. Nor can they be used in such tasks as lexicalization in Natural Language Text Generation, where lexical co-occurrence resources have shown to be of great value (e.g. Wanner, 1997).

Most of the more recent collocation extraction strategies have corrected this generous interpretation of co-occurrence and handle only word occurrences that form valid syntactic structures (Smadja, 1993; Evert and Kermes, 2003; Kilgarriff, 2006).[3] But this is not the end of the story: between the base and the collocate of a collocation not only a syntactic but also a semantic relation holds. This relation is often of abstract nature, such that it applies to 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. A possible label for these features is 'perform'. Obviously, the same label can be used to tag the meaning of *deliver*, *commit*, and *take* in these co-occurrences. The typology of lexical functions (LFs) as proposed in the framework of the Explanatory Combinatorial Lexicology (ECL) (Mel'cuk, 1995) captures this kind of semantic relations between the elements of collocations. The typology consists of about 30 classes of the type 'perform', 'react', 'begin to perform', 'continue to perform', 'take place', 'originate from', 'become involved', 'intense', 'positive', etc.

Table 1 displays, for illustration, examples for ten of these classes. In the first column, we add in parentheses the names of the LFs (Latin abbreviations) as used in the ECL literature and as we will use for the sake of brevity in the paper.

The LF typology is not the only semantic classification of collocates used in lexicography. As already mentioned above, all major collocation dictionaries tend to group collocates of a given lemma in accordance with semantic criteria. Consider, e.g., a fragment of the entry for INITIATIVE in the *Oxford Collocations* dictionary:

> undertake | plan | develop | announce | introduce, launch, set up, start | become involved | lead | approve | reject | sponsor | endorse, support ...

where '|' separates the semantic groupings of collocates.

[3] However, we obviously acknowledge that some researchers prefer to continue to work in the Firthian tradition of the term "collocation" and interpret any pair of tokens which co-occur with statistical significance as collocation. We think that both interpretations can cohabit as long as the authors clearly state the notion that they adopt.

Parallels of this grouping to (an abstracted) LF typology cannot be overlooked. Therefore, we have decided to use the following as reference typologies: (a) the genuine LF typology, because of its clear formal definition and potential of systematic abstraction; and (b) a generalized LF typology which is in its nature very similar to the implicit typologies used in broad distribution collocation dictionaries.

'act'/'perform' (Oper1)	*take* – walk, *give* – talk, *hold* – reception
'undergo'/'meet' (Oper2)	*receive* – blow, *encounter* – obstacle, *run into* – resistance
'act accordingly' (Real1)	*succumb to* – illness, *win* – match, *keep* – promise
'originate from' (Func1)	blow – *come from*, proposal – *stem from*, analysis – *be due to*
'be fulfilled by' (Fact1)	illness – *carry off*, benefit – *proceeds*, generosity – *pay off*
'begin to act/ perform' (IncepOper1)	*open* – dispute, *fall in* – love, *enter* – war
'begin to originate from' (IncepFunc1)	hatred – *come over*, panic – *seize*, routine – *catch up with*
'become more intense' (IncepPredPlus)	love – *grow*, voice – *become louder*, debate – *heat up*
'reduce intensity' (CausPredMinus)	*ease* – shortage, *contain* – inflation, *alleviate* – pain
'intensify' (CausPredPlus)	*increase* – pressure, *augment* – presence, *steer up* – hatred

Table 1: Samples of semantic classes of the LF typology (the collocates are in italics)

3. Experiments

In order to assess to what extent it is possible to identify the semantic labels of collocates in context, we carried out a series of experiments in which we interpreted the task of the semantic label identification as a machine learning-based classification task. As already mentioned above, others (e.g. Wanner, 2004; Wanner et al., 2006a,b) address the same problem using semantic features of the collocation elements from EuroWordNet (Vossen, 1998) to assess the similarity of a candidate co-occurrence with the samples of each given LF class. However, we do not use any external resources. Rather, we intend to explore to what extent semantic knowledge-poor techniques similar to those used for the extraction of collocations can be used for this purpose. In the case of a positive outcome, we furthermore want to explore: (i) whether these techniques also serve for the classification of collocations with respect to a generalized LF typology (of the kind found in broad coverage collocation dictionaries such as the *Oxford Collocations Dictionary* or

McMillan Collocation Dictionary); and (ii) whether lexical features (i.e., concrete words) are crucial for the classifier accuracy, or in other words, how semantic field-specific the classifier needs to be. [4]

3.1 Setup of the experiments

For our experiments on the classification with respect to the genuine LF typology, we focused on the ten LFs listed in Table 1. Table 2 displays the number of samples of each LF in the DICE corpus.

Collocate class	#
Oper1	1470
Oper2	149
Real1	147
Func1	179
Fact1	160
IncepOper1	152
IncepFunc1	244
IncepPredPlus	201
Caus Pred Minus	409
Caus Pred Plus	301

Table 2: Number of samples of each collocate class in the DICE corpus

For the experiments on a generalized fragment of the LF typology, we used five generic collocation categories proposed by colleagues from La Coruña; the generalization, including the subcategories of the general semantic categories, is displayed in Table 3. For readers interested in the actual LFs that compose the categories, they are listed in the Appendix.

For the classification experiments with respect to both typologies, we used the Weka machine learning environment, together with the LibSVM implementation. A linear kernel was chosen to generate the Support Vector Machine (SVM) models since it proved to be adequate for text classification tasks, which usually need to cope with a high amount of features. The following features were used:

- *Lexical features*: all tokens in the sentence + base + collocate + base-collocate pair.[5]
- *POS-features*: 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 +

[4] Recall that the DICE corpus contains only collocations from the field of emotions.

[5] In one of the experiments (see below), we suppressed the base and the base-collocate pair from feature set.

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 pairs of the syntactic dependents of the base and the POS of the base + POS pairs of the POS of the syntactic head of the collocate and the POS of the collocate + POS pairs 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.

Semantic category	**Subcategory**	**# of instances**
Intensity	'high intensity'	50
	'intensity increase'	491
	'intensity decrease'	468
Phase	'preparation'	14
	'initiation'	406
	'continuation'	309
	'termination'	523
Manifest	'manifestation'	1062
	'lack of manifestation'	407
Cause	'causation'	1001
Experimenter	'experimentation'	1478

Table 3: Fragment of the generalized LF typology

The POS and the morphological and syntactic dependency features were obtained by parsing the corpus with Bohnet's (2009) syntactic dependency parser.[6] We trained 10 binary classifiers on separate positive and negative corpora for each of the ten LFs. In the positive corpus, each sentence contained at least one collocation whose collocate was an instance of the given LF. The negative corpus consisted of the sentences with occurrences of the other LFs. Due to the high amount of negative class instances compared to the positive instances, we balanced each set by under-sampling the majority class.

[6] This parser performed best on Spanish in the CoNNL 2009 shared task.

O1	tener 'have' – admiración 'admiration', tributar 'tribute' – respeto 'respect', experimentar 'experience' – disgusto 'annoyance', tener 'have' – pudor 'modesty', sentir 'feel' – bochorno 'embarrassment', pasar 'pass' – apuro 'rush', abrigar 'nourish' – ilusión 'illusion'
O2	gozar 'enjoy' – admiración 'admiration', recibir 'receive' – consideración 'consideration', gozar 'enjoy' – respeto 'respect', sufrir 'suffer' – desprecio 'contempt', tener 'have' – sorpresa 'surprise'
R1	disfrutar 'enjoy' – felicidad 'happiness', degustar 'taste' – felicidad 'happiness', morir 'die' – [de 'of'] pena 'pity', aplicar 'apply' – pena 'sentence', sucumbir 'succumb'– [al 'to'] miedo 'fear'
Fu1	desprecio 'contempt' – anidar 'nest', alborozo 'joy' – reinar 'reign', satisfacción 'satisfaction' – reinar 'reign', felicidad 'happiness' – sonreír 'smile', desazón 'discomfort'– asaltar 'assault'
Fa1	tristeza 'sadness' – sacudir 'shake', pena 'pity' – comer 'eat', desazón 'discomfort' – quemar 'burn', temor 'fear' – paralizar 'paralyze', aprensión 'apprehension' – atenazar 'grip', aflicción 'grief' – azotar 'hit'
IO1	aversión 'aversion' – tomar 'take' , caer 'fall' – [en 'in'] abatimiento 'disheartenment', coger 'catch' – miedo 'fear', cobrar 'gain' – miedo 'fear', tomar 'take' – aprensión 'apprehension'
IF1	sentimiento 'feeling' – invadir 'invade', tristeza 'sadness' – entrar 'enter', desazón 'discomfort'– asaltar 'assault', miedo 'fear' – aparecer 'appear', pasmo 'amazement'– dar 'give', odio 'hatred' – surgir 'surface'
IPP	admiración 'admiration' – aumentar 'augment', respeto 'respect' – crecer 'grow', esperanza 'hope'– aumentar 'augment', angustia 'distress' – crecer 'grow', amistad 'friendship'– intensificar 'intensify'
CPM	enfriar 'freeze' – entusiasmo 'enthusiasm', aliviar 'alleviate' – desprecio 'contempt', paliar 'palliate' – sentimiento 'feeling', mermar 'diminish' extrañeza 'estrangement', frenar 'brake' – euforia 'euphoria'
CPP	aumentar 'augment' – respeto 'respect', reafirmar 'reaffirm' – entusiasmo 'enthusiasm', intensificar 'intensify' – desprecio 'contempt', avivar 'enliven' – aversión 'aversion', promover 'promote' – bienestar 'well-being'

Table 4: Correctly classified individual LF instance samples ('O1' = Oper1, 'O2' = Oper2, 'R1' = Real1, 'Fu1' = Func1, 'Fa1' = Fact1, 'IO1' = IncepOper1, 'IF1' = IncepFunc1, 'IPP' = IncepPredPlus, 'CPM' = CausPredMinus, 'CPP' = CausPredPlus)

For the experiments that targeted the exploration of the semantic field specificity of the classification, we had removed the lexical features from the feature lists.

3.2 Results of the experiments

Due to the context-driven nature of our classification procedure, classification examples should, in fact, always be shown together with their context rather than in isolation. However, in order to keep our presentation as clear and as simple as possible, we nonetheless cite in Tables 4 and 5 a few examples of the output of our LF classification in isolation. Table 4 illustrates some correctly classified samples of individual LFs. Table 5 below displays some of the correctly classified samples of the generalized LF typology.

I	sentir 'feel' – admiración 'admiration', rebajar 'reduce' – exasperación 'exasperation', aumentar 'augment' – bienestar 'well-being', aplacar 'appease' – ira 'anger', mitigar 'mitigate'– nostalgia 'nostalgy'
P	sospecha 'suspicion'– persistir 'persist', conservar 'conserve' – desapego 'indifference', desesperación 'desperation' – invadir 'invade', cariño 'affection' – desaparecer 'disappear', vergüenza 'shame' – entrar 'enter'
M	testimoniar 'testify'– afecto 'affect', satisfacer 'satisfy' – orgullo 'pride', ocultar 'hide' – pudor 'chestity', expresar 'express'– admiración 'admiration', contener 'control' – desencanto 'disappointment'
C	ahogar 'drown'– pena 'pity, despertar 'wake up' – encono 'lingering anger', conseguir 'achieve' – excitación 'excitation', suscitar 'stimulate' – resentimiento 'resentment', causar 'cause' – aprensión 'aprehension'
E	constituir 'form' – felicidad 'happiness', sentir 'feel' – alegría 'joy', tener 'have' – despreocupación 'disregard', abrigar 'harbor' – ilusión 'illusion', poseer 'possess' – temor 'fear'

Table 5: Correctly classified generalized LF instance samples ('I' = Intensity, 'P' = Phase, 'M' = Manifest, 'C' = Cause, 'E' = Experimenter)

If a sample occurs in the corpus several times (which is usually the case), each occurrence is analyzed separately, such that the same sample may be classified differently in different contexts. Sometimes, this is incorrect. Consider, e.g.:

1) ... *por ser oral fundamentalmente, ser transmitida de generación en generación que aumenta el apego del pueblo a su propia lengua*... 'for being basically oral, being transmitted from generation to generation, which strengthens the attachment of the people to their own language'

2) ... *a medida que aumenta el apego al cuerpo, el sufrimiento también aumenta* 'as the attachment to the body increases, the suffering also increases'

In both (1) and (2), *aumentar – apego* 'increase – attachment' is an instance of IncepPredPlus. However, in (1) it has been erroneously classified as CausPredPlus. On the other hand, the distribution-based classification procedure is sensitive to

fine-grained features that are decisive for the distinction between semantically very similar LFs. Thus, in (3), *aumentar – admiración* 'increase – admiration' is an instance of IncepPredPlus, while in (4), the same co-occurrence is an instance of CausPredPlus, such that multiple classification seems necessary.

3) *Su admiración aumenta al recordar la naturalidad con que se dirige a su marino* 'His admiration increases when he remembers the naturalness with which he talks to his seaman'.

4) *... tiene uno buen caldo de cultivo para aumentar su admiración por la hasta entonces controvertida figura del cretense* '... has a fertile breeding ground to augment his admiration for the until then controversial figure of the Cretan'.

The classification procedure correctly classifies the two co-occurrences.

3.3 Evaluation

To test the accuracy of our classifier models, we used a 10-fold cross-validation scheme. Tables 6 and 7 display the results of the classification obtained with respect to the genuine LF typology and the generalized LF typology, respectively.

The second and third columns in Table 6 show the results obtained with classifiers trained on the complete set of features; the fourth and fifth columns show the results obtained with classifiers trained on a set of features that did not contain the lexical tokens of the base. In the second and fourth columns, the accuracy of the classification of a given collocation as the LF in question is indicated; in the third and fifth, the accuracy of the recognition that a given collocation is not an instance of the LF in question is provided.

LF class	F-score (all features)		F-score (no lex. base feature)	
	+	–	+	–
CausPredMinus	0.90	0.99	0.68	0.89
CausPredPlus	0.84	0.98	0.57	0.79
Fact1	0.76	0.99	0.63	0.83
Func1	0.72	0.98	0.61	0.81
IncepFunc1	0.88	0.99	0.55	0.75
IncepOper1	0.85	0.99	0.65	0.86
IncepPredPlus	0.85	0.99	0.68	0.87
Oper1	0.91	0.95	0.64	0.80
Oper2	0.58	0.98	0.52	0.80
Real1	0.69	0.99	0.48	0.76

Table 6: Classification results per LF

For the two LFs with larger numbers of samples, Oper1 and CausPredMinus, we also performed an evaluation with a data split. For this purpose, we split the corresponding corpora into training and testing sets, with an 80% to 20% ratio (using the full set of features). For Oper1 classification, we then obtained a weighted average F-score of 0.93 and for the CausPredPlus an average F-score of 0.97. This is comparable with the performance obtained with the 10-fold cross-validation. For smaller samples, a data split proved to have negative consequences since the training sets of 80% were too small.

Table 7 displays the precision and recall figures of the classification with respect to the generalized LF typology with and without lexical features.

LF class	all features		no lex. base feature	
	p	r	p	r
Intensity	0.947	0.917	0.338	0.388
Phase	0.887	0.909	0.387	0.30
Manifest	0.925	0.904	0.367	0.446
Cause	0.82	0.828	0.442	0.346
Experimenter	0.906	0.92	0.538	0.567

Table 7: Classification results per generalized LF category ('p' = precision; 'r' = recall).

4. Discussion of the Evaluation

4.1 Classification using the LF typology

Table 6 shows that when using the full set of features, i.e., including the lexeme of the base, the classification with respect to the full-fledged LF typology achieves rather high accuracy scores (ranging from 0.58 for the recognition of Oper2-instances to 0.91 for the recognition of Oper1-instances); the variation of the accuracy is first of all due to the varying size of the training sets. The classification of negative instances is even better (between 0.95 and 0.99). This high accuracy is likely to be motivated by the distribution of the collocates of the collocations in a given semantic field (recall that we are dealing with a corpus on emotions here): in accordance with the Zipf law, a small number of collocate lexemes is very frequent, while the large rest occurs with a very limited frequency. Consider, for illustration, Table 8, where the share of the three most frequent collocates for four LFs in the DICE-corpus is given. It remains to be verified whether similar distributions can be observed in other semantic fields; our working hypothesis is that this is the case.

In the light of this distribution, an interesting research question is to what extent semantic field features influence the accuracy of the classification. Since the base lexemes are the most prominent features of a field (in our case, emotion nouns), the outcome of the second experiment in which we removed them from the feature lists is

of relevance; cf. columns 4 and 5 in Table 6. The accuracy is lower for all LFs, but not to an extent that would suggest that for each semantic field, separate collocate classifiers must be used. Since in both experiments positive instance classification turned out to be less accurate than negative instance classification, we focused in our error analysis on false positives.

Oper1	Freq.	**Real1**	Freq.
tener 'have'	26.80%	descargar 'unload'	9.52%
sentir 'feel'	20.74%	dar 'give'	8.84%
ser 'be'	8.57%	disfrutar 'enjoy'	7.48%
Total	**56.11%**		**25.85%**
CausPredMinus		**CausPredPlus**	
aplacar 'soothe'	12.46%	aumentar 'augment'	34.21%
mitigar 'moderate'	10.02%	acrecentar 'increase'	8.97%
aliviar 'alleviate'	9.53%	avivar 'brighten up'	7.64%
Total	**32.01%**		**50.85%**

Table 8: Collocate lexeme distribution in the DICE corpus

Table 9 shows the performance statistics for the classification with respect to four of the LFs using the complete set of features.

LF	**\# Corr.**	**\# Inc.**	**\# FP**
Oper1	4039	262	153
Real1	4220	81	23
CausPredPlus	4205	96	76
CausPredMinus	4228	73	39

Table 9: Error statistics in the individual LF classification

The second column contains the number of correctly classified instances (Corr.), the third the number of incorrectly classified instances (Inc.), and the fourth indicates how many of the incorrectly classified instances are false positives (FP).

A more detailed analysis reveals the following major confusion figures shown in Table 10.

Oper1:	Func 1 (36), Incep Pred Plus (31)
Real1:	Oper1 (6), Real2 (5)
CausPredPlus:	IncepPredPlus (35), CausPredMinus (6)
CausPredMinus:	IncepPredMinus (18), CausPredPlus (17)

Table 10: Classification confusion figures

As expected, the classifiers more commonly confuse LF-instances with very similar syntax. Consider, for instance, Real1 vs. Oper1 vs. Real2: here, we need to capture the semantic difference between, e.g., *keep a promise* vs. *give a promise* vs. *hold* / *fulfill to a promise* – which is hard, although not impossible, using the distributional semantic features we exploited so far. The confusion in the case of CausPredPlus and CausPredMinus is analogous, but still more subtle and thus more difficult to capture: the difference between CausPredPlus respectively CausPredMinus and the LFs with which they are confused consists of a few deep semantic features ('begin to increase' vs. 'increase', 'decrease' vs. 'increase', etc.). Thus, for example, many of the instances of CausPredPlus that have been classified as IncepPredPlus contain the collocate *aumentar* 'augment'; see above, and these examples:

> *aumentar* – placer 'pleasure', *aumentar* – confusión 'confusion', *aumentar* – sensación 'sensation', *aumentar* – admiración 'admiration', *aumentar* – abatimiento 'disheartenment'

4.2 Classification using the generalized typology

A comparison of the figures in Tables 6 and 7 reveals that the balanced F-score achieved during the classification with respect to the generalized LF-typology is persistently higher than the average F-score across the individual LFs that constitute the generalized categories. For instance, the average F-score for recognition of the instances of the three LFs CausPredPlus, IncepPredPlus, and CausPredMinus using lexical features is 0.863, while the recognition of instances of 'Intensity' (which includes, among others, the above three LFs) achieves an F-score of 0.932. This can be interpreted as a sign of quality of the generalized LF-typology: similar LFs that were still confused in the individual LF classification exercise have been gathered into (more) homogeneous semantic categories, with clearer (first of all lexical) discrimination boundaries. However, with the generalized typology confusions obviously also occur. The corresponding confusion matrix in Table 11 reveals that 'Intensity' is confused more with 'Phase' than with other categories, 'Phase' and 'Manifest' with 'Cause', 'Cause' with 'Experimenter' and vice versa. The confusions can be explained by a more detailed analysis of the composition of the generalized categories, or, in other words, by the proximity of the individual LFs that compose the categories. Since this would imply a detailed introduction to the LFs, we refrain from such an analysis here. For the convenience of readers who are familiar with LFs, we provide the list of LFs of which each category is composed in the Appendix.

	I	P	M	C	E
Intensity (I)	944	41	19	16	9
Phase (P)	17	1212	30	48	27
Manifest(M)	19	45	1415	59	28
Cause (C)	11	39	42	985	94
Experimenter (E)	6	29	24	73	1521

Table 11: Confusion matrix in the generalized classification with lexical features

In contrast to the generalized classification which uses lexical features, the classification in which no lexical features have been used cannot compete with individual LF classification; cf. the *p* and *r* figures in columns 4 and 5 of Table 7: both precision and recall are considerably lower. The lack of lexical features penalizes the classification with respect to the generalized LF typology more than it does with respect to the individual LF typology. The confusion matrix in Table 12 shows that the confusion patterns also change. Thus, while 'Intensity' is still mostly confused with 'Manifest', 'Phase' is now confused most often also with 'Manifest' and not with 'Cause', 'Manifest' with 'Experimenter', etc. This is because the syntactic and contextual features of the LFs between these categories are more similar than are the lexical features. A more detailed study is needed to improve on the overall accuracy of generalized classification without the use of lexical features.

	I	P	M	C	E
Intensity (I)	395	150	337	58	98
Phase (P)	254	400	343	126	211
Manifest(M)	250	217	693	127	279
Cause (C)	131	119	238	374	219
Experimenter (E)	138	147	277	161	930

Table 12: Confusion matrix in the generalized classification without lexical features

5. Conclusions and Future Work

We have presented an excerpt of ongoing work on the semantic classification of collocates, which has until now been a largely neglected aspect of collocation processing but which we believe to be very important. To the best of our knowledge, the only existing works on the problem are those presented in Gelbukh (2012), Wanner et al. (2006a,b) and Wanner (2004). In contrast to these previous works, we do not use any external semantic resources and thus avoid two major disadvantages: (i) that the results could be negatively affected by the incompleteness and bias of the Spanish EuroWordNet towards English; and (ii) that an external semantic resource

for a specific language could limit the scalability and porting of the developed tool to other languages. Thus, our approach is much more flexible. The results obtained so far using the corpus of emotions and the genuine LF typology as reference typology are very encouraging, particularly if we take into account that the LF typology is very fine-grained. The preliminary experiments on the generalized LF typology need to be further extended since they have the potential to provide rich (and already appropriately grouped) input material for general public collocation dictionaries. In the immediate future, we plan to extend our experiments to generic corpora and to combine collocate classification with collocation identification, such that automatic semantic labeling of collocates in corpora becomes a realistic task.

6. Acknowledgements

The research reported in this paper has been partially funded by the Spanish Ministry of Economy and Competition (contr. number FFI2011-30219-C02-02) in the framework of the HARENES Project, carried out in collaboration with the DICE Group of the University of La Coruña; many thanks to Margarita Alonso Ramos and Orsolya Vincze for their support. We are also grateful to two anonymous reviewers for their helpful comments. At the time of the reported research, the first and second authors were members of the NLP group, Department of Information and Communication Technologies, UPF.

7. References

Bohnet, B. (2009). Efficient Parsing of Syntactic and Semantic Dependency Structures. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009)*.

Bouma, G. (2010). Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010, Short paper track*. Uppsala.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO*, pp. 34–38.

Church, K.W. & P. Hanks. (1989). Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the ACL*, pp. 76–83.

Evert, S. & H. Kermes. (2003). Experiments on candidate data for collocation extraction. *Companion Volume to the Proceedings of the 10th Conference of the EACL*. 83–86.

Evert, S. (2007). Corpora and collocations. In *Corpus Linguistics. An International Handbook* ed. by A. Lüdeling & M. Kytö. Berlin: Mouton de Gruyter.

Gelbukh, A. & O. Kolesnikova. (2012). *Semantic Analysis of Verbal Collocations with Lexical Functions*. Springer.

Hausmann, F.-J. (1989). Le dictionnaire de collocations. In Hausmann F.J., Reichmann O., Wiegand H.E., Zgusta L. (eds). In *Wörterbücher : ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires*. Berlin/ New York: De Gruyter. 1010-1019.

Kilgarriff, A. (2006). Collocationality (and how to measure it). In *Proceedings of the 12th EURALEX International Congress*. Torino.

Mel'čuk, I.A., (1996). Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In L. Wanner (ed.): *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/ Philadelphia: Benjamins, 37-102.

Mel'čuk, I.A. (1995). Phrasemes in Language and Phraseology in Linguistics. In *Idioms: Structural and Psychological Perspectives* ed. by M. Everaert, E.-J. van der Linden, A. Schenk & R. Schreuder. 167–232. Hillsdale: Lawrence Erlbaum Associates.

Pecina, P. (2008). A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 54–57. Marrakech.

Smadja, F. (1993). Retrieving Collocations from Text: X-Tract. *Computational Linguistics*.19.1:143–177.

Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Wanner, L. (2004). Towards Automatic Fine- Grained Semantic Classification of Verb-Noun Collocations. *Natural Language Engineering Journal*. 10.2:95–143.

Wanner, L. (1997). *Exploring Lexical Resources for Text Generation in a Systemic Functional Language Model*. PhD Dissertation. Universität des Saarlandes.

Wanner, L., B. Bohnet & M. Giereth. L. (2006a) 'What Is Beyond Collocations? Insights from Machine Learning Experiments'. In *Proceedings of the EURALEX Conference*, Turin.

Wanner, L., B. Bohnet & M. Giereth. (2006b). Making Sense of Collocations. *Computer Speech and Language*. 20.4:609–624.

Wible, D. & N.L. Tsao. (2010). Stringnet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL-HLT Workshop on Extracting and Using Constructions in Computational Linguistics*. Los Angeles.

8. Appendix

The following table shows the composition of the generic LF categories by individual LFs. For definitions of the LFs, see, e.g. Mel'čuk (1996).

Semantic category	**LF (# of instances)**
Intensity	Magn+Oper1 (48), Magn+ Caus1Manif (2), CausPredPlus (292), IncepPredPlus (199) CausPredMinus (412), IncepPredMinus (63)
Phase	PreparReal1 (7), IncepOper1 (129), IncepFunc1 (234), Magn + IncepFunc1 (43), ContOper1 (94), CausContFunc0 (82), CausContFunc1 (1), ContFunc0(80), ContFunc1 (52), FinOper1 (113), LiquOper1 (36), Liqu1Func0(256), FinFunc0 (109), FinFunc1 (9)
Manifest	CausManif (610), AntiVer+Caus1Manif (6), Magn+Caus1Manif (2), Caus1Manif (2), Conv21Manif (86), IncepManif (22), PredA1Manif (6), Perm1Manif (3), Real1 (141), Caus1Func2 (5), Mang+Caus1Func2 (1), Fact1 (148), Magn+Fact1 (32), nonPermFact0 (96), nonPerm1Manif (261), nonFact1 (2), AntiReal1 (48)
Cause	V (155), CausFunc0 (186), MagnCausFunc0 (1), Caus2Func1(200), Caus2Func2(116), CausOper1 (102), Magn+CausOper1 (39), Func3 (18), Oper2 (143), Plus+Oper2 (1), Real2 (49)
Experimenter	Oper1 (1311), nonOper1 (3), Func1 (164)