ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER

Computer Speech and Language xxx (2005) xxx–xxx

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Making sense of collocations ☆

Leo Wanner [a,*], Bernd Bohnet [b], Mark Giereth [b]

[a] *ICREA and Pompeu Fabra University, Passeig de Circumvallació, 8, Barcelona 08003, Spain*
[b] *Intelligent Systems Institute, University of Stuttgart, Germany*

Received 22 November 2004; received in revised form 28 July 2005; accepted 5 October 2005

Abstract

Lexico-semantic collocations (LSCs) are a prominent type of multiword expressions. Over the last decade, the automatic compilation of LSCs from text corpora has been addressed in a significant number of works. However, very often, the output of an LSC-extraction program is a plain list of LSCs. Being useful as raw material for dictionary construction, plain lists of LSCs are of a rather limited use in NLP-applications. For NLP, LSCs must be assigned syntactic and, especially, semantic information. Our goal is to develop an "off-the-shelf" LSC-acquisition program that annotates each LSC identified in the corpus with its syntax and semantics. In this article, we address the annotation task as a classification task,viewing it as a machine learning problem. The LSC-typology we use are the *lexical functions* from the Explanatory Combinatorial Lexicology; as lexico-semantic resource, EuroWordnet has been used. The applied machine learning technique is a variant of the *nearest neighbor*-family, which is defined over lexico-semantic features of the elements of LSCs. The technique has been tested on Spanish verb–noun bigrams.

1. Introduction

Lexico-semantic collocations (LSCs) in the sense of Moon (1998), or *simple decomposable MWEs* in the terminology of Baldwin et al. (2003), are a prominent type of *Multiword Expressions* (MWEs). As a rule, an LSC is a combination of two lexical items in which the semantics of one of the items (the *base*) is autonomous from the combination it appears in, while the semantics of the other item (the *collocate*) depends on the semantics of the base. Thus, in take [a] leave,[1] the base is *leave* and the collocate is

☆ This work has been partially supported by the German Academic Exchange Agency in the framework of the Programme *Acciones Integradas Hispano-Alemanas.*

* Corresponding author. Tel.: +34 935422307; fax: +34 935422321.
E-mail addresses: leo.wanner@upf.edu (L. Wanner), bernd.bohnet@iis.uni-stuttgart.de (B. Bohnet), mark.giereth@iis.uni-stuttgart.de (M. Giereth).

[1] Citing verb–noun LSCs, we add, where it appears useful for better readability, the article of the noun, although the article is, strictly speaking, not part of the LSC.

0885-2308/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2005.10.002

[*to*] *take*, in *give* [a] *statement*, the base is *statement* and the collocate is [*to*] *give*, in *rancid butter*, the base is *butter* and the collocate *rancid*, in *confirmed bachelor*, *bachelor* is the base and *confirmed* is the collocate, and so on. Benson (1989) points out that (lexico-semantic) collocations are ‘‘arbitrary recurrent word combinations’’; see, also (Cowie, 1994; Mel’čuk, 1995), among others, for detailed presentations of the idiosyncratic features of LSCs. As a consequence, LSCs tend to be language-specific. For instance, in English one *makes or takes a decision*, in French and Italian one ‘takes’ but does not ‘make’ it (*prendre/*faire une décision*, *prendere/*fare una decisione*), in German, one ‘meets it’ (*eine Entscheidung treffen*), in Spanish one ‘adopts’ or ‘takes’ it (*adoptar/tomar una decisión*), and in Russian one ‘hosts’ it (*prinjat′ rešenie*); in English one gives a *lecture* – as in French (*donner un cours*) and Spanish (*dar una clase*) – in German and Italian one ‘holds’ a lecture (*eine Vorlesung halten, tenere una lezione*), and in Russian one ‘reads’ it (*čitat′ lekciju*); etc.

Due to the idiosyncrasy of LSCs, and thus the need of their explicit listing in a language’s lexical resource, the extraction of LSCs is an increasingly important and prominent issue. The result of most LSC-extraction strategies proposed to date is a list of collocations identified in the corpus, possibly annotated with morpho-syntactic information. But while plain lists of LSCs are a useful resource for manual dictionary construction, their usefulness is rather limited, e.g., for Text Generation, Machine Translation and Text Summarization. In order to be useful in NLP as well as, for instance, in second language learning, an LSC must be supplied with its semantics. In other words, if the meaning of an LSC is not determined during its retrieval, it must be assigned (as a rule, manually) in a subsequent stage – as done, e.g., by Smadja and McKeown (1991).

One way to decide whether a given word combination is an LSC and to determine its semantics is to classify it according to a fine-grained semanticosyntactic typology of collocations. Such a typology is given by *lexical functions*, LFs; cf. Mel’čuk (1996). Wanner (2004) discusses the use of a variant of *instance*-based machine learning (ML) for the classification of verb–noun LSCs according to the typology of LFs. For each LF, a typical semantic feature ‘‘profile’’ (*a centroid*) is constructed. Given that not all eatures of a centroid equally contribute to the distinction of its LF, the features are usually weighted. Once the centroids are constructed in a learning stage, the features of test bigrams are compared with the centroids. A sufficient overlap qualifies a bigram as an instance of the LF in question. Being quite performative (with an average score of 70%), this technique requires an extensive tuning of the weighting variables for each set of test bigrams. This is a serious obstacle for its use in an ‘‘off-the-shelf’’ collocation-classifier.

The goal of this article is twofold:

- to present an easy-to-use standard technique that does not require costly domain-specific tuning, but still ensures good quality LSC-classification;
- to provide further evidence that the automatic compilation of detailed semantically annotated collocation lexica is feasible.

We performed a series of experiments with different ML-techniques. The technique we discuss in detail is a variant of a standard ML-technique known as the *nearest-neighbor* (NN) classification technique. All experiments presented in this article have been carried out with Spanish material. As lexico-semantic descriptions of the lexical elements of the training and test bigrams, their hyperonym hierarchies in the Spanish part of the EuroWordnet (Vossen, 1998), henceforth ‘‘SpanWN’’, have been used.

The remainder of the article is structured as follows. In the following section, we introduce the LF-typology. In Section 3, the theoretical basics of the NN-classification model are discussed. Section 4 presents a description of lexical meanings in terms of hyperonym hierarchies in SpanWN. Section 5 outlines the setup of the experiments, which are then described in Section 6. Section 7 evaluates the quality figures achieved within the experiments; to illustrate the advantage of the NN-model, we briefly contrast its quality figures with the quality figures achieved with a number of other ML-techniques. Section 8 contains an overview of the related work, and Section 9 draws the conclusions from our studies and presents some of the remaining issues for future research.

2. Semantic typology of LSCs: lexical functions

In this article, we presuppose the following three basic features of LSCs:[2]

- an LSC is a binary combination of lexical items;
- an LSC possesses a stable syntactic structure, i.e., in the basic (active) form of a given verb–noun LSC, between the elements of this LSC a syntactic dependency relation holds, and the syntactic dependent always possesses the same grammatical function with respect to the governor;
- an LSC is a lexically restricted word combination and cannot thus be constructed using universal (semantic) selectional restrictions.

The three features are underlying the definition of the typology of lexical functions (LFs). In what follows, we restrict the introduction to LFs to the absolute minimum necessary for the understanding of the content of the article. For a comprehensive overview, see Mel'čuk (1996); for a more detailed presentation of LFs as a classification typology cf. Wanner (2004).

In our context, only the *syntagmatic* LFs are of relevance. A syntagmatic LF is a (directed) standard abstract lexico-semantic relation that holds between the base and the collocate of a given collocation. 'Standard' means that this relation applies to a large number of LSCs. For instance, the relation that holds between *step* and *take* in *Mary takes a step* is the same as the one that holds between *speech* and *deliver, suicide* and *commit, accident* and *have*, and so on. It is the same in the sense that it implies that each collocate provides the same semantic and syntactic linguistic features to its base; cf. Kahane and Polguère (2001). 'Abstract' means that the meaning of this relation is sufficiently general and can therefore be exploited for purposes of generalization and thus classification. In Mel'čuk (1996), about 36 different "simple standard" syntagmatic LFs are distinguished. About 20 of them capture verb noun collocations. Simple LFs can further combine to form "complex LFs"; for a mathematically sound composition calculus, see Kahane and Polguère (2001). In our experiments, we use a subset of both simple and complex LFs.

As names of LFs, abbreviations are used. For instance, 'Oper_1' stands for 'perform', 'do'; 'Oper_2' for 'undergo', 'meet'; Func_0 for 'happen', 'take place'; etc.[3] Consider, for illustration, eight of the most common standard verb–noun LFs in Table 1. The meaning of each LF appears in quotes and its name in parentheses. The arguments of the LFs, i.e., the bases, are written in small capitals, their values, i.e., the collocates, in a *slanted font*. The table illustrates that verb–noun LSCs go well beyond support verb constructions (SVCs),[4] the extraction of which has received considerable attention by researchers working in computational corpus linguistics; cf., e.g., Grefenstette and Teufel (1995); Dras (1995); Tapanainen et al. (1998); Stevenson et al. (2004). Only the LFs 1–5 can be considered SVCs; in the LFs 6–8, the verb expresses full (although possibly idiosyncratic) semantic content.

3. Using NN-classification for classifying LSCs

The task we address in this article can be formulated as follows: Given a plain list of verb–noun LSCs, classify each LSC with respect to the LF-typology. To be able to classify a bigram with respect to the LF-typology T,

[2] These features make clear that the notion of LSC is different from the notion of *collocation* in the sense of Firth (1957) and Halliday (1966), who define a collocation as a high probability association of lexical items in the corpus. See Wanner (2004) for a contrastive discussion of the two notions. A number of works on the extraction of collocations from corpora draws upon Firth's interpretation of the term; cf., e.g., Choueka et al. (1983); Church and Hanks (1989); Justeson and Katz (1995).

[3] The subscripts to the LF-names specify the projection of the semantic structure of the LSCs denoted by an LF onto their syntactic structure. Since we interpret complete LF-names as collocation class labels, we can ignore the semantics of the subscripts and consider them simply as part of LF-names.

[4] *Support* (or *light*) *verb constructions* (Allerton, 1984; Abeillé, 1988; Alonso, 2004b) are verb–noun constructions in which the verb carries little semantic content and is used for the sake of its structural properties only – as in *take* [a] *walk, harbor* [a] *thought, give* [a] *presentation*.

Table 1
Eight standard verb–noun LFs

1.	‘perform’, ‘do’, ‘act’ (Oper$_1$)		5.	‘concern’, ‘apply to’ (Func$_2$)	
	INSULT	*throw*		DISPUTE	*concern*
	PROBLEM	*pose*		DISCUSSION	*center* [*on*]
	OPPOSITION	*mount*		CHANGE	*affect*
	RESPECT	*have*		INFORMATION	*relate* [*to*]
2.	‘undergo’, ‘meet’ (Oper$_2$)		6.	‘act accordingly’ (Real$_1$)	
	INSULT	*suffer*, *endure*		ACCUSATION	*prove*
	PROBLEM	*face*, *encounter*		PROMISE	*keep*
	OPPOSITION	*encounter*, *run* [*into*]		SCHEDULE	*stick* [*to*]
	RESPECT	*command*		THREAT	*fulfil*
3.	‘happen’, ‘take place’ (Func$_0$)		7.	‘react accordingly’ (Real$_2$)	
	INSULT	*fly*, *occur*		DEMAND	*fulfil*, *meet*
	SNOW	*fall*		HINT	*take*
	SUSPICION	*linger*		LAW	*abide* [*to*]
	NEWS	*travel*		CALL	*answer*
4.	‘originate from/with’ (Func$_1$)		8.	‘put an end to’ (Liqu$_1$Func$_0$)	
	ANALYSIS	*be due* [*to*]		SUPPORT	*withdraw*
	IDEA	*originate* [*from*]		RESISTANCE	*put down*
	PROPOSAL	*stem* [*from*]		OBSTACLE	*remove*
	OPPOSITION	*come* [*from*]		EXITEMENT	*die down*

the characteristic features shared by all instances of an LF L in T must be “learned”. Then, the features of the candidate bigram can be compared with the features of the instances of L. If a sufficient similarity is observable, the bigram is likely to be an instance of L as well.

In corpus-based NLP, characteristic features of a word pattern are most often captured in terms of word frequency counts. In contrast, we use *semantic component* (or *concept*) counts, i.e., we assume that the meanings of the elements of the bigrams considered are componential.[5] This has two major advantages. Firstly, we are not bound to the frequency with which a candidate bigram occurs in the corpus. The frequency criterion proved to be a serious obstacle for the identification of less common LSCs. Some authors explicitly reject recurrency as a criterion for a word combination to be considered a collocation; cf. Cowie (1994); Mel’čuk (1995). Secondly, we naturally generalize over collocates with the same meaning. Thus, the concept count allows us to detect the close semantic similarity between [*to*] *brim* [*with*] and [*to*] *exude* in co-occurrence with *confidence* and between *close, intimate* and *deep* in co-occurrence with *friendship*. Such a generalization is a decisive step towards semantically oriented LSC-classification.

We start from a training set of manually compiled disambiguated instances for each of the *n* LFs used in the classification task. Unlike the other ML-techniques, nearest neighbor classification does not include, strictly speaking, a learning stage. In abstract terms, it can be described as a pair of vector space models (Salton, 1980). That is, it can be thought of as consisting of a training material representation stage and a classification stage.

3.1. Representation stage

Assume a training set of instances for each LF $\mathsf{L}_1, \mathsf{L}_2, \ldots, \mathsf{L}_n$ in T. Let $\mathcal{B}$ be the meaning component collection over the base sets of the instances from the training sets of all LFs in T and $\mathcal{C}$ the meaning component collection over the collocate sets of the instances from training sets of all LFs in T. $\mathcal{B}$ and $\mathcal{C}$ naturally map

[5] The componential description of the corresponding words is expected to be available from an external lexical resource. Any sufficiently comprehensive lexicosemantic resource suitable for NLP can be used. As already pointed out in Section 1, we use SpanWN, the Spanish part of the EuroWordnet.

onto multidimensional vector spaces $V_{\mathscr{B}}$ (the *base description space*) and $V_{\mathscr{C}}$ (the *collocate description space*). Each component $b \in \mathscr{B}$ and each component $c \in \mathscr{C}$ provide a distinct dimension in $V_{\mathscr{B}}$ and $V_{\mathscr{C}}$, respectively. Each training instance I is thus represented as a pair of vectors $(\vec{v}_{b_I}, \vec{v}_{c_I}) \in (V_{\mathscr{B}}, V_{\mathscr{C}})$. In the most simple realization of the model, $\vec{v}_{b_I}$ and $\vec{v}_{c_I}$ will contain a '1' for dimensions (components) available in I and a '0' for dimensions that are not available in I. Obviously, realizations with a weighting schema are possible to take into account the varying importance of dimensions for the description of an LSC-instance.We use the binary weighting schema.

Before applying this representation in the classification stage, those samples that are "unreliable" are removed from $(\mathscr{B}, \mathscr{C})$. We consider a sample unreliable if it is nearest to an instance of a different LF than it is itself. To determine which instance is nearest, we use Eq. (1) from the classification stage; see below.

3.2. Classification stage

Given a candidate word bigram $K: = (N, V)$ that is to be classified according to the LF-typology, the classification stage consists of (a) the decomposition of the meaning of N into the component set N and of the meaning of V into the component set V; (b) mapping of (N,V) onto $(V_{\mathscr{B}}, V_{\mathscr{C}})$. The LF-label of the instance I whose vector pair $(\vec{v}_{b_I}, \vec{v}_{c_I})$ is nearest to the vector pair $(\vec{v}_{n_K}, \vec{v}_{v_K})$ of K is assigned to the candidate.

To determine the similarity between $(\vec{v}_{b_I}, \vec{v}_{c_I})$ and $(\vec{v}_{n_K}, \vec{v}_{v_K})$, the *cosine* or any other suitable metric can be used. In our experiments, we used the following set-based metric:

$$sim(I,K) = \beta \frac{f_b}{f_{b_{\max}}|N|} + \gamma \frac{f_c}{f_{c_{\max}}|V|} \tag{1}$$

with f_b as $|\vec{v}_{b_I} \cap \vec{v}_{n_K}|$, i.e., the number of dimensions shared by $\vec{v}_{b_I}$ and $\vec{v}_{n_K}$;$f_{b_{\max}}$ as the maximal number of dimensions shared by $\vec{v}_{n_K}$ and a base vector of any instance in the training set for the LF of which I is an instance;f_c as $|\vec{v}_{c_I} \cap \vec{v}_{v_K}|$, i.e., the number of dimensions shared by $\vec{v}_{c_I}$ and $\vec{v}_{v_K}$, and $f_{c_{\max}}$ as the maximal number of dimensions shared by $\vec{v}_{c_K}$ and a collocate vector of any instance in the training set for the LF of which I is an instance. $|N|$ stands for the number of components in the description of the noun of K and $|V|$ for the number of components in the description of the verb of K. β and γ are constants that can be used to tune the importance of the base and collocate, respectively, for the classification. In our experiments (Section 6), we used $\beta: = 1$, $\gamma: = 1.5$; that is, we assigned higher importance to the collocate meaning than to the base meaning. If $f_{c_{\max}} = 0$ (which means that $\vec{v}_{c_I}$ and $\vec{v}_{v_K}$ do not share any dimension), the second summand in Equation (1) becomes invalid and the candidate bigram is rejected as an LSC of the type L of I. The candidate bigram can also be rejected if $sim(I,K)$ is smaller than a given threshold for all instances of L in the training set.[6]

To reduce the number of vector pair comparisons in the classification stage, the vector pairs of similar instances can be merged beforehand. Experiments show that an improvement of the processing time of about 20% can be achieved. However, such a merge always implies a decrease of the classification quality.

4. SpanWN as the source of the semantic description of lexical items

For the componential description of the LF-instances in the training sets as well as for the description of the candidate bigrams, we use the hyperonym hierarchies provided by SpanWN, the Spanish part of the lexical database EuroWordNet (Vossen, 1998). SpanWN is a middle-size lexical database organized in terms of sets of synonymous or quasi-synonymous word senses (the sets are called *synsets* and their elements *variants* of a synset). The average number of senses distinguished in SpanWN for nouns is about four; that of verbs about seven (among the most ambiguous verbs are *dar* 'give' with 17 senses, *hacer* 'do' with 19 senses, and *llevar* 'carry' with 25 senses). In contrast to the Princeton WN (Fellbaum (ed.), 1998), where the hyperonym hierar-

[6] There is still some room for improvement of the metric. Thus, we achieved better quality figures with the following metric: $sim(I,K) = \beta \frac{f_b}{f_{b_{\max}}} + \gamma \frac{f_c}{f_{c_{\max}}} + \frac{f_{b_{\max}}}{|N|} + \frac{f_{c_{\max}}}{|V|}$. However, since it appeared less motivated than Eq. (1), we used (1).

```
(7. RECLAMACIÓN3
    communication
    6. INSTANCIA2 PETICIÓN1 PEDIDO1 MANIFESTACIÓN2 RELACIÓN5
       communication
       5. MENSAJE2 CONTENIDO3
          communication Communication|Usage|Mental
          4. COMUNICACIÓN1
             Tops 3rdOrderEntity|Social|Purpose|Mental|Communication
             3. RELACIÓN_SOCIAL1
                Tops Relation|Social
                2. RELACIÓN1
                   Tops Relation
                   1. ABSTRACCIÓN1
                      Tops)
(6. PRESENTAR3
    communication
    5. SOMETER2
       communication
       4. PEDIR1
          communication Agentive|BoundedEvent|Communication|Purpose
          3. COMUNICAR2
             communication Agentive|Communication|UnboundedEvent
             2. INTERACTUAR1
                social Agentive|Dynamic|Social
                1. ACTUAR4 LLEVAR-A-CABO2 HACER15
                   social Agentive|Dynamic)
```

Fig. 1. Hyperonym hierarchies for PRESENTAR**3** and RECLAMACIÓN in the collocation *presentar [una/la] reclamación* (lexical items are written in small capitals, BCs in *sans serif*, and the TCs start with a capital; individual TCs are separated by the '|' sign).

chy of a lexical item is purely lexical (i.e., contains only hyperonyms), in SpanWN the hyperonym hierarchy of each lexical item consists of:

- its hyperonyms and synonyms (i.e., words that combine with the lexical item in question to form a *synset*),
- its own *Base Concepts* (BCs) and the BCs of its hyperonyms,
- the *Top Concepts* (TCs) of its BCs and the TCs of its hyperonyms.

BCs are general semantic labels that subsume a sufficiently large number of synsets. Examples of such labels are: change, feeling, motion, and possession. Thus, DECLARACIÓN**3** 'declaration' is specified as communication, MIEDO**1** 'fear' as feeling, PRESTAR**3** 'lend' as possession, and so on.[7] Unlike *unique beginners* in the original WN, BCs are mostly not "primitive semantic components" (Miller, 1998); rather, they can be considered labels of semantic fields.[8] The set of BCs used across different WNs in the EuroWN consists of 1310 different tokens. The language-specific synsets of these tokens constitute the cores of the individual WNs in EuroWN.

Each BC is described in terms of TCs – language-independent features such as Agentive, Dynamic, Existence, Mental, Location, Social, etc. (in total, 63 different TCs are distinguished). For instance, the BC change is described by the TCs Dynamic, Location, and Existence.

Consider, for illustration, Fig. 1, which shows the hyperonym hierarchies (including synonyms, BCs and TCs) of PRESENTAR**3** 'present' and RECLAMACIÓN**3** 'complaint' from the collocation *presentar* [una] *reclamación* lit. 'present [a] complaint'.

In *presentar* [una] *reclamación*, it is the third SpanWN-sense of *reclamación* and the third SpanWN-sense of *presentar* that apply. PRESENTAR**3** does not possesses any synonymous senses. The BC of the corresponding one-element synset is communication, which does not display any TC-features. The immediate hyperonym of PRESENTAR**3** is SOMETER**2** 'submit', which in turn possesses the hyperonym PEDIR**1** 'request'. Both are communication

[7] The numbers indicate the corresponding senses in SpanWN.

[8] Note, however, that unique beginners of Princeton WN are part of the BC-set.

lexemes. PEDIR**1** carries the TC-features Agentive, Bounded Event, Communication and Purpose. And so on. The root the hyperonym hierarchy of PRESENTAR**3** is given by the synset {ACTUAR**4**, LLEVAR-A-CABO**2**, HACER**15**}.

The BC of RECLAMACIÓN**3** is equally communication and its parent hyperonym synset consists of INSTANCIA**2**, PETICIÓN**1**, PEDIDO**1**, MANIFESTACIÓN**2**, and RELACIÓN**5** . The BC of this synset is again communication with no TC-features. The root of the hyperonym hierarchy of DECLARACIÓN**3** is ABSTRACCIÓN**1** 'abstraction' with the generic TC-feature Tops.

5. Setting up the experiments

To validate the proposed NN-classification technique and to compare its performance with Wanner (2004) and with other common ML-techniques used in computational lexicography, we conducted two experiments with different training and test material. We used the same LFs and the same data as in Wanner (2004). In the first experiment, we trained on and classified candidate verb–noun bigrams the nouns of which belong to the same semantic field, namely to the field of *emotion nouns*. In the second experiment, we classified verb–noun bigrams with no consideration of field constraints. A separate experiment on mono-field material is of value because the meanings of the nouns that belong to the same semantic field are a priori homogeneous at a certain level of abstraction; the lexical-semantic description of the instances of the same LF can thus be assumed to be very similar. This allowed us expect reasonably good quality figures for single-field classification. We have chosen emotion nouns because they are rich in collocations and because for emotion nouns, lists of LF-instances are already available for French (Mel'čuk et al., 1984, 1988, 1992, 1999), German (Mel'čuk and Wanner, 1996), and, what is more important, for Spanish (Alonso, 2004a). Obviously, the availability of these resources facilitated the compilation of the training material.

5.1. Choosing LFs for the experiments

The LFs used in the experiments must be chosen so that they illustrate, on the one hand, the range of different types of verb–noun LSCs that we are able to recognize, and, on the other hand, the potential of the techniques to distinguish between similar types of collocations. We consider two verb–noun collocations to be similar if their semantics are similar and/or their *government patterns* are the same, i.e., if they project the semantic actant structure of the noun onto the syntactic structure of the verb in the same way. Therefore, for both experiments, we selected several LFs with similar semantic features and the same government pattern and at least one LF that was sufficiently different from the others (either in terms of its syntactic structure or in terms of its semantics). To judge the semantic similarity between several LFs we examined their glosses provided in Mel'čuk (1996) and then relied on our intuition.

As will become clear below, LFs with the same government pattern may be semantically very similar. This raises the question whether these LFs should be combined to form one LF. We refrain from such a merge. First, because these LFs can still be clearly distinguished by humans; see also Polguère (forthcoming) on criteria for the definition of a distinct LF. And second, our goal was to see how the NN-classifier performs when applied to the stock of LFs being used by lexicographers.

5.1.1. LFs used in experiment 1

The following five LFs were considered in Experiment 1: Oper_1, ContOper_1, $\text{Caus}_2\text{Func}_1$, IncepFunc_1 and FinFunc_0. Note the glosses and examples for each:

Oper_1 'experience an emotion'; e.g.:

sentir [*la*] *admiración* lit. 'feel [the] admiration', [*la*] *alegría* lit. 'have [the] joy', *experimentar* [*un*] *odio* lit. 'experience [a] hatred', *tener* [*un*] *odio* lit. 'have [a] hatred'.

ContOper_1 'continue to experience an emotion'; e.g.:

guardar [*el*] *entusiasmo* lit. 'keep [the] enthusiasm', *guardar* [*la*] *esperanza* lit. 'keep [the] hope', *conservar* [*el*] *odio* lit. 'conserve [the] hatred'.

Caus$_2$Func$_1$ 'cause (by the object of emotion) the emotion to be experienced'; e.g.:
causar [*un*] *horror* lit. 'cause [a] horror', *dar* [*una*] sorpresa lit. 'give [a] surprise', *provocar* [*la*] *indignación* lit. 'provoke [the] indignation', *despertar* [*el*] *odio* lit. 'awake [the] hatred'.

IncepFunc$_1$ 'an emotion begins to be experienced'; e.g.:
[*la*] *desesperación entra* [*en N*] lit. '[the] despair enters [in *N*]', [*el*] *odio se apodera* [*de N*] lit. '[the] hatred gets hold [of *N*]', [*la*] *ira invade* [*N*] lit. '[the] rage invades [*N*]'.

FinFunc$_0$ 'an emotion ceases to be experienced'; e.g.:
[*la*] *aprensión se disipa* lit. '[the] apprehension evaporates' [*el*] *odio desaparece* lit. '[the] hatred disappears', [*el*] *entusiasmo se desvanece* lit. '[the] enthusiasm vanishes'.

Oper$_1$ and ContOper$_1$ are very similar in terms of their semantics, and possess the same government pattern: the first semantic actant of the noun is the Subject of the verb, and the noun itself its Object. The government pattern of Caus$_2$Func$_1$ slightly deviates from that of Oper$_1$ and ContOper$_1$: it is the second actant of the noun which becomes the Subject of the verb.[9] The semantics of Caus$_2$Func$_1$ differs considerably from the semantics of Oper$_1$ and ContOper$_1$ (see the glosses).

The government patterns of IncepFunc$_1$ and FinFunc$_0$ have in common that the noun is the Subject of the verb – in contrast to the previous three LFs, in which the noun is the Object. However, the structure of IncepFunc$_1$ also requires the Agent (or, Experiencer in the case of emotions) of the noun to be expressed as Object, while the verbal value of FinFunc$_0$ is intransitive. The semantics are, again, rather different: the semantic features of IncepFunc$_1$ are closer to the semantic features of Caus$_2$Func$_1$ than to those of FinFunc$_0$.

5.1.2. LFs used in experiment 2

As in Experiment 1, in Experiment 2, the classification techniques were tested with respect to five different LFs. These LFs were: CausFunc$_0$, Oper$_1$, Oper$_2$, Real$_1$ and Real$_2$. Consider, again, the glosses and examples for each:

CausFunc$_0$ 'cause the existence of the situation, state, etc.'; e.g.:
dar alarma lit. 'give alarm', *celebrar elecciones* lit. 'celebrate elections', *provocar* [*una*] *crisis* lit. 'provoke [*a*] *crisis*', *publicar* [*una*] *revista* lit. 'publish [a] review/journal'.

Oper$_1$ 'perform', 'experience', 'carry out', etc.; e.g.:
dar [*un*] *golpe* lit. 'give [a] blow', *presentar* [*una*] *demanda* lit. 'present [*a*] demand', *hacer* [*una*] *campaña* lit. 'do [a] campaign', *dictar* [*la*] *sentencia* lit. 'dictate [the] sentence'.

Oper$_2$ 'undergo', 'be source of', etc.; e.g.:
someterse [*a un*] *análisis* lit. 'submit (oneself to an) analysis', *afrontar* [*el*] *desafío* lit. 'face [the] challenge', *hacer* [*un*] *examen* lit. 'do [a] examination', *tener* [*la*] *culpa* lit. 'have [the] blame'.

Real$_1$ 'act accordingly to the situation', 'use as foreseen'; e.g.:
ejercer [*la*] *autoridad* lit. 'exercise [the] authority', *utilizar* [*el*] teléfono lit. 'use [the] telephone', *hablar* [*una*] *lengua* lit. 'speak [a] language', *cumplir* [*la*] *promesa* lit. 'fulfil [the] promise'.

Real$_2$ 'react accordingly to the situation'; e.g.:
responder [*a la*] *objeción* lit. 'respond to the objection', *satisfacer* [*el*] *requisito* lit. 'satisfy [*the*] requirement', *atender* [*la*] *solicitud* lit. 'attend [the] petition', *rendirse* [*a la*] *persuasión* lit. 'render (oneself) [to the] conviction'.

The meanings of Oper$_1$ and Oper$_2$ are very similar, and so are those of Real$_1$ and Real$_2$. Also, in some cases, we found virtually no distinction between the semantic description of the instances of CausFunc$_0$ and Oper$_1$. Consider, for instance, *rendir* [*un*] *homenaje* lit. 'render [an] homage', *dar* [*una*] *explicación* lit. 'give [an] expla-

[9] Thus, in *Los comentarios de algunos politicos provocan la indignación de los vecinos del Carmel* lit. 'The comments of some politicians cause indignation of the neighbors of Carmel', *comentarios de algunos politicos* is the second actant of INDIGNACIÓN (with *los vecinos del Carmel* being the first actant).

Table 2
Sizes of the LSC-sets used in Experiment 1

$Caus_2Func_1$	$ContOper_1$	$FinFunc_0$	$Oper_1$	$IncepFunc_1$
63	14	40	37	23

Table 3
Sizes of the LSC-sets used in Experiment 2

$CausFunc_0$	$Oper_1$	$Oper_2$	$Real_1$	$Real_2$
53	87	48	52	53

nation', *hacer* [*un*] comentario lit. 'do [a] comment' and *poner* [*una*] *queja* lit. 'put [a] complaint', which have been classified as $Oper_1$-instances by human experts. However, to a certain extent, they also express a 'causation of existence'. Wanner (2004) rated the classification as correct if one of such $Oper_1$-instances was classified as $CausFunc_0$. In our current experiments, we applied a more rigorous evaluation rating classifications of this kind as false. This was done in order to keep up with the classification granularity suggested by lexicographers (see also above).

5.2. Data used in the experiments

For Experiment 1, a collection of Spanish LSCs already classified in terms of LFs in the *Diccionario de colocaciones del español* (Alonso, 2004a) has been used; cf. Table 2 for the number of instances available for each of the LFs in Experiment 1.

The data for Experiment 2 have been compiled drawing upon various sources: (i) informants (native speakers of Spanish): two linguists working within the framework of the *Explanatory Combinatorial Lexicology* and a layperson with a pronounced intuition with respect to the acceptability of idiosyncratic combinations; (ii) *Collins* bilingual English–Spanish dictionary; (iii) corpora, where we looked up verb–noun combinations for sets of predetermined nouns (choosing combinations that were instances of one of the relevant LFs).

Table 3 summarizes the sizes of the LSC-sets used in Experiment 2.

When dividing the available material into training and test material, the following two observations should be kept in mind.

- In certain corpora, the material for specific LFs will be scarce. Stevenson et al. (2004) argue that even the *British National Corpus* does not give a broad coverage of SVCs, which are the most common LSCs.
- The optimal size of a training set for a given LF depends on the semantic heterogeneity of the collocations to be classified. In general, it can be assumed that collocations that belong to the same semantic field (such as, e.g., emotions, speech acts, communicative actions, movement actions, etc.) are more homogeneous than collocations that belong to different semantic fields. For instance, for $ContOper_1$ in the field of Spanish emotion nouns, only two values are available: *conservar* lit. 'conserve' and guardar lit. 'keep', which possess in SpanWN the same semantic description. This means that even a very small sized training set can be assumed to suffice. For $IncepFunc_1$ in the same field, we have three different values (*apoderarse, entrar and invadir*) with rather different semantic descriptions. That is, a larger training set is needed to achieve a comparable quality of classification. With the increasing number of semantic fields to be covered, the size of the training set further increases. The results of the experiments give information on this issue.

To explore the minimal and optimal sizes of admissible training sets, experiments with different sizes of the training sets are necessary. We accomplished this as follows. In both experiments, for each LF, x% of the available LSC-set has been used as training material. The remaining $100 - x$% of the LSC-sets of all five

Table 4
Experiment 1: The quality figures (as $p|r$) of the NN-classification of emotion bigrams over different ratios of the training set size

LF	Ratio of the training set size					
	5%	10%	25%	50%	75%	95%
$\text{Caus}_2\text{Func}_1$	0.67|0.75	0.95|0.99	0.78|0.79	0.84|0.81	0.88|0.84	0.84|0.84
ContOper_1	0.67|0.74	0.93|0.87	0.79|0.70	0.83|0.73	0.87|0.77	0.95|0.75
FinFunc_0	0.96|0.62	1.0|0.94	0.89|0.69	0.92|0.71	0.95|0.73	0.95|0.76
IncepFunc_1	0.54|0.51	0.73|0.97	0.65|0.80	0.70|0.92	0.71|0.95	0.70|0.96
Oper_1	0.77|0.79	1.0|0.76	0.81|0.86	0.83|0.89	0.85|0.92	0.87|0.93

LFs drawn upon in an experiment made up the test material. i.e., from the perspective of a specific LF, the test material consisted in $100 - x$% of its LSC-set as positive test data and in $100 - x$% of the LSC-sets of the other four LFs as negative test data. Tests have been performed with $x = 5\%$, 10%, 25%, 50%, 75% and 95%.

6. Experiments

All experiments were carried out with non-disambiguated test material.[10] In SpanWN, the elements of test bigrams usually have more than one sense. Therefore, we had to build the cross-product of all possible readings of each test bigram. In other words, if we assume that for a given bigram (N, V), the noun N encounters s_N senses and the verb Vs_V senses, $\{Se_1^N, Se_2^N, \ldots, Se_{s_N}^N\} \times \{Se_1^V, Se_2^V, \ldots, Se_{s_V}^V\}$, where $Se_i^N (1 \leqslant i \leqslant s_N)$ is one of the nominal senses and $Se_j^V (1 \leqslant j \leqslant s_V)$ one of the verbal senses, was used. To classify a given candidate word bigram as an instance of one of the LFs in the typology, each sense bigram (Se_i^N, Se_j^V) of this word bigram has been examined. Obviously, only one of the (Se_i^N, Se_j^V) may qualify the word bigram as an instance of a specific LF.[11] However, as is well-known, the distinction of word senses in SpanWN is biased towards English, which means that sense distinctions are made for a Spanish word if the corresponding readings are available for the English original – even if they are not available in Spanish; cf. Wanner et al. (2004) for examples. As a result, Spanish words are often assigned several incorrect senses – which has negative consequences for the quality of the classification procedure. To minimize these consequences, we used the so-called *voting* strategy: instead of choosing ONE sense bigram as evidence that the word bigram is instance of the LF L, each sense bigram "voted" for an LF; the word bigram was assigned the LF-label with most votes.

To eliminate a distortion of the experiment outcomes by the selection of the training samples, for each ratio of the training set size (i.e., 5%, 10%, 25%, 50%, 75% and 95%), experiments were performed in 200–500 iterations. In each iteration, the training samples were chosen randomly. The quality figures cited below reflect the average performance over all iterations.

Table 4 shows the performance of the NN-classification for the field of emotion nouns; here and henceforth, 'p|r' stands for 'precision|recall'.[12] For all LFs, except for Oper_1, the ratio of 10% provides the highest f-score: 0.97 for $\text{Caus}_2\text{Func}_1$, 0.9 for ContOper_1, 0.97 for FinFunc_0, and 00.83 for IncepFunc_1.[13] This means that when 10% of the material available for the LF L is taken for training, the share of training instances for the LF L′ which are semantically similar to candidate bigrams for L is the smallest. For Oper_1, the ratio of 95% led to a slightly better f-score than 10%, which is the second best (0.9 compared to 0.86).

[10] Recall, however, that we train on manually disambiguated LF-instances.

[11] For instance, *take [a] rest* is an instance of the Oper_1-LF only for *rest* in the sense of 'relaxation' (and not in the sense of 'peace', 'support', or 'remainder') and *[to] take* in the "emptied" sense of a support verb (and not in the sense of 'remove', 'steal', 'capture', 'accept', 'buy', or any other of its other numerous senses).

[12] As usual, we define p(recision) and r(ecall) as $p(i) = \frac{|LF_{ci}|}{|LF_{pe}|}$ and $r(i) = \frac{|LF_{ci}|}{|LF_i|}$, where $|LF_{ci}|$ is the number of testset elements correctly classified as the LF i, $|LF_{pe}|$ is the total number of testset elements classified as the LF i, and $|LF_i|$ is the total number of testset elements available for the LF i.

[13] We use an equal weighting of p and r to calculate the f-score, i.e., $f = \frac{2pr}{p+r}$.

Table 5
Experiment 2: The quality figures (as $p|r$) of the NN-classification of field-independent bigrams over different ratios of the training set size

LF	Ratio of the training set size					
	5%	10%	25%	50%	75%	95%
CausFunc_0	0.34|0.40	0.45|0.54	0.52|0.64	0.56|0.75	0.59|0.78	0.59|0.79
Oper_1	0.34|0.38	0.41|0.40	0.47|0.49	0.53|0.52	0.55|0.52	0.65|0.55
Oper_2	0.34|0.35	0.44|0.35	0.55|0.49	0.60|0.60	0.61|0.66	0.62|0.71
Real_1	0.35|0.30	0.42|0.40	0.47|0.55	0.56|0.47	0.58|0.44	0.58|0.44
Real_2	0.32|0.34	0.39|0.41	0.49|0.43	0.55|0.46	0.58|0.51	0.56|0.55

Table 5 displays the performance of NN-classification for field-independent candidate bigrams. Given the heterogeneous semantics of both the training and test material samples, it is not surprising that the overall quality figures are lower than in Experiment 1. Unlike in Experiment 1, both p and r generally increase for all LFs with the increasing ratio. Contrary to this general trend are the recall for Real_1, which slightly decreases with the ratios of 75% and 95% when compared to 50%, and the precision for Real_1 with the ratio of 95% (when compared to p with 75%). This is due to the similar semantics of Real_1 and Real_2: with the increasing ratio, the share of Real_1 training instances that are similar to Real_2-instances inevitably increases, as does the share of Real_2 training instances that are similar to Real_1-instances.

7. Evaluation of the experiments

Experiments 1 and 2 provide information on the following two topics:[14]

(1) Should LF-oriented collocation classification be pursued separately for each semantic field, or can we avoid the cost of grouping candidate bigrams into semantic fields?
(2) What can be said concerning the training set size?

The experiments show a considerably better performance of the NN-classifier when it is applied to single field material than when it is applied to multiple field material i.e., semantically sufficiently homogeneous training and test material will always lead to a higher quality LF-classification. However, it must be also taken into account that the field of emotion nouns is extremely homogeneous. We hypothesize that rarely any other field will be as homogeneous as the field of emotion nouns – with the consequence that the quality figures will be lower. In other words, at this stage, we cannot make any reliable statement on the general preference of single field collocation classification. Our experiments are only a first indication that it might be so. Experiments with other semantic fields are needed to buttress this indication.

The experiments also reveal interesting details concerning the size of the training sets: although training sets must contain a sufficient number of samples for a ML-technique to perform well, larger training sets do not automatically stand for a better performance.

In a diffierent run of Experiment 2, we restricted the size of all training sets to 28 – independently of how many instances of an LF were present in our material. Table 6 shows the performance of the NN-classifier with this setup and LSC-set cardinalities as listed in Table 3.

In general, 28 training instances turned out to be too few to achieve optimal accuracy. However, this has already been demonstrated in the previous section. A more interesting issue is how the equal size for all training sets influences the performance. For CausFunc_0, Oper_2, Real_1, and Real_2, the training set of 28 samples approximately corresponds to the 50% ratio in Section 6, and for Oper_1 to the 25% ratio. That is, compared to the Experiment 2 run with the 25% ratio, the uniform size run contains more training instances for Caus-

[14] A further topic which is certainly also of outmost relevance concerns the suitability of SpanWN as an external lexico-semantic resource. For the evaluation of SpanWN in the context of collocation classification, we refer the interested reader to Wanner (2004).

Table 6
Performance of the NN-classifier (as $p|r$) with a training set size of 28 for each LF

CausFunc_0	Oper_1	Oper_2	Real_1	Real_2
0.31|0.78	0.85|0.49	0.33|0.62	0.44|0.46	0.46|0.42

Table 7
The *f*-scores for LFs in Experiments 1 and 2 with the 95% training size ratio

	$\text{Caus}_2\text{Func}_1$	ContOper_1	FinFunc_0	IncepFunc_1	Oper_1
Experiment 1	0.84	0.84	0.84	0.81	0.90
Experiment 2	0.68	0.60	0.66	0.50	0.55

Table 8
Typical verbs in collocations covered by the LFS in Experiment 2

CausFunc_0:	INICIAR 'initiate', CREAR 'create'
Oper_1:	HACER 'do', DAR 'give', EXPERIMENTAR 'experience'
Oper_2:	PERCIBIR'perceive', SUFRIR 'undergo'
Real_1:	UTILIZAR 'apply, utilize, employ', PONER 'put'
Real_2:	RESPONDER 'respond', SEGUIR 'follow'

Table 9
The *f*-scores for LFs in Experiment 2 with the 95% training size ratio using ID3-, NB-, and TAN-classifiers, the *f*-scores achieved in (Wanner, 2004) (abbreviated as 'LW') with manually disambiguated data, and the baseline performance

	CausFunc_0	Oper_1	Oper_2	Real_1	Real_2
Baseline	0.42	0.40	0.12	0.19	0.33
ID3	0.58	0.68	0.46	0.44	0.51
NB	0.59	0.74	0.30	0.45	0.47
TAN	0.50	0.59	0.55	0.49	0.45
LW	0.76	0.60	0.75	0.74	0.58

Func0, Oper_2, Real_1 and Real_2; compared to the run with the 50% ratio, it contains less Oper_1 training instances.

The reduced uniform training set size led to a (partially) considerably lower *f*-score for all LFs except for Oper_1. For Oper_1, in particular *p* was significantly higher with the uniform size. Compare the figures in Table 6 with the corresponding figures in Section 6.

The quality figures gained in the experiments and the above evaluation allow for a concluding assessment of the NN-classifier in the context of LSC-classification with respect to the LF-typology using external semantic resources. Table 7 shows the *f*-scores in Experiments 1 and 2 with the 95% ratio (see Table 8).

To examine the relative performance quality of the NN-classifier in the task of LSC-classification, we carried out Experiment 2 with three further ML-techniques, namely the decision tree algorithm ID3 (Quinlan, 1986), the Naïve Bayes (NB) classifier (Mitchell, 1997) and the *Tree Augmented Bayes Network* (TAN) classifier (Friedman et al., 1997). For all three techniques, the total set of attribute variables was assumed to be given by $\mathscr{B} \cup \mathscr{C}$, i.e., each meaning component was considered an attribute. Table 9 summarizes the quality figures obtained for these three techniques, the accuracy achieved in comparable experiments in Wanner (2004),[15] along with a baseline. The baseline is the match of the verb of a candidate bigram with one of

[15] A word of caution is in order here: strictly speaking, we cannot directly compare the results of the experiments described in this article with the results in Wanner (2004) since in Wanner (2004), we used manually disambiguated test data.

the most common collocates of the LF in question. The most common collocates used for the LFs drawn upon in Experiment 2 are summarized in Table 8. We have taken this baseline because in Explanatory Combinatorial Lexicology Mel'čuk et al. (1995), most common collocates of an LF tend to be considered adequate glosses of the meaning of this LF.

The TAN-classifier performs best, when clear component correlations in both the training and test samples can be identified. The NB-classifier is suitable if the instances of the individual LFs have distinctive meaning components (as, e.g., the synonyms of *disiparse* lit. '[to] evaporate', which is typical of FinFunc_0 in the emotion noun field). The ID3-algorithm is unreliable in single field experiments, but outperforms, e.g., TAN in the experiments with more heterogeneous material. However, in general, the NN-classifier proved to be the most reliable ML-technique for our task. All techniques examined perform considerably better than the baseline.

8. Related work

Sag et al. (2002) call the problem of handling MWEs "a pain in the neck for NLP". An increasing number of works attempts to contribute to its cure. In this section, we discuss mainly those of them that deal with the problem of collocation recognition in corpora and collocation classification. The collocation recognition task is immediately relevant to this article because, as shown in Wanner et al. (2005b), the techniques discussed can be well applied for the extraction of collocations from corpora; see also Section 9. It should be pointed out that our work is also related to research in such areas *as acquisition of co-occurrence restrictions* (or *selectional preferences*); see, e.g., (Resnik, 1993; Ribas, 1995; Sanfilippo, 1997; McCarthy, 1997; Li and Abe, 1998; McCarthy, 2000; Clark and Weir, 2002); and *semantic classification* of either single lexical items or binary relations between lexical items using machine learning techniques; consider, e.g., (Siegel, 1999; Merlo and Stevenson, 2001; Rosario and Hearst, 2001). See Wanner (2004) for an overview and their relation to the task of collocation classification using semantic information defined in WordNet.

The overwhelming majority of the approaches to automatic identification and extraction of collocations is based on the interpretation of the notion of collocation as a sequence of words that frequently appear together – either adjacently or interrupted by other words; see, e.g., (Choueka et al., 1983; Church and Hanks, 1989; Smadja, 1993; Justeson and Katz, 1995; Merkel and Andersson, 2000).[16] As a rule, these approaches provide plain lists of presumed collocations, possibly enriched with POS-information. Due to purely statistical techniques applied, no semantic information on the combinations extracted can be provided. Lin (1998) combines statistical techniques with syntactic processing – arguing, as we do, that although collocations are *reccurrent* combinations, they are not necessarily *frequent* combinations. Lin's approach consists of three steps: (1) collection of dependency triples (he also considers the article of the noun in such collocations as *file a lawsuit, weather a storm*, etc.), (2) (automatic) correction of the erroneous frequency counts of the triples that result from parser mistakes, (3) filtering of the triples with mutual information. For (2), syntactic features derived from the WordNet are used. Pearce (2001) proposes the evaluation of the frequency of the co-occurrence of lexical items with synonymous lexemes: if a word W_1 co-occurs with a word W_2 n times and with a synonym of W_2, W_3, m times, and $m < n - 1$, then $W_1 + W_2$ is considered a potential collocation. Between W_1 and W_2 as well as between W_1 and W_3 a specific dependency relation (e.g., 'modifier') must hold. To determine the synonyms of a given lexeme, Pearce uses WordNet.

A considerable number of researchers focus on the extraction of support (or *light*) verb constructions, SVCs, (Grefenstette and Teufel, 1995; Dras, 1995; Tapanainen et al., 1998; Stevenson et al., 2004), which constitute the most prominent kind of verb–noun LSCs.[17] The basic difference between these works and ours is that we rely upon *semantic similarity* of a candidate bigram to samples in reference (training) sets, while they use either stan-

[16] This interpretation is attractive to automatic processing because it allows for the use of well-developed statistical models and does not require other linguistic preprocessing than part of speech tagging.

[17] SVCs are represented by the Oper-LFs, i.e., Oper_i $i = 1, 2, \ldots$, and are thus a subset of LSC-types we draw upon in our classification experiments.

dard statistical measures extended to capture the linguistic properties of SVCs (as Stevenson et al. (2004), who use the *pointwise mutual information* (Church et al., 1991)), or combine frequency counts with morpho-syntactic information on the deverbal nature of verbal complements as Grefenstette and Teufel (1995) and Dras (1995). Furthermore, while they either attempt to find possible deverbal noun complements for a given set of support verbs (as Stevenson et al. (2004)) or probable support verbs for deverbal noun complements (as Grefenstette and Teufel (1995); Dras (1995); Tapanainen et al. (1998)), our approach is perspective-neutral: we consider rather the semantically motivated correlation between the elements of a given bigram.

In general, most approaches to collocation extraction as discussed in the literature can be considered to be complementary to our approach: once binary combinations of lexical items assumed to be collocations have been extracted by the former, our approach can either assign a semantics to them (by identifying the LF to which a given combination belongs) or reject their collocational status. The latter is achieved by introducing into the LF-typology an additional "pseudo LF" that comprises free verb–noun bigrams; see Wanner et al. (2005b) for theoretical details and experiments.

A few recent works draw upon LFs when identifying collocations in the corpus. Thus, Daille (2003) uses morpho-syntactic variations of words to detect instances of mainly derivative LFs such as **Mult** 'multitude' (cf. Mult(FISH) = *school*, Mult(SHEEP) = *flock*), **Gener** 'generic name' (cf. Gener(CARROT) = *vegetable*, Gener(LOVE) = *emotion*), $\mathbf{S}_1$ 'first actant's typical name' (cf. S_1 (SONG) = *singer*, S_1 (STUDY) = *student*), etc. Claveau and L'Homme (2004) exploit the syntagmatic context attempting to detect N–V pairs that qualify for any LF from the **Real**-group Real_i ($i = 1, 2, \ldots$) with the meaning 'act appropriately with respect to the situation', for an LF from the **Fact**-group (Fact_j $j = 0, 1, \ldots$) with the meaning 'be dealt with appropriately', etc. However, to our knowledge, none of the previously cited works proposed techniques for an actual classification of bigrams with respect to the fragment of the LF-typology we are working with. Neither did they achieve such a classification granularity and accuracy.

9. Conclusions and remaining issues

In this article, we described the application of the NN-classifier to the task of the classification of verb–noun collocations, contrasting its performance with the performance of several other ML-techniques. We used the typology of lexical functions as the classification schema and the hyperonym hierarchies provided by SpanWN as the source for the semantic componential description of the lexical items involved. The techniques have been implemented and applied to Spanish material. Experiments have been carried out on material from the emotion noun field and on material with no field restrictions.

The experiments demonstrated that the techniques proposed are able to provide a high quality classification of verb–noun collocations. In the experiments described elsewhere (see Wanner et al. (2005b)), we show that these techniques can be used to classify in terms of LFs any verb–noun bigrams extracted from a corpus, i.e., not only bigrams that a *priori* are known to be an instance of an LF (although, not of which LF). As pointed out in the previous section, this requires the extension of the LF-typology by a pseudo LF that subsumes free verb–noun combinations. Obviously, this implies the consideration of training instances for this pseudo LF during the learning stage.

We plan to use the developed system within the DICE-Project (Alonso, 2004a). Our work can also be used in a broader scenario, for example, for the following purposes:

(i) classifying verb–noun bigrams with a specific syntactic structure that have been acquired from a corpus by partial parsing in terms of LFs;
(ii) assignment of semantics to collocations listed in (collocation) dictionaries or extracted from a corpus by a technique that provides plain lists of collocations (this corresponds to the experiment setting described in the paper);
(iii) filtering out word combinations that have erroneously been classified as collocations (these will be instances falling into the class of "non-collocation"-LF).

An interesting by-product of the use of EuroWordnet is the semantic disambiguation of the bigram elements by the classification procedure. This is because the result of the classification is not a claim that a particular word-bigram is an instance of a given LF, but a claim that particular SpanWN-senses of the words in this bigram form an instance of this LF.

Several issues still remain to be tackled at this point. The most important of them being, first, the use of additional LFs from the LF-typology, including further verb–noun LFs such as **Son** 'typical sound': *dog barks, teeth chatter, hurricane roars*, etc. and **Degrad** 'deteriorate': *teeth decay, temper frays, discipline decays*, etc. as well as adverb–verb LFs. Experiments on the classification of adjective–noun LSCs are described in Wanner et al. (2005a). Second, the ease of the dependency on external semantic information as given in EWN. The goal is to use a mixture of contextual and lexico-semantic information for LF-oriented collocation classification. Currently, experiments are under way with German material. Third, extention of our work to English and French. English is well-suited for our experiments. Thus, the Princeton WN is, in combination with the English part of the EuroWN, the most detailed and exhaustive lexico-semantic resource available to date for a language. Furthermore, well-balanced extensive training sets for all LFs can be readily compiled for English from the LF-base that is publicly available from I. Mel'čuk. For French, a machine readable LF-dictionary is already available Polguère (2000). LF-instances in this dictionary can be used as bootstrapping seeds for the classification algorithm. This would make the compilation of the training sets obsolete and thus contribute to the efficiency of the approach.

References

Abeillé, A., 1988. Light verb constructions and extraction out of NP in a tree adjoining grammar. In: Papers of the 24th Regional Meeting of the Chicago Linguistics Society. Chicago.

Allerton, D.J., 1984. Three or four levels of co-occurrence relations. Lingua 63, 17–40.

Alonso, M., 2004a. Elaboración del diccionario de colocaciones del español y sus aplicaciones. In: Battaner, P., DeCesaris, J. (Eds.), Symposium Internacional de Lexicografía. IULA, Universitat Pompeu Fabra, Barcelona, pp. 149–162.

Alonso, M., 2004b. Las construcciones con verbo de apoyo. Visor Libros, Madrid.

Baldwin, T., Bannard, C., Tanaka, T., Widdows, D., 2003. An empirical model of multiword expression decomposibility. In: Proceedings of the Workshop on Multiword Expressions at the 41st Annual Meeting of the ACL, pp. 89–96.

Benson, M., 1989. The structure of the collocational dictionary. International Journal of Lexicography 2 (1), 1–13.

Choueka, Y., Klein, T., Neuwitz, E., 1983. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. Journal of Literary and Linguistic Computing 4 (1), 34–38.

Church, K.W., Hanks, P., 1989. Word association norms, mutual information, and lexicography. In: Proceedings of the 27th Annual Meeting of the ACL, pp. 76–83.

Church, K.W., Gale, W., Hanks, P., Hindle, D., 1991. Using statistics in lexical analysis. In: Zernik, U. (Ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. Erlbaum, Hillsdale, NJ, pp. 116–164.

Clark, S., Weir, D., 2002. Class-based probability estimation using a semantic hierarchy. Computational Linguistics 28 (2), 187–204.

Claveau, V., L'Homme, M.C., 2004. Discovering specific semantic relationships between nouns and verbs in a specialized French corpus. In: Proceedings of the 3rd International Workshop on Computational Terminology, pp. 39–46.

Cowie, A.P., 1994. Phraseology. In: Asher, Simpson (Eds.), The Encyclopedia of Language and Linguistics, vol. 6. Pergamon, Oxford, pp. 3168–3171.

Daille, B., 2003. Concept structuring through term variations. In: Proceedings of the Workshop on Multiword Expressions at the 41st Annual Meeting of the ACL, pp. 9–16.

Dras, M., 1995. Automatic identification of support verbs: a step towards a definition of semantic weight. In: Proceedings of the 8th Australian Joint Conference on Artificial Intelligence. World Scientific, Singapore, pp. 451–458.

Fellbaum, Ch. (Ed.), 1998. WordNet. An Electronic Lexical Database. The MIT Press, Cambridge, MA.

Firth, J.R. (Ed.), 1957. Papers in Linguistics 1934–1951. Oxford University Press, Oxford.

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Machine Learning 29 (2–3), 131–163.

Grefenstette, G., Teufel, S., 1995. Corpus-based method for automatic identification of support verbs for nominalizations. In: Proceedings of the Biannual Meeting of the EACL, pp. 27–31.

Halliday, M.A.K., 1966. Lexis as a Linguistic Level. In: Bazell, C.E. et al. (Eds.), Memory of J.R. Firth. Longman, London.

Justeson, J.S., Katz, S.M., 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering 1, 9–27.

Kahane, S., Polguère, A., 2001. Formal foundation of lexical functions. In: Proceedings of the Workshop Collocation: Computational Extraction, Analysis and Exploitation at the 39th Annual meeting of the ACL, Toulouse, pp. 8–15.

Li, H., Abe, N., 1998. Generalizing case frames using a thesaurus and the MDL principle. Computational Linguistics 24 (2), 217–244.

Lin, D., 1998. Extracting collocations from text corpora. In: Proceedings of the First Workshop on Computational Terminology, pp. 57–63.

McCarthy, D., 1997. Word sense disambiguation for acquisition of selectional preferences. In: Proceedings of the Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources at the 35th Annual Meeting of the Association for Computational Linguistics, pp. 52–60.

McCarthy, D., 2000. Using semantic preferences to identify verbal participation in role switching. In: Proceedings of the 1st Conference of the NAACL, pp. 256–263.

Mel'čuk, I.A., Wanner, L., 1996. Lexical functions and lexical inheritance for emotion lexemes in German. In: Wanner, L. (Ed.), Lexical Functions in Lexicography and Natural Language Processing. Benjamins Academic Publishers, Amsterdam, pp. 209–278.

Mel'čuk, I.A., Clas, A., Polguère, A., 1995. Introduction à la lexicologie explicative et combinatoire. Duculot, Louvain-la-Neuve.

Mel'čuk, I.A., et al., 1984, 1988, 1992, 1999. Dictionnaire explicatif et combinatoire du français contemporain, vol. I–IV. Presses de l'Université de Montréal, Montréal, Canada.

Mel'čuk, I.A., 1995. Phrasemes in language and phraseology in linguistics. In: Everaert, M., van der Linden, E.-J., Schenk, A., Schreuder, R. (Eds.), Idioms: Structural and Psychological Perspectives. Lawrence Erlbaum Associates, Hillsdale, pp. 167–232.

Mel'čuk, I.A., 1996. Lexical functions: a tool for the description of lexical relations in a lexicon. In: Wanner, L. (Ed.), Lexical Functions in Lexicography and Natural Language Processing. Benjamins Academic Publishers, Amsterdam, pp. 37–102.

Merkel, M., Andersson, M., 2000. Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In: Proceedings of the RIAO Conference, pp. 737–746.

Merlo, P., Stevenson, S., 2001. Automatic verb classification based on statistical distributions of argument structure. Computational Linguistics 27 (3), 373–408.

Miller, G., 1998. Nouns in WordNet. In: Fellbaum, C. (Ed.), WordNet. An Electronic Lexical Database. The MIT Press, Cambridge, MA, pp. 23–46.

Mitchell, T., 1997. Machine Learning. McGraw-Hill.

Moon, R., 1998. Fixed Expressions and Idioms in English: A Corpus-based Approach. Clarendon Press, London.

Pearce, D., 2001. Synonymy in collocation extraction. In: Proceedings of the Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations at the Annual Meeting of the NAACL, Pittsburgh, pp. 41–46.

Polguère, A., 2000. Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. In: Proceedings of EURALEX'2000, pp. 517–527.

Polguère, A., forthcoming. Lexical function standardness. In: L. Wanner (Ed.), Selected Lexical and Grammatical Topics in the Meaning-Text Theory. In honour of Igor Mel'čuk. Benjamins, Amsterdam.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1, 81–106.

Resnik, P., 1993. Selection and Information: A Class-based Approach to Lexical Relationships. PhD thesis, University of Pennsylvania.

Ribas, F., 1995. On learning more appropriate selectional restrictions. In: Proceedings of the Biannual Meeting of the EACL, pp. 112–118.

Rosario, B., Hearst, M., 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In: Proceedings of the 6th Conference on Empirical Methods in NLP, pp. 82–90.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D., 2002. Multiword expressions: a pain in the neck for NLP. In: Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City, Mexico, pp. 1–15.

Salton, G., 1980. Automatic term class construction using relevance: a summary of work in automatic pseudoclassification. Information Processing and Management 16 (1), 1–15.

Sanfilippo, A., 1997. Using semantic similarity to acquire cooccurrence restrictions from corpora. In: Proceedings of the Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources at the 35th Annual Meeting of the ACL, pp. 82–89.

Siegel, E.V., 1999. Corpus-based linguistic indicators for aspectual classification. In: Proceedings of the 37th Annual Meeting of the ACL, pp. 112–119.

Smadja, F., McKeown, K., 1991. Using collocations for language generation. Computational Intelligence 7 (4), 229–239.

Smadja, F., 1993. Retrieving collocations from text: X-tract. Computational Linguistics 19 (1), 143–177.

Stevenson, S., Fazly, A., North, R., 2004. Statistical measures of the semi-productivity of light verb constructions. In: Proceedings of the Workshop on Multiword Expressions: Integrating Processing at the 42nd Annual Meeting of the ACL, pp. 1–8.

Tapanainen, P., Piitulainen, J., Järvinen, T., 1998. Idiomatic object usage and support verbs. In: Proceedings of the COLING/ACL '98, pp. 1289–1293.

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

Wanner, L., 2004. Towards automatic fine-grained semantic classification of verb–noun collocations. Natural Language Engineering Journal 10 (2), 95–143.

Wanner, L., Alonso, M., Martí, A., 2004. Enriching the Spanish Wordnet with collocations. In: Proceedings of the LREC '04, Lisbon, pp. 1087–1090.

Wanner, L., Bohnet, B., Alonso, M., Vázquez, N., 2005a. The true, deep happiness: towards the automatic semantic classification of adjective–noun collocations. In: Kiefer, F., Kiss, G., Pajzs, J. (Eds.), Papers in Computational Lexicography. COMPLEX 2005. Hungarian Academy of Sciences, Budapest, pp. 255–265.

Wanner, L., Bohnet, B., Giereth, M., Vidal, V., 2005b. The first steps towards the automatic compilation of specialized collocation dictionaries. Terminology 11 (1), 137–174.