Natural Language Engineering **10** (2): 95–143. © 2004 Cambridge University Press
DOI: 10.1017/S1351324904003328 Printed in the United Kingdom

Towards automatic fine-grained semantic classification of verb-noun collocations

LEO WANNER

Computer Science Department, University of Stuttgart, Universitätsstr. 38, 70569 Stuttgart, Germany
e-mail: wanner@informatik.uni-stuttgart.de

(*Received 12 February 2002; revised 23 September 2002*)

Abstract

Plain lists of collocations as provided to date by most approaches to automatic acquisition of collocations from corpora are useful as a resource for dictionary construction. However, their use is rather limited in the case of NLP-applications such as Text Generation, Machine Translation and Text Summarization if not enriched by information on the grammatical function of the collocation elements and by information on the semantics of the collocations as multiword units. In this article, we describe an approach to a fine-grained classification of verb-noun bigrams according to a semantically motivated typology of collocations and illustrate this with Spanish material. The typology of collocations that underlies our classification is based on verb-noun Lexical Functions (LFs) from the Explanatory Combinatorial Lexicology. In the first stage of the approach, the program learns the semantic features of each LF from training data. In the second stage, it examines the semantic features of verb-noun candidate bigrams and compares them with the features of all the LFs taken into account. A candidate whose features are sufficiently similar to those of a specific LF is considered to be an instance of this LF. The semantic features of both the training material and the candidate bigrams are derived from the hyperonymy hierarchies provided by the EuroWordNet. In the experiments carried out to validate the approach, we achieved an average f-score of about 70%.

1 Introduction

The present article deals with the task of classifying verb-noun bigrams according to a semantically motivated typology of collocations. This task can be considered a particular stage in the process of the automatic acquisition of verb-noun collocations from the corpus, which takes place after a partial parsing stage, in which verb-noun patterns are retrieved, and, possibly also after a disambiguation stage, in which the elements of the retrieved verb-noun combinations are disambiguated. It can equally be considered a stand-alone NLP-application that processes lists of syntactically annotated verb-noun pairs provided by external sources.[1]

[1] For instance, for Engish, such a verb-noun list has been compiled using the Fidditch parser (Hindle, 1983); this list is available for various NLP-applications. Such lists can equally be obtained for German by tools described, for example, in Christ *et al.* (1999) and Klatt (1997), for French by tools described, for instance, in Giguet and Vergne (1997), and for a number of other languages.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 1. *Some high probability associations of lexical items with the adjectives* STRONG *and* POWERFUL *from Manning and Schütze (1999:156) (C stands for frequency in the corpus)*

w	*C*(STRONG,*w*)	*w*	*C*(POWERFUL,*w*)
support	50	force	13
safety	22	computers	10
sales	21	position	8
opposition	19	men	8
showing	18	computer	8
sense	18	man	7
message	15	symbol	6
defense	14	military	6
gains	13	machines	6
…	…	…	…
case	11	Germany	5
…	…	…	…
signal	9	neighbor	4
man	9	magnet	4

Since the term "collocation" has at least two incompatible interpretations, let us, before sketching the goals and the model underlying our approach, clarify the usage of the term in this paper.

1.1 On the notion of collocation

The term "collocation" as introduced by Firth (1957) and made more explicit by Halliday (1961; 1966) encompasses the statistical distribution of lexical items in context; cf. *Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item x, the items a, b, c …* (Halliday, 1961:276). In other words, lexical items that form high probability associations are collocations. To illustrate this definition, Manning and Schütze (1999) cite a list of nouns that form high probability associations with the lexical items STRONG and POWERFUL. A part of this list, which has been retrieved by Justeson and Katz's (1995) collocation extraction program, is reproduced in Table 1.

It is this interpretation that underlies, e.g., the COBUILD dictionary (Sinclair, 1994) and that is drawn upon in *contextual* lexical models, i.e. in association nets in which each lexical item is linked to all other lexical items it co-occurs with in a representative corpus (e.g. see Williams 1999).

Along with the statistical interpretation of the term, in contemporary lexicography and lexicology, an interpretation that stresses the idiosyncratic nature of collocations has gained acceptance. Thus, Benson (1989) states that "collocations should be defined not just as 'recurrent word combinations', but as 'ARBITRARY recurrent word combinations'". "Arbitrary" as opposed to "regular" means that collocations are not predictable and cannot be translated word by word. For instance, in English one

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 2. *Some idiosyncratic associations of lexical items*

	Base	Collocate		Base	Collocate
1.	⟨NOUN⟩	⟨ADJ⟩	3.	⟨VERB⟩	⟨ADV⟩
	bachelor	confirmed		hit	hard
	rain	heavy		breath	heavily
	demand	legitimate		walk	steadily
	smoker	heavy		cut	neatly
2.	⟨NOUN⟩	⟨VERB⟩	4.	⟨VERB⟩	⟨PREP⟩
	accusation	prove		struggle	with
	hint	take		complain	of
	analysis	carry out		draw	on
	attention	receive		recall	from

takes a walk, in French, German, and Italian one 'makes' it (*faire une promenade*, *einen Spaziergang machen*, *fare una passeggiata*), and in Spanish one 'gives' it (*dar un paseo*); in English one *gives a lecture* – 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.

Many authors stress the particular syntactic and/or semantic features of lexical items that form a collocation. According to Cowie (1993), collocations are "associations of two or more lexemes (or roots) recognized in and defined by their occurrence in a specific range of grammatical constructions".[2] Hausmann (1984) and Mel'čuk (1995) explicitly reject recurrency as a criterion for a word combination to be considered a collocation.[3] For them, a collocation is a combination of two lexical items in which the semantics of one of the lexical items (the *base*) is autonomous from the combination it appears in, and where the other lexical item (the *collocate*) adds semantic features to the semantics of the base. Thus, in *take a walk*, the base is *walk* and the collocate is *take*; in *give a lecture*, the base is *lecture* and the collocate is *give*. As the examples show, the semantics of a word in its function as collocate does not need to be same as the semantics of the same word in isolation.

As Firth's interpretation, Mel'čuk and Hausmann's interpretation covers not only 'verb-noun' combinations, but also 'verb-adverb', 'noun-adjective', 'noun-noun', 'adverb-adjective', and, with some restrictions, also 'noun-preposition' and 'verb-preposition' combinations. For illustration, consider a short list of collocations in Hausmann's and Mel'čuk's sense in Table 2. For a detailed discussion of the syntactic patterns of collocations in general, consult, for instance, Heid (1992) and Fontenelle (1997).

As the examples in Tables 1 and 2 suggest, the two interpretations of the term 'collocation' do not necessarily lead to a different judgement with respect to the collocation status of a word combination. Just the contrary: two lexical items that

[2] 'Lexeme' stands here for a word taken in one well-defined sense.
[3] See, among others, Heid (1992) for a contrastive presentation of Hausmann's and Mel'čuk's approaches.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

form an "arbitrary" recurrent (or idiosyncratic) combination are likely to occur together in the corpus. For instance, *strong tea* is a typical combination of this kind (and is thus a collocation in the sense of Hausmann and Mel'čuk). It is arbitrary because (a) we cannot predict that *strong* (and not *powerful*) must be used in combination with *tea* and *coffee* to express the meaning 'with much caffeine' (we must learn this by heart), and (b) the meaning of *strong* in combination with *tea* and *coffee* deviates from the literal meaning of *strong* (as, in free combination with, for instance, *person*: *strong person*). The combination *strong tea* will also occur in the corpus much more often than *powerful tea* (which we will probably not find at all). That is, *strong tea* is also a collocation in Firth's sense.[4] However, such combinations as *strong man*, *powerful Germany*, and *powerful neighbor* also co-occur significantly often in the corpus (see Table 1), which suggests that they are collocations in terms of their statistical distribution. In the idiosyncratic sense, they would not be considered collocations, but rather recurrent free combinations. On the other hand, the combination *feeble imagination* is certainly a collocation in the idiosyncratic sense. However, in the Web pages covered by Google on March 17th 2003, it occured only 293 times (whereas *wild imagination* occured 12.000 times and *vivid imagination* 32.600 times), and, therefore, would most likely not be considered a collocation in the statistical sense.

To summarize, the idiosyncratic interpretation of the term "collocation" covers all non-free combinations also covered by its distributional interpretation, but, in addition, it covers rare idiosyncratic combinations and rejects all recurrent free combinations. Therefore, it is the interpretation of the term "collocation" as an idiosyncratic combination that underlies our study. More specifically, we make use of the following properties implied by this interpretation:

- a collocation is a binary combination of lexical items,
- a collocation possesses a coherent syntactic structure, i.e. the base and the collocate always possess the same grammatical function with respect to each other,
- a collocation is a lexically restricted word combination, i.e. it cannot be constructed using universal (semantic) selectional restriction rules; rather, the base predetermines the set of lexical items (collocates) it may appear with in a combination based on idiosyncratic, collocation type-specific grounds.

Word combinations such as *take/give* [*a*][5] *book*, *make* [*a*] *lamp*, *hold* [*a*] *pen* are binary combinations with a coherent syntactic structure. However, they are not lexically restricted by any of their elements and can thus be constructed in accordance with a set of selectional restrictions. Therefore, they are not collocations (in contrast to, for example, *take* [*a*] *walk*, *give* [*a*] *hint*, etc., which possess all three of the above properties). The phenomenon of collocations is extremely wide-spread. We focus on verb-noun collocations, which go well beyond *support verb constructions*

[4] As a matter of fact, Halliday cites it as an example of a prototypical collocation.
[5] When citing verb-noun combinations we introduce henceforth, where it appears useful for better readability, the article of the noun in brackets. Note, however, that the article is, strictly speaking, not part of the collocation.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

(Allerton, 1984), i.e. 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*. Compare *raise hope*, *grab attention*, *dissolve* [*an*] *assembly*, *meet* [*a*] *demand*, *discipline decays*, *temper frays*, *wind howls*, etc., which all are not support verb constructions.

1.2 The goal: classification of bigrams according to a semantic typology of collocations

Some conventional monolingual general purpose dictionaries provide lists of collocations in which the lemma in question is involved. As a rule, they do not indicate their meanings, which are needed for language understanding and language production. To account for these needs, in an increasing number of collocational dictionaries, the meaning of collocations is specified either implicitly by a predefined order of collocations in an entry (Benson *et al*., 1997) or explicitly (Mel'čuk *et al*., 1984, 1988, 1992, 1999).

Even if we assume that a human speaker can cope with plain lists of collocations, for such NLP-applications as Machine Translation and Text Generation, collocations are hardly of any use if simply listed without any further grammatical or semantic information. To put it in Heid's (1998) words: "A collocational dictionary should be not only a dictionary *of collocations*, but also a dictionary *about collocations*." If the meaning of a collocation is not determined during its retrieval from the corpus, it must be assigned (as a rule, manually) in a subsequent stage – as done, for instance, by Smadja and McKeown (1991).

One way to decide whether a given word combination is a collocation and to determine its meaning is to classify it according to a semantico-syntactic typology of collocations. One such typology is based on *lexical functions* (LFs) as introduced in the *Explanatory Combinatorial Lexicology* (Mel'čuk, 1996). In what follows, we present an *instance-based* machine learning approach to the classification of verb-noun combinations according to LFs. The approach is based on computing the similarity between the meaning of a candidate verb-noun bigram and the meaning of a "prototypical" collocation (i.e. *centroid*) of each LF, using a metric suggested by Ogawa *et al.* (1991) in Information Retrieval. The prototypical collocations of the LFs are computed from training data collected for each LF. The approach has been tested on Spanish material. The Spanish part of the EuroWordNet (Vossen, 1998) is used as the source of meaning description of the Spanish data.

In the following section, we introduce the LF-typology. In Section 3, the theoretical aspects of the LF learning and classification model are outlined. Section 4 presents the description of lexical meanings in terms of hyperonym hierarchies in SpanWN (henceforth, 'SpanWN' stands for the Spanish part of the EuroWordNet). Section 5 describes the prerequisites of the experiments carried out to validate the model proposed, the data which have been used in these experiments, and the LFs which have been drawn upon. Section 6 presents these experiments. Section 7 discusses the results obtained and examines the parameters that had a negative influence on the results. Section 8 contains an overview of the work related to ours, and, finally,

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

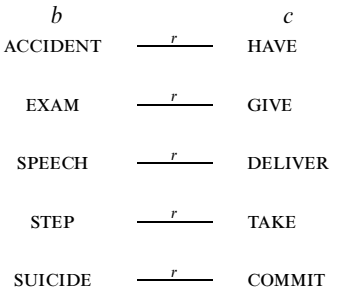


Fig. 1. Illustration of an LF relation.

Section 9 draws the conclusions from our study and presents some of the remaining issues for future research.

2 Lexical functions as a typology of collocations

To introduce the notion of lexical function (LF), we need a more formal definition of the notion of collocation used in this article. Consider the following definition derived from Mel'čuk (1995):

Definition 1 (Collocation)
A collocation A⊕B is a semantic phraseme such that the meaning composition 'X' expressed by A⊕B is constructed out of the signified 'A' and the meaning 'C', i.e. 'X' = 'A⊕C', such that the lexical item B expresses 'C' contingent on A.[6]

A is here the base which keeps its semantics, and B is the collocate which adds semantic features combined in 'C' to 'A' (that is, Definition 1 reflects the notion of collocation as an idiosyncratic combination, rather than as a frequent combination). Each 'C' is denoted by a distinct *syntagmatic* LF – the type of LF that is relevant in our context.[7] A syntagmatic LF is defined as follows:

Definition 2 (Syntagmatic Lexical Function)
A syntagmatic LF is a (directed) standard abstract relation that holds between the base A and the collocate B of the collocation A⊕B and that denotes 'C' ⊂ 'A⊕C' with 'A⊕C' being expressed by A⊕B.

'Standard' means that this relation applies to a large number of collocations. 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). In Figure 1, where *b* stands for 'base' and *c* stands for 'collocate', each *c* contributes to the semantics of the respective *b* the component 'do' or 'perform'. That is, the directed relation *r* that holds between *b* and *c* is 'do' or

[6] '⊕' stands for meaning composition and is thus a union operator.
[7] We consider only regular "standard" LFs. See Mel'čuk (1996) for a more detailed definition of the notion of LF and a comprehensive overview of all types of LFs – including *paradigmatic* LFs that capture the semantic derivatives of lexical items.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

'perform'. The relation *r* also implies in this case that the 'Doer' is the grammatical Subject and *b* is the Direct Object in the default minimal syntactic structure that involves *b*, *c* and a 'Doer'.

'Abstract' means that the meaning of this relation is sufficiently general and can therefore be exploited for purposes of classification.

LFs have been established by I. Mel'čuk and his colleagues from the *Moscow Semantic School* in cross-linguistic empirical studies since the early sixties (see Mel'čuk (1996) for a historical presentation and lists of LF-instances from English, French, Russian, Persian, Arabic, etc.).

In Mel'čuk (1996), about 60 different "simple standard" LFs are distinguished. Compared to the only other existing semantico-syntactic typology we know of, the typology suggested by Benson *et al.* (1997), which consists of eight different types of "grammatical" collocations (= combinations of an open class lexical item with one or several closed class lexical items or with a clause) and seven different types of "lexical" collocations (= combinations of open class lexical items), the LF-typology is thus very fine-grained indeed.

The 60 standard simple LFs are assumed to provide an exhaustive semantico-syntactic typology of the stock of idiosyncratic binary word combinations in natural languages. About twenty of them capture verb-noun collocations. The other forty capture various paradigmatic relations (such as *doctor*∼*patient*, *shoot*∼*target*, *escape*∼*fugitive*, etc.), noun-noun collocations (such as *peak* [*of the*][8] *crisis*, *climax* [*of the*] *struggle*, ...), adjective-noun and adverb-verb collocations (such as *stark naked*, *infinite patience*, *loyal citizen*, *cut neatly*, *fruitful analysis*, etc.), and noun-preposition and verb-preposition collocations (such as *by mail*, *via satellite*, *out of love*, etc.).

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, Latin abbreviations are used. Consider, for illustration, eight of the most common standard verb-noun LFs in Table 3. 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 subscript to the LF-name specifies how the argument of the LF and the semantic actant structure of the LF are projected onto the syntactic structure of the verbal value of the LF. For instance, in connection with 'Oper' and 'Real', '$_1$' means that the first actant of the argument (the Doer, the Agent, etc.) is the verb's Subject and the argument itself its Object, and '$_2$' means that the second actant of the argument (the Undergoer, the Patient, etc.) is the verb's Subject and the argument its Object. In connection with 'Func', '$_1$' indicates that the argument is the Subject of the verbal value and the first actant of the argument its Object, while '$_2$' means that the argument is the verb's Subject and the second actant of the argument its Object. '$_0$' in Func$_0$ signals that the argument of the LF is the Subject of an intransitive verbal value. In the complex LF Liqu$_1$Func$_0$, '$_1$' stands for the realization of the first actant of the argument as Subject, and '$_0$' for the realization of the argument of

[8] Again, for better readability, we introduce the preposition and the article of the noun.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 3. *Eight of the most common standard verb-noun* LFs

1.	'perform', 'do', 'act' ($Oper_1$)		5.	'concern', 'apply to' ($Func_2$)	
	BLOW	*deal*		ANALYSIS	*concern*
	OBSTACLE	*pose*		BLOW	*fall* [*upon*]
	RESISTANCE	*put up*		CHANGE	*affect*
	SUPPORT	*lend*		LECTURE	*be* [*on*]
2.	'undergo', 'meet' ($Oper_2$)		6.	'act accordingly' ($Real_1$)	
	BLOW	*receive*		ACCUSATION	*prove*
	OBSTACLE	*encounter*		PROMISE	*keep*
	RESISTANCE	*meet*, *run* [*into*]		SCHEDULE	*stick* [*to*]
	SUPPORT	*receive*		THREAT	*fulfil*
3.	'happen', 'take place' ($Func_0$)		7.	'react accordingly' ($Real_2$)	
	ACCIDENT	*happen*		DEMAND	*fulfil*, *meet*
	RAIN	*fall*		HINT	*take*
	RUMOUR	*circulate*		LAW	*abide* [*to*]
	SMELL	*linger*		CALL	*answer*
4.	'originate from' ($Func_1$)		8.	'put an end to' ($Liqu_1Func_0$)	
	ANALYSIS	*be due* [*to*]		SUPPORT	*withdraw*
	BLOW	*come* [*from*]		RESISTANCE	*put down*
	PROPOSAL	*stem* [*from*]		OBSTACLE	*remove*
	SUPPORT	*come* [*from*]		MEETING	*end*

the LF as Object. Since we interpret complete LF-names as collocation class labels, we can ignore the semantics of the subscripts and consider them simply as parts of LF-names.

An automatic classification of verb-noun combinations extracted from the corpus according to LFs would obviously be extremely helpful for both (computational) lexicography and NLP-applications.

3 The theoretical model of classification

In what follows, we outline the general theoretical aspects of our model for the classification of verb-noun combinations according to the typology of LFs. First, we present how LF-specific relations are learned from collocations in the training sets. The collocations in the training sets are LF-instances gathered manually to train the system. Then, we show how candidate verb-noun combinations are classified according to LF-typology drawing on the relations learned.

3.1 Learning LF-characteristic relations

We presuppose a componential description of the meaning of the bases and collocates of the LF-instances in the training sets. The description of the meaning of a base **B** is thus given by the set of components $\{b_1, b_2, \ldots, b_{N_b}\}$ (with N_b being the number of the components in the description of **B**); and the description of the meaning of a collocate **C** is given by the set of components $\{c_1, c_2, \ldots, c_{N_c}\}$ (with N_c being

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

the number of the components in the description of **C**). The componential meaning description allows us to compute the dependency between the base and the collocate of a collocation as a count over meaning components (or *concepts*) rather than over words, as is most often the case in approaches to automatic recognition of collocations.[9] More precisely, the componential meaning description allows us to calculate the relative pairwise co-occurrence frequency of the components in the base and in the collocate descriptions for a given training set. Or, in other words, to estimate how reliably we can conclude from the occurrence of a component b_i ($1 \leq i \leq N_b$) in the description of a base that the component c_j ($1 \leq j \leq N_c$) will occur in the description of the corresponding collocate. We call the relative pairwise co-occurrence frequency of b_i and c_j the *association strength* of the pair $\langle b_i, c_j \rangle$.[10] For the calculation of the association strength of a given $\langle b_i, c_j \rangle$ we adopt the formula from Ogawa *et al.* (1991).

Let $M = B \times C$ be the relative co-occurrence matrix, where B is the set of the components of all bases in the training set for a given LF, and C the set of the components of all collocates in the same training set. Then, the association strength m_{ij} of a pair $\langle b_i, c_j \rangle$, $b_i \in B$, $c_j \in C$ ($1 \leq i \leq |B|$, $1 \leq j \leq |C|$) is given by:

$$m_{ij} = \frac{a_{ij}}{a_i + a_j - a_{ij}} \tag{1}$$

where a_{ij} is the number of the common occurrences of b_i and c_j, a_i the number of the occurrences of b_i, and a_j the number of the occurrences of c_j. $|B|$ stands for the total number of different base components in the training set, and $|C|$ stands for the total number of different collocate components in the training set.

Equation (1) is closely related to the *pointwise mutual information* formula often used in corpus-based NLP. (1) performs somewhat better on sparse data than *mutual information* – see Manning and Schütze (1999) and the references therein for a discussion on the problems encountered with *mutual information* in the case of data sparseness. Furthermore, unlike *mutual information*, *log-likelihood ratio* and other similar measures, (1) does not presuppose that the two tokens w_1 and w_2 for which dependency is computed are members of the same sequence of tokens (e.g. as words in the corpus) such that $N_{w_2} \leq N - N_{w_1}$ (with N being the total number of tokens in the sequence, N_{w_1} being the number of w_1 tokens and N_{w_2} the number of w_2 tokens).

The ratio between a given m_{ij} and $\sum_{r=1,r\neq i}^{|B|} \sum_{s=1,s\neq j}^{|C|} m_{r,s}$, i.e. the sum of all other m_{ij}s of the training set (with $|B|$ being, as above, the total number of different base components in the training set, and $|C|$ being the total number of different collocate components in the training set), gives us a hint about the importance (the *weight*) of the corresponding $\langle b_i, c_j \rangle$ for the description of the LF-instances in the training set in question, and, therefore, for the LF we want to identify. The weight of any $\langle b_i, c_j \rangle$

[9] Since, as mentioned in Section 1, we use semantically disambiguated data, the computing is rather straightforward.

[10] Henceforth, component pairs will be written in angle brackets '$\langle\rangle$', component pairs of collocations will be referred to as '$\langle b_i, c_j \rangle$', component pairs of candidate noun-verb bigrams as '$\langle n_i, v_j \rangle$'. '(n, v)' will be used to refer to candidate noun-verb bigrams.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

in the description of a training set is given by:

$$E_{\langle b_i,c_j\rangle} = \frac{\tau(b_i) \times \tau(c_j) \times m_{ij}}{\sum_{r=1,r\neq i}^{|B|} \sum_{s=1,s\neq j}^{|C|} m_{rs}} \tag{2}$$

with $\tau(x)$ as the component type significance function. $\tau(x)$ takes into account that in a componential description of a lexical item, not all types of components may equally contribute to the lexico-semantic discrimination of this item. If only one type of component is distinguished, τ is a constant.

Given the weight of each $\langle b_i, c_j\rangle$ in the training set at hand, we can calculate the total weight, or the score, W of an individual LF-instance $\mathscr{C}_r$ in the training set. Symplifying, this is the sum of the weights of all $\langle b_i, c_j\rangle$s in the componential description of this LF-instance. However, a simple sum would lead to an unbalanced picture since LF-instances with fewer $\langle b_i, c_j\rangle$-pairs in their description would be unproportionally penalized. To avoid this distortion, the sum is normalized, i.e. divided by the number of $\langle b_i, c_j\rangle$-pairs that the description of the LF-instance $\mathscr{C}_r$ in question possesses:

$$W(\mathscr{C}_r) = \frac{\sum_{i=1}^{N_b} \sum_{j=1}^{N_c} E_{\langle b_i,c_j\rangle_{\mathscr{C}_r}}}{N_b \times N_c} \tag{3}$$

(with N_b being the number of the components in the description of the base, and N_c being the number of the components in the description of the collocate).

The average of the scores of all LF-instances in the training set provides the score of the prototypical collocation (the *centroid*) captured by the LF in question:

$$\overline{W}(LF) = \frac{\sum_{r=1}^{p} W(\mathscr{C}_r)}{p} \tag{4}$$

(where p stands for the number of the LF-instances in the training set).

The difference between the score W of an LF-instance and the centroid $\overline{W}(LF)$ of an LF thus signals how typical this LF-instance is for the LF under consideration.

3.2 The basic classification procedure

Once the centroid for each LF used in the classification task is determined, disambiguated verb-noun bigrams can be classified, i.e. assigned to one of the LFs or rejected. As mentioned in Section 1, the candidate verb-noun bigrams can either be retrieved and disambiguated during the preceding stages in the overall process of the automatic acquisition of collocations from the corpus, or taken from existing lists of verb-noun bigrams.

The classification procedure consists of the calculation of the score of a candidate bigram (n, v) relative to each centroid. If the relative deviance between the candidate's score and the centroid in question is greater than a specific Δ, the candidate is rejected as an instance of the corresponding LF, otherwise, it is considered an instance of it. Δ is an empirical figure (see also Section 6). That is, our approach can be considered an implementation of a specific version of the *instance-based* machine learning technique: Each class (= LF) is represented in terms of a prototype (*centroid*). For any new candidate, we calculate the distance between the candidate and the

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

prototypes. The candidate is then assigned to all classes whose prototypes are closer to the candidate than a predefined threshold. Obviously, this means that a bigram can be "recognized" as an instance of several LFs. In this case, the result of the classification is a disjunction of LFs rather than a single LF.

The componential descriptions of the noun n and the verb v of the candidate bigram (n, v) serve as input to the classification procedure. Let the componential description of the noun be $\{n_1, n_2, \ldots, n_{N_N}\}$ (N_N = the number of components in the description of n) and the componential description of the verb $\{v_1, v_2, \ldots, v_{N_V}\}$ (N_V = the number of components in the description of v). In classifying each component pair $\langle n_i, v_j \rangle$ (with $1 \leq i \leq N_N$ and $1 \leq j \leq N_V$) with respect to a specific LF, the following parameters are taken into account:

- the weight of $\langle n_i, v_j \rangle$ [if $\langle n_i, v_j \rangle \in M$, i.e. $\langle n_i, v_j \rangle$ is a $\langle b_k, c_l \rangle$-pair of the training set],
- the maximal similarity of $\langle n_i, v_j \rangle$ to a $\langle b_k, c_l \rangle \in M$ and the weight of the latter [if $\langle n_i, v_j \rangle \notin M$],
- a penalty score [if $\langle n_i, v_j \rangle \notin M$ and there is no similar $\langle b_k, c_l \rangle \in M$, or if $\langle n_i, v_j \rangle \in M^-$].

($M = B \times C$ is the relative co-occurrence matrix as defined in Section 3.1, $1 \leq k \leq |B|$, $1 \leq l \leq |C|$, and $M^- = B^- \times C^-$ is the relative co-occurrence matrix of a *negative* training set, i.e. a set which contains some verb-noun combinations that are clearly not instances of the LF to which the candidates are assigned.)

How the similarity between the descriptions of two lexical items is estimated depends on the nature of the decompositional description of the lexical items involved. In our case, it is of a hierarchical nature. We assume that the description of each lexical item consists of a set of its hyperonyms (i.e. of a hyperonymy hierarchy). Each hyperonym in the hierarchy can be further characterized by one or several semantic features.

As any taxonomy, a hyperonymy hierarchy is an abstraction hierarchy. In our model, abstraction is essential for both learning and classification. In the learning stage, we abstract from the individual features of the LF-instances in a training set in order to identify the features common to all of them. In the classification stage, we abstract from the individual features of a candidate combination in order to assess whether its generalized description is sufficiently similar to the descriptions of the LF-instances in the training set in question. That is, the hyperonymy hierarchy forms the backbone of the description of lexical items.

As Leacock and Chodorow (1998), we calculate the similarity between the descriptions of two lexical items α and β using path length, or, in other words, the degree to which the hyperonym hierarchies of α and β overlap.

The formula used to identify the maximal similarity between the description of α and the description of β is:

$$S_{max}(\alpha, \beta) = arg\, \max_{\lambda} \left(\frac{\max_{\beta \in H(\lambda)} (D(H_\alpha \cap H_\beta))}{D(H_\alpha \cap H_\lambda)} \right) \tag{5}$$

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

with $\langle \beta, x \rangle \in M$ if β is a base hyperonym and $\langle x, \beta \rangle \in M$ if β is a collocate hyperonym, D being the depth or path length function, and H_α, H_β, H_λ the hyperonym hierarchy of α, β and λ, respectively.

$H_\alpha \cap H_\lambda$ provides the intersection between the hyperonym hierarchies of α and of a $\lambda \in B \cup C$ that is chosen in such a way that the intersection is maximal.[11] $H_\alpha \cap H_\beta$ (with $\beta \in H(\lambda)$) provides the maximal intersection between the hyperonym hierarchies of α and of a β such that β possesses a pair complement in M.

The weight of $\langle \beta, x \rangle$ (if α is a noun component), or the weight of $\langle x, \beta \rangle$ (if α is a verb component), is then used to estimate the contribution of $\langle \alpha, v_j \rangle$ (respectively $\langle n_i, \alpha \rangle$) to the score of the candidate bigram. Given a $\langle n_i, v_j \rangle$ where n_i is a hyperonym and $\langle n_i, v_j \rangle \notin M$, the weight of $\langle n_i, v_j \rangle$ is calculated using (5) as follows:

$$W_{sim}^{n_i}(\langle n_i, v_j \rangle) = arg \max_{b_k} \left(\frac{\max_{b_{k'} \in H(b_k)} (D(H_{n_i} \cap H_{b_{k'}}))}{D(H_{n_i} \cap H_{b_k})} \right) E(\langle b_{k'}, x \rangle) \tag{6}$$

with $b_k, b_{k'} \in B$ and x as defined above.

If v_j is a hyperonym and n_i is not, the same formula with appropriate parametrization (n_i is replaced by v_j, b_k by c_k, and $b_{k'}$ by $c_{k'}$, $c_k, c_{k'} \in C$) is applied. If both n_i and v_j are hyperonyms, W_{sim} is calculated for both and the minimum W_{sim} is taken. This formula also captures the case where one of the elements in $\langle n_i, v_j \rangle$ appears in the corresponding component list of the training set and the other element does not. In this case, $b_{k'}$ (respectively $c_{k'}$) is equal to n_i (respectively v_j).

If $\langle n_i, v_j \rangle \notin M$, n_i and/or v_j are hyperonyms but have no similar hyperonyms in the compontential description of the training set, or neither n_i nor v_j is a hyperonym, the function Pe calculates the penalty by which the score of the candidate bigram is to be decremented:

$$Pe(\langle n_i, v_j \rangle) = \rho(\langle n_i, v_j \rangle) Mean \tag{7}$$

where $Mean$ is the average weight of all $\langle b_i, c_j \rangle$s in M, and ρ a component type significance function (similar to τ in Formula 2). When $\langle n_i, v_j \rangle \notin M$ and $\langle n_i, v_j \rangle \in M^-$ (recall that M^- is the negative training set co-occurrence matrix) its weight is also subtracted from the score of the candidate.

The algorithm for computing the score of a candidate bigram with respect to the training set of an LF is sketched in Figure 2.

Note that the weight of a $\langle n_i, v_j \rangle \in M$ is added to the score of the candidate bigram even if $\langle n_i, v_j \rangle$ also occurs in M^-. This is done in order not to distort the characteristics of the LF in question. However, this practice may lead to too much noise in the case of "universal" $\langle n_i, v_j \rangle$s that appear in descriptions of the instances of many LFs. In such a case, it may be advisable not to take a $\langle n_i, v_j \rangle$ into account or even to substract its weight from the score of the candidate.

The algorithm is iterative. After the first run of the classification procedure, bigrams that have been recognized as collocations of the LF in question are added

[11] In other words, λ is the hyperonym in the description of the collocations in the training set whose hierarchy overlaps with that of α more than the hierarchies of all other hyperonyms.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

$W_c := 0$; // initialize the score of the candidate bigram
for each $\langle n_i, v_j \rangle$ of the candidate bigram **do**
  $w_s := 0$ // initialize the weight variable
  if $\langle n_i, v_j \rangle \in M$
    then $w_s := E(\langle n_i, v_j \rangle)$
    elseif $\langle n_i, v_j \rangle \in M^-$
      then $w_s := -E^-(\langle n_i, v_j \rangle)$
      elseif is-hyperonym(n_i) $\wedge$ is-hyperonym(v_j)
        then $w_s := min(W_{sim}^{n_i}(\langle n_i, v_j \rangle), W_{sim}^{v_j}(\langle n_i, v_j \rangle))$
        elseif is-hyperonym(n_i)
          then $w_s := W_{sim}^{n_i}(\langle n_i, v_j \langle)$
          elseif is-hyperonym(v_j)
            then $w_s := W_{sim}^{v_j}(\langle n_i, v_j \rangle)$
          endif
        endif
      endif
  endif
  if $w_s = 0$
    then $W_c := W_c - Pe(\langle n_i, v_j \rangle)$ //penalize the disparity
    else $W_c := W_c + w_s$
  endif
$W_c := W_c/(N_N \times N_V)$ //normalize the score

Fig. 2. Algorithm for the calculation of the score of a candidate bigram with respect to the training set of an LF.

to the training set, the centroid is recalculated and the algorithm is again performed on the remaining candidate set. The algorithm stops when no new candidates are identified as collocations.

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 hyperonymy hierarchies provided by the Spanish part of the lexical database EuroWordNet (Vossen 1998; 1999). SpanWN is a middle-size lexical database with 23,370 *synsets*, i.e. sets of synonymous lexical items, and with 50,526 word meanings. 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). In contrast to the Princeton WN (Miller, 1990; Beckwith *et al*., 1991; Fellbaum, 1998), where the hyperonymy hierarchy of a lexical item is purely lexical (i.e. contains only hyperonyms), in SpanWN (as in most WNs in the EuroWN), 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.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

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.[12] 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.[13] 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`.

It is the union of all three types of components (hyperonyms, BCs and TCs) in the semantic description of a lexical item that makes this description sufficiently fine-grained and discriminatory to be usable in the process of classification. Consider, for illustration, Figure 3, which shows the hyperonym hierarchies (including synonyms, BCs and TCs) of PRESTAR**1** 'lend' and DECLARACIÓN**3** 'declaration' from the collocation *prestar declaración* lit. 'lend declaration'.

In SpanWN, four different senses of *declaración* are distinguished: (1) 'declaration in written form'; (2) 'declaration as statement of argumentation', (3) 'declaration as official statement', and (4) 'declaration as unsworn statement in a legal transaction'. In *prestar declaración*, it is the third sense that applies. In SpanWN, *prestar* has one sense only: 'lend'.[14]

The BC of DECLARACIÓN**3** is `communication`, which carries the TC-features `Agentive`, `Social`, `Purpose`, `Manner`, `Communication` and `Bounded Event`. For DECLARACIÓN**3**, the parent hyperonym is the synset whose lexical units all roughly mean 'statement' and that consists of AFIRMACIÓN**3**, EXPOSICIÓN**8**, INFORME**1**, MANIFESTACIÓN**2**, and RELACIÓN**5**. The BC of this synset is again `communication` with specific TC-features. The root of the hyperonym hierarchy of DECLARACIÓN**3** is ABSTRACCIÓN**1** 'abstraction'.

PRESTAR**1** possesses one synonym – DEJAR**14**. The BC of the corresponding synset is `possession`, which does not display any TC-features. The immediate hyperonym of both PRESTAR**1** and DEJAR**14** is DAR**8** 'give', which in turn possesses the hyperonym TRANSFERIR**3** 'transfer'.

5 Setting up the experiments

To validate the proposed model, we conducted two experiments. In the first experiment, we classified candidate verb-noun bigrams whose nouns belong to the same semantic field. The semantics of the nouns that belong to the same semantic field are *a priori* homogeneous at a certain level of abstraction. The lexical-semantic

[12] The numbers indicate the corresponding senses in SpanWN.
[13] Note, however, that unique beginners in Princeton WN are part of the BC-set.
[14] Clearly, the sense 'lend' of *prestar* does not cover its reading as support verb in *prestar declaración*.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

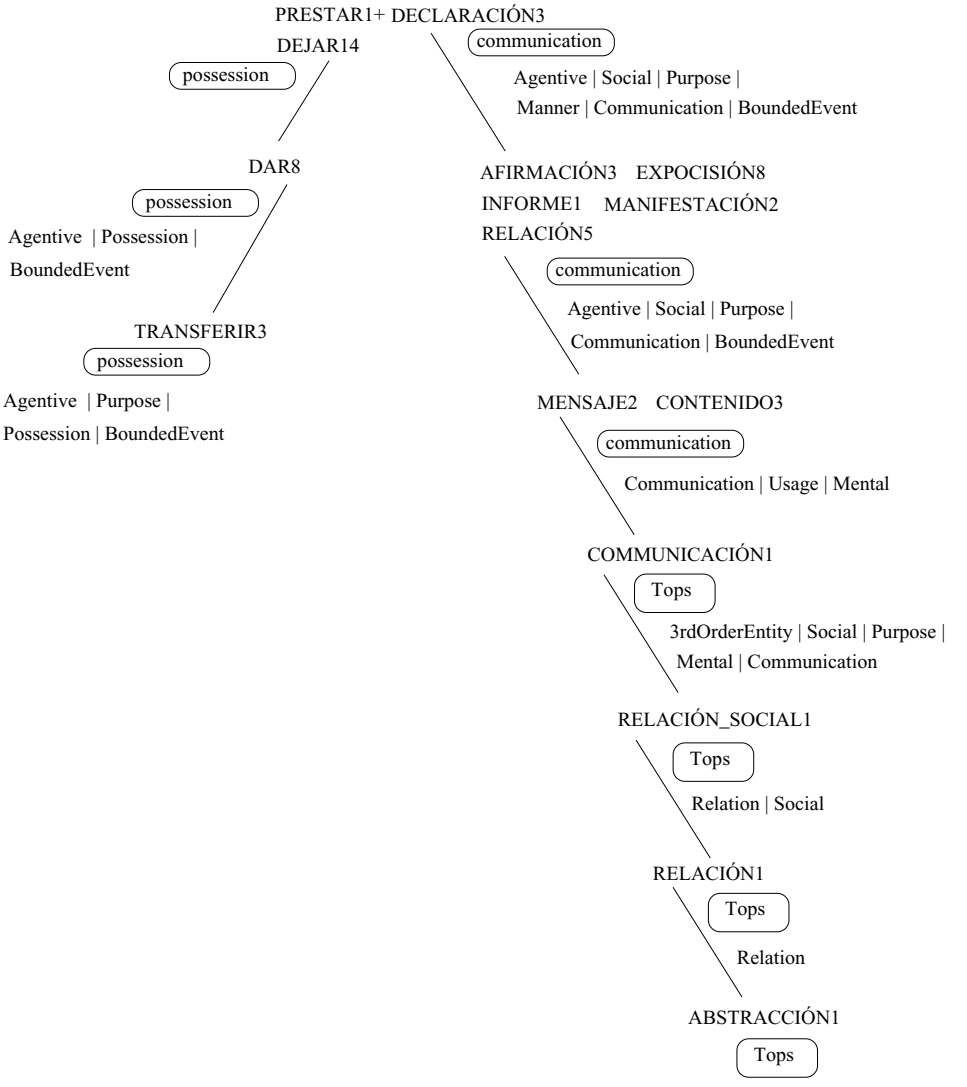


Fig. 3. Hyperonym hierarchies for PRESTAR**1** and DECLARACIÓN**3** in the collocation *prestar* [*una*/*la*] *declaración* (lexical items are written in small capitals, BCS are boxed, and the TCS start with a capital; individual TCS are separated by the '|' sign).

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. The field we have chosen in order to verify this expectation was the field of emotion nouns, which includes, among others, the following nominals:

admiración 'admiration', *alegría* 'joy', *aprensión* 'apprehension', *desesperación* 'despair', *enfado* 'annoyance', *entusiasmo* 'enthusiasm', *esperanza* 'hope', *indignación* 'indignation', *ira* 'rage', *miedo* 'fear', *odio* 'hatred', *pánico* 'panic'.

We have chosen the field of emotion nouns for two reasons: (i) emotion nouns are rich in collocations, (ii) for emotion nouns, lists of LF-instances are already available

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

for French (Mel'čuk *et al*., 1984, 1988, 1992, 1999), German (Mel'čuk and Wanner, 1996), and, what is more important, for Spanish (Sanromán, 1998) – which facilitates the compilation of the training sets.

In the second experiment, we classified verb-nouns bigrams with no consideration of field constraints.

In this section, we first sketch the steps that led to the classification of candidate verb-noun bigrams according to LF-typology. Next, we introduce the LFs, and training and test data used in the experiments.

5.1 Steps in collocation classification of verb-noun bigrams

The classification procedure of candidate bigrams according to LF-typology based on the model introduced in Section 3 consists of two stages: the learning stage **L** and the candidate classification stage **C**. In stage **L**, the system learns the relation between the bases and the collocates of LF-instances in the training sets. The steps in **L** can be summarized as follows:

- **L.1** Compilation of a training set of instances for each LF used for classification.
- **L.2** Acquisition of the componential meaning description of the base and the collocate of each LF-instance in the training sets.
- **L.3** Calculation of association strengths between base and collocate components, determination of the score for each LF-instance, and derivation of the centroid for each LF from the scores for LF-instances in the corresponding training set as outlined in Section 3.1.

In stage **C**, lists of candidate verb-noun bigrams are processed and classified as follows:

- **C.1** Acquisition of the componential description of the noun n and the verb v in each verb-noun bigram (n, v).
- **C.2** Calculation of the score W_c of each (n, v) as outlined in Section 3.2.
- **C.3** Comparison of W_c with the centroids $\overline{W}(LF_i)$ (i = 1,2,…, n). If $W_c/\overline{W}(LF_i) - 1 < \Delta$ (with Δ being empirically determined) (n, v), is classified as a collocation of LF_i.

The classification stage (more precisely, steps **C.2** and **C.3**) is performed for each LF on all candidate bigrams (recall that a candidate bigram can be classified as an instance of several LFs). As mentioned in Section 3.2, the classification procedure is iterative. In the course of assigning candidate bigrams to a given LF-class, the centroid is recalculated after each iteration because bigrams that have been identified during an iteration as instances of this LF are added to the corresponding training set. That is, in the worst case, the centroid is calculated for each LF $|Cand|$-times (with $|Cand|$ being the number of candidate bigrams).

5.2 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 collocations that our model is able to recognize, and, on the other hand, the potential of the model to distinguish between

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

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. However, 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 our approach performs when applied to the stock of LFs being used by lexicographers in the Explanatory Combinatorial Lexicology.

The following five LFs were considered in Experiment 1: $Oper_1$, $ContOper_1$, $Caus_2Func_1$, $IncepFunc_1$ and $FinFunc_0$. Note the glosses and examples for each:

$Oper_1$ 'experience an emotion'; e.g.:
sentir admiración lit. 'feel admiration'
tener alegría lit. 'have joy'
experimentar odio lit. 'experience hatred'
tener odio lit. 'have hatred'

$ContOper_1$ 'continue to experience an emotion'; e.g.:
guardar entusiasmo lit. 'keep enthusiasm'
guardar esperanza lit. 'keep hope'
conservar odio lit. 'conserve hatred'

$Caus_2Func_1$ 'cause (by the object of emotion) the emotion to be experienced'; e.g.:
causar horror lit. 'cause horror'
dar sorpresa lit. 'give surprise'
provocar indignación lit. 'provoke indignation'
despertar odio lit. 'awake hatred'

$IncepFunc_1$ 'an emotion begins to be experienced'; e.g.:
desesperación entra [*en* N] lit. 'despair enters [in N]'
odio se apodera [*de* N] lit. 'hatred gets hold [of N]'
ira invade [N] lit. 'rage invades [N]'

$FinFunc_0$ 'an emotion ceases to be experienced'; e.g.:
aprensión se disipa lit. 'apprehension evaporates'
odio desaparece lit. 'hatred disappears'
entusiasmo se desvanece lit. '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

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

of the verb, and the noun itself its Object. The government pattern of $Caus_2Func_1$ 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. The semantics of $Caus_2Func_1$ 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, Senser 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 $Caus_2Func_1$ than to those of $FinFunc_0$.

As in Experiment 1, in Experiment 2, the classification procedure was 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 [*la*] *alarma* lit. 'give alarm'
celebrar elecciones lit. 'celebrate elections'
provocar [*una*] *crisis* lit. 'provoke crisis'
publicar [*una*] *revista* lit. 'publish review/journal'

$Oper_1$ 'perform', 'experience', 'carry out', etc.; e.g.:
dar golpe lit. 'give blow'
presentar [*una*] *demanda* lit. 'present demand'
hacer campaña lit. 'do campaign'
dictar [*la*] *sentencia* lit. 'dictate sentence'

$Oper_2$ 'undergo', 'be source of', etc.; e.g.:
someterse [*a un*] *análisis* lit. 'submit (oneself to an) analysis'
afrontar [*un*] *desafío* lit. 'face challenge'
hacer examen lit. 'do examination'
tener [*la*] *culpa* lit. 'have blame'

$Real_1$ 'act accordingly to the situation', 'use as foreseen'; e.g.:
ejercer [*la*] *autoridad* lit. 'exercise authority'
utilizar [*el*] *teléfono* lit. 'use telephone'
hablar [*la*] *lengua* lit. 'speak language'
cumplir [*la*] *promesa* lit. 'fulfil promise'

$Real_2$ 'react accordingly to the situation'; e.g.:
responder [*a*] *objeción* lit. 'respond to objection'
satisfacer [*un*] *requisito* lit. 'satisfy requirement'
atender [*la*] *solicitud* lit. 'attend petition'
rendirse [*a*] *persuasión* lit. 'render (oneself) [to] 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 $Oper_1$ that could be expected to be recognized by the program. Consider, for instance, *rendir homenaje* lit. 'render

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

homage', *dar explicación* lit. 'give explanation', *hacer comentario* lit. 'do comment' and *poner queja* lit. 'put complaint', which have been classified as $Oper_1$-instances by human experts. However, to a certain extent, they also express a 'causation of existence'. Therefore, in such cases, we rated the classification as correct if a candidate bigram that was *de facto* an $Oper_1$-instance was suggested to be an instance of $CausFunc_0$, and if, *vice versa*, $CausFunc_0$-instances were classified as $Oper_1$-instances. As outlined in Section 2, the government patterns of $Oper_1$ and $Real_1$ and of $Oper_2$ and $Real_2$ are the same. The government pattern of $CausFunc_0$ realizes the noun as Object of the verb; the Subject of the verb is an external participant (i.e. a participant that is not an actant of the noun).

5.3 Compiling the training and test sets

A training set must contain a sufficient number of representative instances for the LF in question. To ensure that the members of the training sets are indeed instances of the respective LFs and that the training sets are well-balanced, the training sets must be compiled either manually by an LF-expert or by using an on-line LF-dictionary. So far, only one on-line LF-dictionary of a considerable size is available: the DiCo for French (Polguère, 2000). For Spanish, an LF-dictionary (*Diccionario de colocaciónes del español*, DiCE) is under construction (Alonso Ramos, 2002). However, DiCE does not contain as yet enough material to serve as source for training sets in our experiments. We were thus forced to use other manually compiled resources.

With respect to the size of the training sets it should be noted that:

- The actual size of a given training set depends on the semantic heterogeneity of the collocation types to be retrieved. 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 emotion nouns, only two values are available: *conservar* lit. 'conserve' and *guardar* lit. 'keep'. As we will see below, *conservar* and *guardar* possess the same semantic description. Therefore, 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.
- Since, according to our definition, the phenomenon of collocation is not related to the frequency of the occurrence of words in combination, we must assume that the material for specific LFs will be scarce in certain corpora. Therefore, experiments with small training and test sets are necessary to explore the minimal size of the training and test sets. If the training sets turn out to be too small, *n*-fold cross-validation may be used. Unfortunately, we are not in the position here to vary the sizes of the training sets to observe the change in assessment figures.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 4. *Sizes of the sets used in Experiment 1*

LF	#training set (positive tr. set + negative tr. set)	#test set (posit. samples + negat. samples)
$Caus_2Func_1$	(27 + 42)	(36 + 72)
$ContOper_1$	(5 + 64)	(9 + 99)
$FinFunc_0$	(14 + 55)	(26 + 82)
$Oper_1$	(16 + 53)	(21 + 87)
$IncepFunc_1$	(7 + 62)	(16 + 92)

For Experiment 1, we used a collection of Spanish collocations already classified in terms of LFs (Sanromán, 1998). We divided Sanromán's sets of LF-instances into training subsets and positive candidate subsets, i.e. samples of collocations used to evaluate the performance of the classification procedure with respect to the recognition of positive candidate bigrams as instances of the corresponding LF. The negative training set for a given LF was composed of the (positive) training subsets of the other four LFs. To evaluate the performance of the procedure with respect to the rejection of negative candidate bigrams for a given LF we used the (positive) candidate test subsets of the other four LFs. The distribution figures were as listed in Table 4.

To compile the training sets for Experiment 2, we drew upon the following sources:

1. informants (native speakers of Spanish), whom we asked to provide us with lists of LF-instances used for classification: 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;
2. the *Collins* bilingual English-Spanish dictionary, where we looked up Spanish equivalents for English collocations known to be instances of specific LFs;
3. the POS-tagged corpus LEXESP (Sebastián, 2000), where we looked up verb-noun combinations for sets of predetermined nouns (choosing those combinations which were instances of one of the LFs used for classification).

The negative training set for each LF was composed of the (positive) test subsets of the other four LFs. The negative test set for each LF was composed of the positive training sets of the other four LFs. Table 5 summarizes the sizes of the training and test sets in Experiment 2.

Note that the compilation of the training sets is entirely separate from the potential corpus processing stage in which candidate verb-noun combinations are retrieved from the corpus in order to be automatically classified in the subsequent stage. Also, as indicated above, for the compilation of the training set elements from the corpus, we did not use partial parsing techniques as done in the corpus processing stage. This is because the search for instances of a given LF starting from a set of predefined nouns allowed us to use our intuition with respect to the availability of this LF for specific kinds of nouns.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 5. *Sizes of the sets used in Experiment 2*

LF	#training set (positive tr. set + negative tr. set)	#test set (posit. samples + negat. samples)
$Caus_2Func_0$	(38 + 60)	(15 + 143)
$Oper_1$	(35 + 60)	(15 + 146)
$Oper_2$	(33 + 60)	(15 + 148)
$Real_1$	(37 + 60)	(15 + 144)
$Real_1$	(38 + 60)	(15 + 143)

5.4 Processing members in the training and test sets

Once the training sets and the test sets have been compiled, the senses of the bases and the collocates of the members are disambiguated and their hyperonymy hierarchies are retrieved from SpanWN.

LF-instances in the training sets must be disambiguated in order to ensure precise centroids and high quality classification. This is best done manually by a human expert. Since the average size of the training sets is less than 40 for field-independent classification and less than 20 for field-restricted classification, this appears to be feasible for all available verb-noun LFs.

Although we disambiguate the test sets, our strategy also allows for the use of non-disambiguated elements of candidate bigrams. In this case, if the noun of a given bigram has in SpanWN s_n senses and the verb s_v senses, we obtain $s_n \times s_v$ logically possible pairs of componential sets as the description of this bigram. Each pair may serve as input to the classification procedure. In accordance with our strategy, the bigram will then be rejected as an instance of the LF in question if the respective pair of componential verb-noun descriptions does not give rise to an acceptable score for the candidate. Since all pairs are taken into account in sequence, there is no qualitative difference between the use of disambiguated and non-disambiguated candidate bigrams. However, evidently, the processing of non-disambiguated candidate bigrams is computationally much more costly. Therefore, in our experiments, we have so far used disambiguated bigrams (experiments with non-disambiguated test sets are currently being carried out). The disambiguation has been performed manually by our informants.

A possible strategy for avoiding the disambiguation stage, is to exploit the SpanWN-internal ranking of senses. In SpanWN (as in the Princeton WN), senses are listed in order of their frequency in terms of usage. Therefore, instead of disambiguating the senses of the elements in the candidate combinations, one could by default take the most common sense of each element.[15]

For the sake of simplicity, we flatten out the hyperonym hierarchies.[16] That is, the description of a verb-noun combination (be it a member of a training set or a

[15] This strategy has been suggested by one of the JNLE reviewers.
[16] Note, however, that for the calculation of the similarity between candidate combinations and collocations in the training sets, we use the hierarchical information.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

member of a test set) is given in terms of two SETS of components. Consider, for illustration, the description of *aprensión se disipa* lit. 'apprehension evaporates' (as in Figure 3, hyperonyms are written in small capitals, TCs with a capital and BCs in small letters):

```
((feeling APRENSION2 TEMOR1 RECELO4 PAVOR1
  feeling ESPANTO1 MIEDO1
  feeling Dynamic Mental Experience EMOCION1
  Dynamic Mental Experience SENTIMIENTO1
     Mental Property rasgo-psicologico1)
 (perception DISIPARSE2 DESVANECERSE4
  change Dynamic Location Existence DESAPARECER1 ESFUMARSE1))
```

To build the relative co-occurrence matrix $M = B \times C$ (see Section 3.1), first, repeated elements in the descriptions of the tokens in the training sets are eliminated. Then, the descriptions of all bases are combined to form set B, while the descriptions of the collocates are combined to form set C.

The result of processing LF-instances in the training set is the matrix E_{LF}, where $E_{\langle b_i,c_j\rangle} \in E_{LF}$ is the weight of the base-collocate component pair $\langle b_i, c_j\rangle$, $b_i \in B, c_j \in C$. See Formula (2) for the calculation of $E_{\langle b_i,c_j\rangle}$.

5.5 *Component significance functions*

In (2) and (7), two different component significance functions are used: τ and ρ, respectively.

As already mentioned, τ takes into account that different types of components in the description of a lexical item may contribute to a different degree to the lexical-semantic discrimination of this item. Recall that our descriptions contain three different types of components: hyperonyms, BCs and TCs. Since in Experiment 1 the decompositional descriptions of the training set instances as well as of the positive candidate bigrams are, to a major extent, homogeneous, the variation of τ had no relevance. τ was thus set at 1.0.[17]

In Experiment 2, both initial training sets and candidate bigrams are perforce semantically quite heterogeneous. This suggests that the different types of components in the descriptions of the lexical items should be given different prominence. In other words, the optimal τ is likely to be different for hyperonyms, BCs and TCs. It is also likely to be different for each LF. After a number of trial runs, in Experiment 2, the τs summarized in Table 6 have been used. Due to the already rather small size of the test sets, we refrained from dividing them into *held out* subsets on which we could tune τ (and also ρ, see below) and "real" test subsets and tuned on test data instead.

[17] With $\tau = 1.0$, the weight $E_{\langle b_i,c_j\rangle}$ of a base-collocate component pair $\langle b_i, c_j\rangle$ in Formula (2) is the relative association strength of $\langle b_i, c_j\rangle$: $E_{\langle b_i,c_j\rangle} = \frac{m_{ij}}{\sum_{r=1,r\neq i}^{|B|} \sum_{s=1,s\neq j}^{|C|} m_{rs}}$. Note that a different constant would simply proportionally reduce, respectively multiply, the weights of all base-collocate pairs in the training set without affecting the quality of the learning and classification procedures.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 6. *The τ-parameter in Experiment 2 ('HP' stands for 'hyperonym')*

	CausFunc_0		Oper_1		Oper_2		Real_1		Real_2	
	$\tau(b_i)$	$\tau(c_j)$	$\tau(b_i)$	$\tau(c_j)$	$\tau(b_i)$	$\tau(c_j)$	$\tau(b_i)$	$\tau(c_j)$	$\tau(b_i)$	$\tau(c_j)$
b_i/c_j is a HP	1.5	7.5	2.0	3.75	1.5	7.5	1.5	7.5	2.0	15.0
b_i/c_j is a BC	5.0	5.5	5.0	7.5	5.0	5.5	5.0	7.5	5.0	7.5
b_i/c_j is a TC	1.0	1.0	1.0	1.5	1.0	1.0	1.0	4.5	1.0	2.5

ρ reflects the impact on the bigram score the occurrence of components in the description of a candidate bigram has when these are not available in the descriptions of the LF-instances in a training set. The impact depends, again, on the type of components involved, and on the similarity of these components to the components available in the descriptions of the LF-instances in the training set.

If for a given candidate bigram component pair $\langle n_i, v_j \rangle$ (with $1 \leq i \leq N_N$, $1 \leq j \leq N_V$, N_N = number of components in the description of the noun, N_V = number of components in the description of the verb) either n_i or v_j is a hyperonym and at least one lexical item is similar to any hyperonym in the description of the instances noted in the respective training set, the impact of this pair on the score is calculated using Formula (6). Otherwise, one of the following five cases comes to bear:

1. both elements of the pair are hyperonyms, but in neither case are the corresponding lexical items similar to any of those in the descriptions of the collocations in the training set;
2. the noun component is a hyperonym, but there is no similar lexical item in the descriptions of the bases in the training set;
3. the verb component is a hyperonym, but there is no similar lexical item in the descriptions of the collocates in the training; set
4. The verb component is a BC which occurs in the descriptions of the collocates in the training set;
5. all other cases.

For each of these cases, the impact on the score is given by a distinct ρ-parameter (ρ_1 to ρ_5). For Experiment 1, $\rho_1 - \rho_5$ were set to 1 – as τ was. For Experiment 2, the following ρ-settings proved to provide good results: $\rho_1 = 15.0$, $\rho_2 = 7.5$, $\rho_3 = 7.5$, $\rho_4 = 2.5$, and $\rho_5 = 7.5$ for all five LFs involved. As mentioned above, the tuning of ρ was done on test data.

6 Experiments

In this section, we present the results of Experiments 1 and 2 along with the data described above and compare these results with a baseline. The baseline is the match of the verb of a candidate bigram with one of the most common collocates of the LF in question. The most common collocates used for the LFs drawn upon in Experiments 1 and 2 are summarized in Table 7.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 7. *Typical verbs in collocations covered by the* LFS *in Experiments 1 and 2*

$CausFunc_0$:	INICIAR 'initiate', CREAR 'create'
$Caus_2Func_1$:	CAUSAR 'cause', ESTIMULAR 'stimulate'
$IncepFunc_1$:	TOMAR 'take', ENTRAR 'enter'
$FinFunc_0$:	CESAR 'cease', FINALIZAR 'end'
$Oper_1$:	HACER 'do', DAR 'give', EXPERIMENTAR 'experience'
$ContOper_1$:	CONSERVAR 'keep', GUARDAR 'keep'
$Oper_2$:	PERCIBIR'perceive', SUFRIR 'undergo'
$Real_1$:	UTILIZAR 'apply, utilize, employ', PONER 'put'
$Real_2$:	RESPONDER 'respond', SEGUIR 'follow'

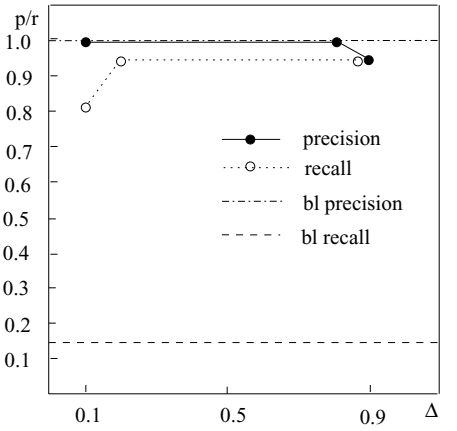


Fig. 4. Precision and recall of $Oper_1$-classification of test set candidate bigrams for various Δs.

In the Appendix, we present, for illustration, the deviance of the scores of some candidate bigrams from the centroids of the LFs in Experiments 1 and 2.

6.1 Experiment 1

Figures 4, 6, 8, 10, and 12 show the dependence of precision and recall on the similarity threshold Δ for the recognition of a candidate bigram as an instance of a given LF. Both figures are contrasted with baseline performance.

Figures 5, 7, 9, 11, and 13 show the distance of the scores of the candidate bigrams from the initial centroid of the respective LF. The precision and recall curves clearly reflect the structure of a cluster of candidate bigrams around the respective centroid.

The $Oper_1$-cluster in Figure 5 shows a very pronounced contrast between candidate bigrams that are grouped around a short distance from the centroid and the rest of the candidate bigrams. The precision and recall curves suggest that all candidates that are very close to the centroid are instances of $Oper_1$. For $\Delta = 0.2$, the recall increases from 80.95% to 95.23%. In other words, candidates at a slightly greater distance from the centroid are also $Oper_1$-instances. With a further increase of Δ, recall and precision remain stable. This suggests that the Δ-interval [0.3, 0.8] covers

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

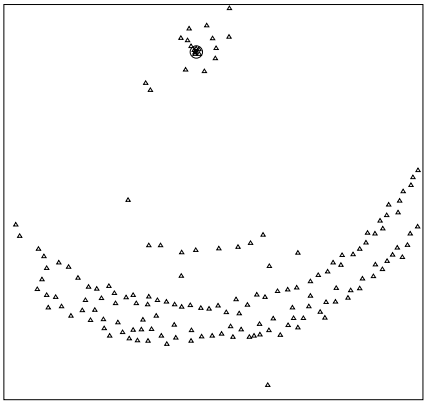


Fig. 5. The distance of the scores of candidate bigrams from the centroid (= 12.7431) of $Oper_1$.

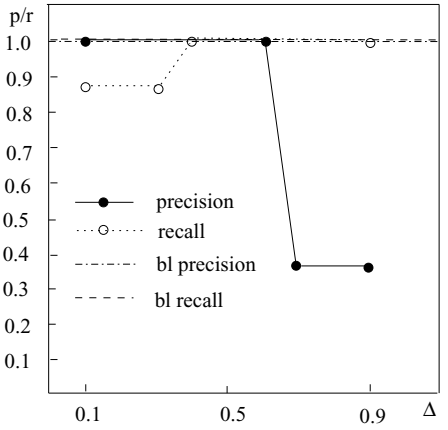


Fig. 6. Precision and recall of $ContOper_1$-classification of test set candidate bigrams for various Δs.

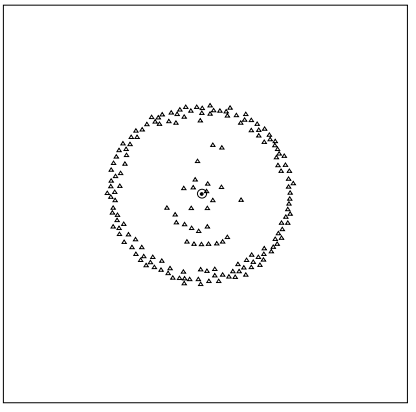


Fig. 7. The distance of the scores of candidate bigrams from the centroid (= 3.4816) of $ContOper_1$.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

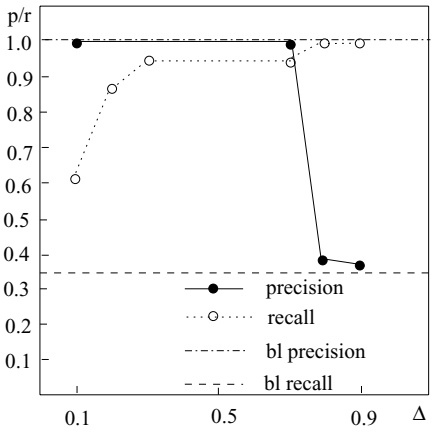


Fig. 8. Precision and recall of $Caus_2Func_1$-classification of test set candidate bigrams for various Δs.

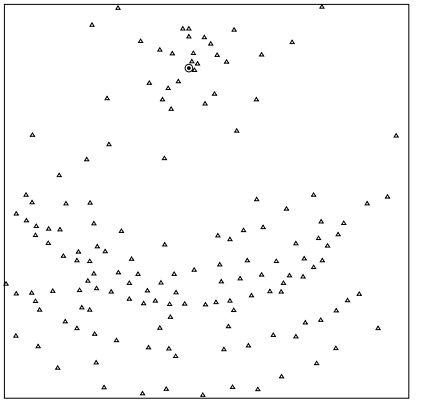


Fig. 9. The distance of the scores of candidate bigrams from the centroid (= 15.1310) of $Caus_2Func_1$.

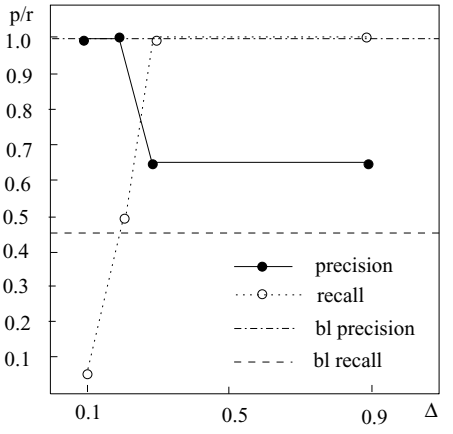


Fig. 10. Precision and recall of $IncepFunc_1$-classification of test set candidate bigrams for various Δs.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

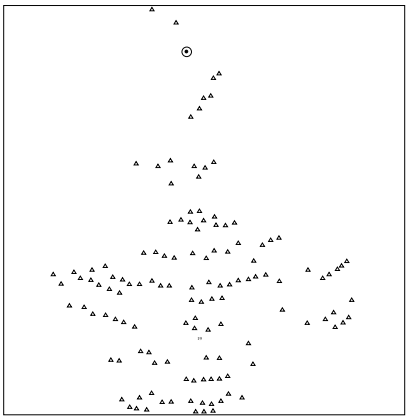


Fig. 11. The distance of the scores of candidate bigrams from the centroid (= 1.3375) of IncepFunc_1.

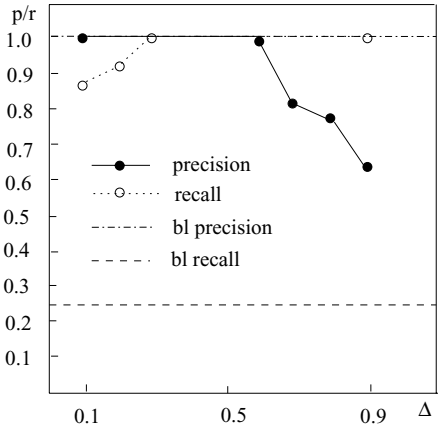


Fig. 12. Precision and recall of FinFunc_0-classification of test set candidate bigrams for various Δs.

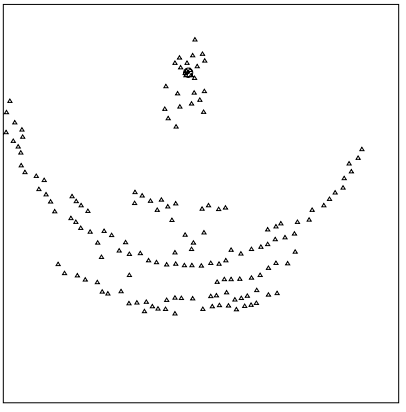


Fig. 13. The distance of the scores of candidate bigrams from the centroid (= 10.8635) of FinFunc_0.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 8. *Maximal f-scores achieved in Experiment 1 vs. baseline f-scores*

	$FinFunc_0$	$IncepFunc_1$	$Caus_2Func_1$	$Oper_1$	$ContOper_1$
Exp 1. *f*-score	100.00	78.04	97.14	97.55	100.00
baseline *f*-score	37.50	60.86	53.06	25.00	100.00

the empty space between the inner "belts" and the more distant "belts" of the candidate bigrams: the difference between the scores of the positive candidates and the scores of the negative candidates is too large to affect the precision figure when Δ is increased. At $\Delta = 0.9$, precision decreases from 100% to 95.23%. Obviously, there are a few negative candidate bigrams whose scores deviate from the centroid by less than 0.9. But there are also a few positive candidate bigrams which are not captured because their scores are vey different from the centroid.

The precision curves of $ContOper_1$, $Caus_2Func_1$, $IncepFunc_1$ and $FinFunc_0$ are similar: up to a specific Δ they remain unchanged at 100%, at which point they fall abruptly. In the case of $ContOper_1$ and $FinFunc_0$, this is at $\Delta = 0.6$; in the case of $Caus_2Func_1$ this is at $\Delta = 0.7$; and in the case of $IncepFunc_1$ at $\Delta = 0.2$. That is, for each of the LFs in question, there are several positive candidates positioned close enough to the centroid with no negative candidate among them. Starting from a certain distance from the centroid, positive and negative candidates with similar scores are identified. This leads to a decrease in precision, and, at the same time, to an increase in recall. The recall increases and reaches 100% at $\Delta = 0.4$ for $ContOper_1$, at $\Delta = 0.8$ for $Caus_2Func_1$, at $\Delta = 0.3$ for $IncepFunc_1$ and for $FinFunc_0$.

Table 8 contrasts the *f*-scores (with precision and recall being equally weighted) achieved in Experiment 1 with baseline *f*-scores. It shows that for field-oriented classification, where one would expect a good baseline, our approach performs considerably better – except for $ContOper_1$. For $ContOper_1$ both the baseline and our approach achieve the optimal *f*-score.

6.2 Experiment 2

Experiment 1 demonstrated that our classification procedure provides high precision and recall rates when the training set instances belong to the same semantic field and the positive candidates are expected to belong to this field as well. Now, let us see how the classification procedure performs on the heterogeneous data in Experiment 2. As in the previous section, the following pairs of Figures show precision and recall in relation to Δ and the distance of candidate bigrams scores from the initial centroid of the respective LF. Again, precision and recall are contrasted with the baseline performance.

The clusters of candidate bigrams around the centroids of $CausFunc_0$ and $Oper_2$ are similar to the clusters plotted for all LFs in Experiment 1: an inner belt of candidates whose scores are close to the respective centroid is positioned around the centroid. Then, there appears an outer belt that consists of candidates whose scores show a great distance from the centroid. This is the majority. Between the inner belt and the outer belt some candidates with intermediate scores are located.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

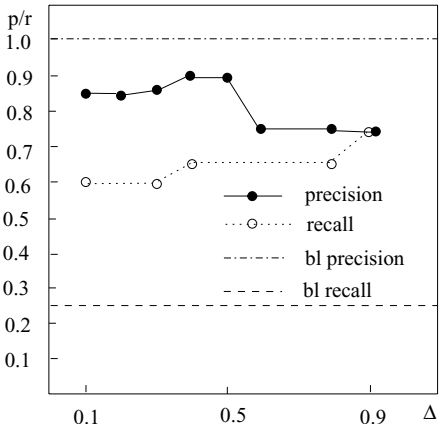


Fig. 14. Precision and recall of CausFunc_0-classification of test set candidate bigrams for various Δs.

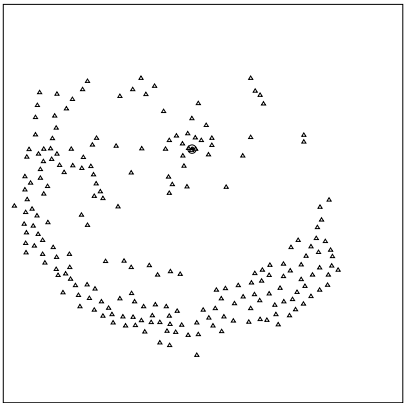


Fig. 15. The distance of the scores of candidate bigrams from the centroid (= 5.4740) of CausFunc_0.

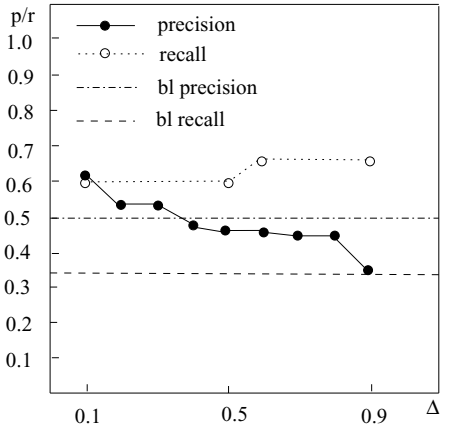


Fig. 16. Precision and recall of Oper_1-classification of test set candidate bigrams for various Δs.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

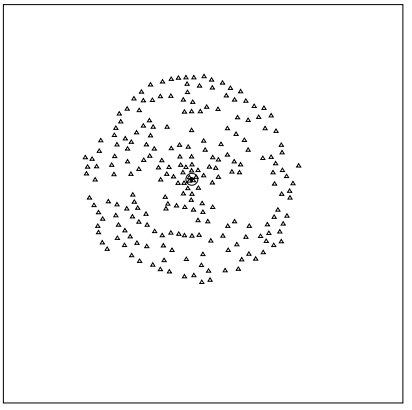


Fig. 17. The distance of the scores of candidate bigrams from the centroid (= 3.0417) of $Oper_1$.

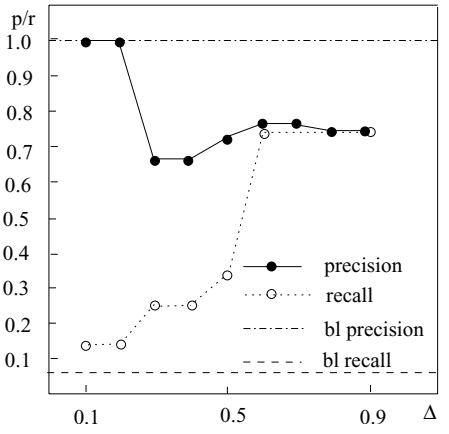


Fig. 18. Precision and recall of $Oper_2$-classification of test set candidate bigrams for various Δs.

However, unlike in Experiment 1, the inner belt of $CausFunc_0$ contains both positive and negative candidate bigrams. Therefore, for $\Delta = 0.1$, a precision of (merely) 87.50% is achieved. This increases to 90% for $\Delta = 0.4$ and then drops (to achieve 73.33% for $\Delta = 0.9$). The recall increases and reaches 73.33% for $\Delta = 0.9$. Again, this shows that in the case of $CausFunc_0$, no clear cut borderline exists between the scores of positive candidates and the scores of negative candidates.

In contrast to $CausFunc_0$, the inner belt of $Oper_2$ does not contain any negative candidate bigrams. However, as the recall figure indicates, the number of candidates in the inner belt is very low. For $\Delta = 0.3$, the precision of $Oper_2$-classification drops to 66.66%, while the recall increases to merely 26.66%. That is, a considerable number of the candidates at a rather short distance from the centroid are not $Oper_2$-instances. As Δ increases, precision increases slightly. The recall increases as well, but much faster than the precision. Thus, here we see another correspondence with the $CausFunc_0$-classification. The precision and recall of $Oper_1$ show a similar, although less pronounced picture.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 9. *Confusion matrix of Experiment 2*

LF drawn upon for classification	# of LF-instances accepted during classification					
	$CausFunc_0$	$Oper_1$	$Oper_2$	$Real_1$	$Real_2$	Δ
$CausFunc_0$	18	(–)	–	2	–	0.4
$Oper_1$	5 (4)	9	–	1	2	0.1
$Oper_2$	–	1	11	–	2	0.6
$Real_1$	–	–	–	10	2	0.2
$Real_2$	–	–	1	1	7	0.4

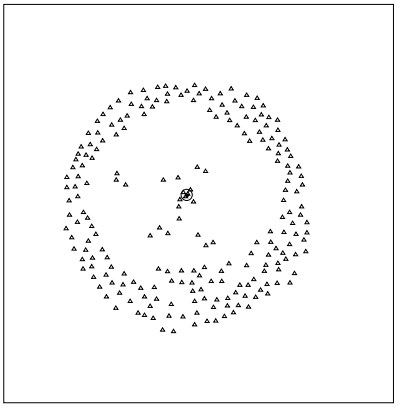


Fig. 19. The distance of the scores of candidate bigrams from the centroid (= 2.3011) of $Oper_2$.

The clusters of $Oper_1$ and especially those of $Real_1$ and $Real_2$ do not show the typical "belt"-like shape of the previous LF-clusters. This suggests that the scores of candidate bigrams with respect to $Oper_1$, $Real_1$, and $Real_2$ are much more heterogeneous. The precision and recall figures support this hypothesis: the majority of the candidates located very close to the centroid ($\Delta = 0.1$) are positive candidates, but there is also a considerable number of negative candidates; compare the precision of 61.90% for $Oper_1$, of 88.88% for $Real_1$ and of 80.00% for $Real_2$. The number of negative candidates increases—as Δ increases.

Unlike in Experiment 1, in Experiment 2 no recall of 100% was reached for any LF. This suggests that among the positive candidates, there were those with a score that drastically deviated from the centroid. Or, in other words, that the decompositional description of certain positive candidates deviated considerably from the descriptions of the LF-instances in the training sets.

Table 9 contains the confusion matrix for each LF at a Δ that resulted in the best *f*-score (with precision and recall being again equally weighted). The first column lists the LFs used for classification. The next five columns indicate how many instances of a given LF were accepted as samples of the LF drawn upon during classification.

'(–)' for $Oper_1$ in the $CausFunc_0$-line means that some $Oper_1$-instances were classified as $CausFunc_0$-instances, but that these can also be admitted as

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

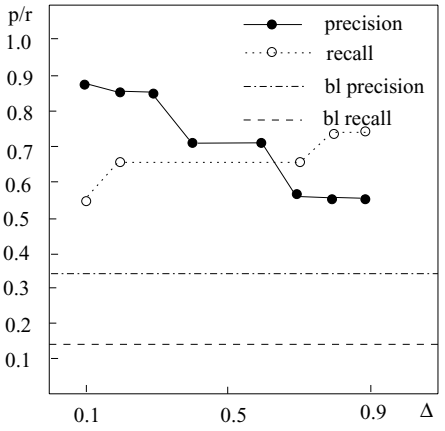


Fig. 20. Precision and recall of $Real_1$-classification of test set candidate bigrams for various Δs.

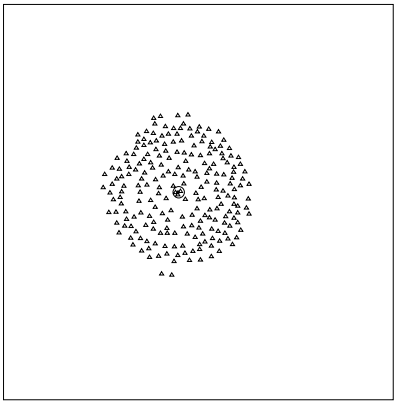


Fig. 21. The distance of the scores of candidate bigrams from the centroid (= 1.6136) of $Real_1$.

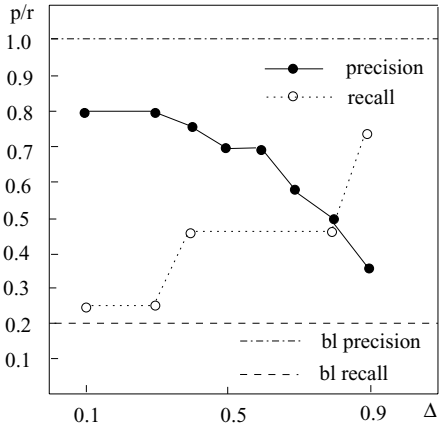


Fig. 22. Precision and recall of $Real_2$-classification of test set candidate bigrams for various Δs.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

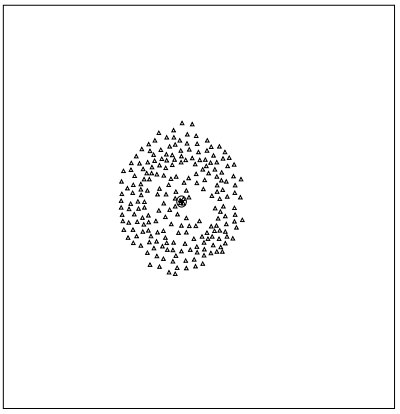


Fig. 23. The distance of the scores of candidate bigrams from the centroid (= 1.2525) of $Real_2$.

Table 10. *Maximal f-scores achieved in Experiment 2 vs. baseline f-scores*

	$CausFunc_0$	$Oper_1$	$Oper_2$	$Real_1$	$Real_2$
Exp 2. *f*-score	76.58	60.93	75.85	74.06	58.32
baseline *f*-score	42.10	40.00	12.50	19.04	33.33

$CausFunc_0$-instances. '(4)' for $Causfunc_0$ in the $Oper_1$-line means that four $CausFunc_0$-instances have been classified as $Oper_1$, but that, again, this has been judged as acceptable. See Section 5.2 on the similarity between these two LFs. $Oper_1$-samples rated as $CausFunc_0$-instances include: *hacer comentario* lit. 'do comment', *dar explicación* lit. 'give explanation', and *hacer favor* lit. 'do favor'. The set of $CausFunc_0$-instances classified as $Oper_1$ contained *celebrar elecciones* lit. 'celebrate elections', *sacar apuntes* lit. 'take out notes', *celebrar entrevista* lit. 'celebrate interview', etc. The classification of the other 5 $CausFunc_0$-instances as $Oper_1$, was considered incorrect.

The confusion matrix shows that most often the noise is caused by $Real_1$- and $Real_2$-instances. This can be explained by the level of detail in the decomposition of the meaning of lexical items in SpanWN. At the level given, the meaning of $Real_1$-collocates is similar to the meaning of the collocates of $CausFunc_0$ and $Oper_1$, while the meaning of $Real_2$-instances is similar to the meaning of the collocates of $Oper_1$ and $Oper_2$. Also, $Real_1$ and $Real_2$ have, by definition, very similar semantic features.

Table 10 compares the *f*-scores achieved in Experiment 2 with baseline *f*-scores. As we can see, for field-independent classification, our approach is also always considerably better than the baseline.

7 Discussion

In this section, we summarize first the validation of the classification procedure in Experiments 1 and 2 and then discuss the general parameters on which the classification procedure depends.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

7.1 Summary of the experiments

In Experiments 1 and 2, we tested the classification procedure on negative and positive candidate bigrams.

The negative candidate bigrams were verb-noun combinations where (i) the noun occured as the base in the training set for the respective LF, but the verb (in the given reading) did not occur as collocate, or (ii) neither the noun nor the verb occured in the training set.

The positive candidate bigrams were verb-noun combinations where (i) the verb did not occur as collocate in the respective training set; (ii) the noun did not occur as base; (iii) neither the noun nor the verb occured as base/collocate, or (iv) both the noun and the verb were available, but not together in one combination.

In Experiment 1, the overwhelming majority of candidate bigrams that fell into one of the categories (i) to (iv) has been classified correctly. This was so because even if the elements of a candidate bigram did not occur in the training set, semantically similar or semantically identical lexical items did. Consider, for instance, *encender* 'inflame' (as in *encender ira* lit. 'inflame rage') and *dar* 'give' (as in *dar sorpresa* lit. 'give surprise'), which were not available in the training set for $Caus_2Func_1$ but occured in the test set. The semantic description of *encender* is given in Figure 24a. It subsumes the description of *despertar* 'arouse' (Figure 24b), which was available in the training set as collocate. That is, *encender* is just like *despertar*, only somewhat more specific.

The reverse situation occurs in the case of *dar*. Its semantic description is more abstract than, for example, the description of *producir* 'produce', which occured in the training set as collocate. Consider the componential descriptions of *dar* and *producir* in Figure 25.

(a) *encender*:

```
(emotion ENCENDER3 INFLAMAR2
 emotion Dynamic Stimulating Mental Experience DESPERTAR5 INCITAR9
         ESTIMULAR7
 creation BoundedEvent Physical Existence Cause CAUSAR3 REALIZAR2 CREAR2)
```

(b) *despertar*:

```
(emotion Dynamic Stimulating Mental Experience DESPERTAR5 INCITAR9
         ESTIMULAR7
 creation BoundedEvent Physical Existence Cause CAUSAR3 REALIZAR2 CREAR2)
```

Fig. 24. Componential descriptions of *encender* and *despertar*.

(a) *dar*:

```
(creation DAR5
 creation BoundedEvent Physical Existence Cause CAUSAR3 REALIZAR2 CREAR2)
```

(b) *producir*:

```
(creation PRODUCIR5 CAUSAR4 OCASIONAR1
 creation BoundedEvent Condition Cause HACER6
 creation BoundedEvent Physical Existence Cause CAUSAR3 REALIZAR2 CREAR2)
```

Fig. 25. Componential descriptions of *dar* and *producir*.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

(a) *guardar*:
(possession Possession Relation GUARDAR4 CONSERVAR2 MANTENER5
possession Possession Relation TENER4)

(b) *concervar*:
(possession Possession Relation CONSERVAR2 MANTENER5 GUARDAR4
possession Possession Relation TENER4)

(c) *entrar*:
(possession ENTRAR5
possession ACCEDER3
possession RECOBRAR1 RECUPERAR5
possession Dynamic Possession CONSEGUIR5 TOMAR9 SACAR20 OBTENER4)

Fig. 26. Componential descriptions of *guardar*, *conservar*, and *entrar*.

Again, the overlap between the semantic descriptions is large enough to allow us to identify co-occurrences of emotion nouns with *dar* as collocations of the same type as co-occurrences of emotion nouns with *producir*.

In Experiment 2, the componential descriptions of LF-instances in the training sets and the descriptions of candidate bigrams were much more diverse. Instances in a training set seldom shared hyperonyms, and candidates seldom contained a verb or a noun already occuring as base/collocate in an LF-instance in the training set. The different parametrization of the classification procedure in the case of one field and multiple-field classifications (in terms of the parameters τ and ρ) takes into account the different profiles of the training sets; cf. Section 5.

The classification procedure is sensitive to the description of the LF-instances in the training sets. Consider, for instance, the classification with respect to IncepFunc_1 in Experiment 1 (see also Section 7.2). Some of the candidate bigrams that contain either *conservar* or *guardar* have been erroneously classified in the runs with $\Delta > 0.7$ as IncepFunc_1-collocations. During the first iteration of the classification procedure, *conservar esperanza* lit. 'conserve hope' and *guardar esperanza* lit. 'keep hope' were identified as IncepFunc_1-collocations; the other candidate bigrams in the subsequent iterations were identified because of the similarity of their descriptions to the descriptions of *conservar esperanza* and *guardar esperanza*.

The two initial erroneous classifications occured because a common IncepFunc_1-collocate is *entrar* 'enter', which received a semantic description similar to *conservar* and *guardar* (see Figure 26), and also because the description of *esperanza* 'hope' contains a small number of hyperonyms. As a result, abstract BCs and TCs contribute overproportionally to the description.

7.2 Parameters of influence

The experience gained in Experiments 1 and 2 suggests that the results of the classification are greatly influenced by three factors: (i) the similarity of the LFs used for classification, (ii) the composition of the training sets, and (iii) the quality of the descriptions in SpanWN.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

If the LFs drawn upon are very similar, the semantic description of their instances will also be similar. Therefore, the error rate of the classification procedure will inevitably be significantly higher in cases where the LFs are similar.

The composition of the training sets is decisive in terms of the results of the classification procedure. For one-field classification, it is crucial that a training set for a given LF contain

- one or several representative bases for each of the field's semantic characteristics,
- a representative cross-section of the collocates (i.e. values for the LF in question).

These requirements imply that the minimal size of training sets may vary from LF to LF. This is also reinforced by the results of Experiment 1. In the case of multiple-field classification, these requirements are not realistic. Therefore, weaker requirements must be applied. Here, it is crucial that a training set for a specific LF contain

- several representative bases for each field for which collocations are being acquired,
- the most prominent collocates covered by this LF in each relevant field.

As Experiment 2 illustrated, acceptable quality figures can be achieved with training sets of a comparatively small size. However, it is sufficiently obvious that the quality of the classification is directly related to the quality of the descriptions in SpanWN. Cross-field classification is especially sensitive to the lack of the proper sense of a word and to an imprecise description of a sense of a word. Both cases were encountered in the process of classification. Let us consider each in turn.

Lack of the proper sense. For instance, for *observar* as in *observar* [*la*] *costumbre* 'observe the custom' seven readings are given in SpanWN:

OBSERVAR**1** ≡ 'explain'
OBSERVAR**2** ≡ 'comment', 'point out'
OBSERVAR**3** ≡ 'notice'
OBSERVAR**4** ≡ 'note'
OBSERVAR**5** ≡ 'detect'
OBSERVAR**6** ≡ 'watch attentively'
OBSERVAR**7** ≡ 'watch', 'keep an eye on'

None of these readings corresponds to that of *observar* in the collocation *observar* [*la*] *costumbre*. For *celebrar* as in *celebrar* [*una*] *entrevista* lit. 'celebrate an interview', two readings are given:

CELEBRAR**1** ≡ 'host', 'entertain'
CELEBRAR**2** ≡ 'celebrate'

Again, none of these readings covers the meaning of *celebrar* in *celebrar* [*una*] *entrevista*. In the same vein, the meanings of, for instance, *tomar* as in *tomar notas*

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

'take notes', of *guardar* as in *guardar* [*la*] *esperanza* 'keep hope', of *pisar* as in *pisar* [*una*] *mina* 'strike mine', and of *responder* as in *responder* [*a un*] *tratamiento* 'respond to a treatment' are not covered in SpanWN. For certain nouns (such as *domicilio* in *establecer* [*un*] *domicilio* 'establish a home', *record* in *establecer* [*un*] *record* 'establish a record', *bar* in *llevar* [*un*] *bar* lit. 'carry a bar'), the proper senses are likewise not available.

In cases where the proper sense of a lexical item was not available, the most similar sense was taken as substitute. But quite often the SpanWN-description of the most similar sense deviated rather significantly from the description of the proper sense, which led then to a distorted score of the combination. When this combination was an LF-instance in a training set, the result was an imprecise centroid. In the case of a candidate bigram, the result was a faulty score possibly leading to an incorrect classification.

Imprecise Description. An imprecise description of a lexical item led to a distorted score of the combination in which it occured and, therefore, had the same consequences for the classification as in the previous case.

Among the imprecise descriptions that negatively influenced classification we encountered, for example, the labeling of *disparar* in *disparar* [*una*] *pistola* 'fire a pistol' as 'competition', of *seguir* in *seguir* [*un*] *curso* 'follow a course' as 'change', and of *ejercer* in *ejercer autoridad* 'exercise authority' and of *utilizar* in *utilizar* [*un*] *teléfono* 'use a phone' as 'consumption'.

Further obstacles to classification were the different descriptions of semantically similar words such as *periodico* 'newspaper' and *libro* 'book', *gastos* 'charges' and *costas* 'costs', etc. Figure 27 shows the descriptions of *periodico* and *libro*.

(a) *libro*:

```
(communication LIBRO3
 communication Part Representation APARTADO2 SECCION3
 communication Artifact LanguageRepresentation ESCRITO1 ESCRITURA5
 communication Artifact LanguageRepresentation COMUNICACION-ESCRITA1
               LENGUAJE-ESCRITO1
 3rdOrderEntity Social Purpose Mental Communication COMUNICACION1
 Relation Social RELACION-SOCIAL1
 Relation RELACION1 ABSTRACCION1)
```

(b) *periodico*:

```
(communication PERIODICO2 DIARIO3
 communication PRENSA3 PRENSA-PUBLICA1
 communication MEDIO-DE-COMUNICACION-ESCRITA1
 communication Representation MEDIO-DE-COMUNICACION1
 artifact Functional MANERA2 METODO1 MEDIO2
 artifact Artifact Instrument UTILLAJE1
 Artifact Object ARTEFACTO1
 Natural Object COSA1 OBJETO-INANIMADO1 OBJETO-FISICO1 OBJETO1
   ENTIDAD1)
```

Fig. 27. Componential descriptions of *libro* and *periodico*.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Despite such obvious obstacles, which are unavoidable in any large scale resource such as SpanWN, the results show that SpanWN is a valuable resource for the task of fine-grained semantic classification of word combinations.

8 Related work

The interpretation of the notion of collocation as a sequence of words that frequently appear together – either adjacently or interrupted by other words – is attractive for 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. Therefore, it is natural that it is this interpretation that underlies the vast majority of the approaches to automatic identification and extraction of collocations; e.g. see Choueka *et al.* (1983), Church and Hanks (1989), Smadja (1993) and Justeson and Katz (1995). As a rule, these approaches provide plain lists of presumed collocations, possibly enriched with POS-information. No semantic information is given.

A number of approaches focus on the extraction of specific types of collocations such as support verb constructions, SVCs (Grefenstette and Teufel, 1995; Tapanainen *et al.*, 1998).[18] They classify, to an extent, binary word combinations into SVCs and "non-relevant" combinations. In this respect, our goals are similar: we also aim at a classification of binary combinations as collocations – however, our classification is more detailed, uses explicit semantics, and is not restricted to SVCs.

In general, approaches to the extraction of collocations 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.

Another strand of research that is related to ours is the acquisition of cooccurrence restrictions (or selectional preferences) (Resnik, 1993; Ribas, 1995; Sanfilippo, 1997; McCarthy, 1997; Li and Abe, 1998; McCarthy, 2000; Clark and Weir, 2001). When interpreted from the perspective of verb-noun collocation extraction, the acquisition of cooccurrence restrictions aims to determine classes of nominal bases (or semantic restrictions thereof) for (possibly predefined) collocate readings of verbs (LFs in our application). Most approaches to the acquisition of cooccurrence restrictions use WordNet's hyperonymy hierarchy for their classifications. This makes them especially interesting in the context of our approach. In particular, Sanfilippo (1997) is similar to our approach. Sanfilippo's algorithm consists of three stages. In the first stage, syntactic patterns under consideration are retrieved from the corpus by partial parsing, lexical items involved in the patterns are lemmatized, and the dependents of each individual syntactic head and the heads of each individual dependent are collected into separate sets. In the second stage, the set associated with a head, respectively dependent, is partitioned into subsets with congruous meanings using Resnik's notion of similarity (Resnik, 1995; Resnik, 1999) applied to the WN

[18] SVCs are represented by the $Oper_i$- and $Func_i$-LFs (i = 0,1,2,3).

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

hyperonym hierarchy. The result of this stage are combinations of the form 'head–collection of (disambiguated) semantically similar dependents' and 'dependent–collection of (disambiguated) semantically similar heads'. In the third stage, the individual heads and dependents are then disambiguated and the semantically similar disambiguated combinations are grouped. The third stage thus results in pairs of sets (B, C) where for any $b_i, b_j \in B$ and any $c_k, c_l \in C$ $sim(b_i, b_j) > \sigma, sim(c_k, c_l) > \sigma$ (with σ being a similarity threshold), and any $b_i \in B$, and any $c_k \in C$ cooccur. If a specific (B, C) is a set of collocations (recall that a (B, C) can also contain sets of freely cooccurring elements), these collocations are per force all of the same type. More precisely, they are all instances of the same LF. However, since no explicit typology is used, no explicit semantic features can be assigned to them. Also, no distinction can be made between collocations and free word combinations.

Approaches presented elsewhere (Resnik, 1993; Ribas, 1995; Li and Abe, 1998; McCarthy, 2000; Clark and Weir, 2001) differ from ours mainly in that they consider the semantic description of one of the elements of a binary word combination. Starting from a concrete verbal (or, possibly, any other predicative) lexical item, they attempt to find the most general class in the WN-hierarchy that subsumes the selectional preferences of this item with respect to one of its argument roles. The measures applied to calculate the suitable level of generalization vary. For instance, Resnik uses the *association score* (a measure of how precisely a concept class denoted by a synset satisfies the selectional restrictions of the lexical item in question), Li and Abe and McCarthy use different implementations of the *Minimum Description Length*, and Clark and Weir use the χ^2- and G^2-measure, respectively. From the viewpoint of the classification of word combinations in terms of collocations, these approaches determine, for a given verbal collocate, a class of bases. However, no field-independent generalization seems possible. Furthermore, since usually no explicit semantic typology for the predicative lexical items is supplied, what has been said above with respect to Sanfilippo's (1997) approach applies here as well.

Other work to which our approach is similar is the semantic classification of either single lexical items or binary relations between lexical items using machine learning techniques (Siegel, 1999; Merlo and Stevenson, 2001; Rosario and Hearst, 2001). As mentioned in Section 3.2, our approach can be considered as an implementation of the *instance-based* (or, more precisely, *nearest neighbor*) learning technique. This technique is simple, but in our case it leads to results of rather high quality.

In the classification approaches known from the literature, the most popular techniques are decision tree induction, rule learning and neural networks. Often, the performance of several techniques for a given task is evaluated contrastively. Thus, Merlo and Stevenson (2001) use for the task of verb classification that is based on the distribution of several grammatical and lexico-semantic features (such as transitivity, causativity, and animacy) decision trees and rule sets (generated by C5.0, a newer version of Quinlan's (1992) C4.5 program). Siegel (1999) explores four different ML-techniques (decision tree induction, clustering, log-linear regression, and genetic programming) for aspectual classification of verbs. Similar to the approach by Merlo and Stevenson, Siegel's approach is based on the distribution of surface indicators (in his case: occurrence of certain tenses, passive voice, and occurrence of

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

adverbials of a certain kind), rather than on semantic features of the classified items (as in our approach). In both approaches, the indicators are tested in isolation and in combination.

In Rosario and Hearst (2001), predetermined semantic relations between the constituents in medical noun compounds are learned by neural networks. To generalize across the compound constituents, Rosario and Hearst use the MeSH semantic hierarchy (Lowe and Barnett, 1994), which is organized following the *International Decimal Classification*. For instance, the noun *flu* has the descriptor 'D4.808.54.79.429.154.349', and *vaccination* the descriptor 'G3.770.670.310.890'. Each substring of a descriptor that is separated by a '.' stands for a class at a distinct level of the hierarchy. Rosario and Hearst experiment with different depths of the classification of nouns (e.g. the description for *flu* at depth 3 is 'D4.808.54', that of depth 4 is 'D4.808.54.79'). A noun compound is represented by a vector that consists of the concatenation of the descriptors of both constituents. This representation is thus similar to the representation of binary word combinations in terms of WN-hyperonymy hierarchies in our approach. Thanks to the use of neural networks, Rosario and Hearst are able to simultaneously classify across all 36 relations on which their approach is based. In contrast, our algorithm takes each LF in sequence classifying the candidate bigrams with respect to it one by one.

9 Conclusions and remaining issues

In this paper, we presented a strategy for the classification of verb-noun combinations that uses the typology of LFs as the classification schema and the hyperonym hierarchies provided by SpanWN as the source for the semantic componential description of the lexical items involved. To our knowledge, this is the first attempt to automatize the recognition of LF-instances among candidate combinations on a broader scale.

The strategy has been implemented and applied to Spanish material. The experiments that have been carried out demonstrated that the proposed techniques achieve a rather high quality classification and can thus be used, for example, for the following purposes:

(i) classifying verb-noun combinations acquired from corpora by partial parsing in terms of LFs;
(ii) assignment of semantics to collocations listed in dictionaries or acquired from corpora by a non-semantically driven collocation extraction approach;
(iii) filtering out word combinations that have erroneously been classified as collocations;
(iv) discovering stable semantic patterns (other than LFs) between lexical items in a corpus by bootstrapping from small (training) sets of positive samples.

The main linguistic lesson we can learn from the success of our experiments is that a model for the composition and recognition of the semantics of collocations exists, and that this model is based on semantic analogy. This lesson contradicts the traditional view that collocations are strictly non-compositional.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Several issues still remain to be tackled at this point. The four most important ones are:

- The experiments showed that classification which is restricted to a single semantic field is considerably more accurate and more comprehensive than non-restricted classification. This seems due to the wide semantic variation of the candidate bigrams in the non-restricted case. However, so far we do not have any experience of how our classification performs in general on single field classification. Further experiments with other fields need to be carried out to obtain more data. If they turn out to support this hypothesis, the higher costs of field-specific classification (with field-specific training sets) are justified.
- In Experiments 1 and 2, we worked with disambiguated candidate bigrams. Thus, the description of a bigram used as input to the classification procedure proper was a pair of sets $(C_n^{s_i}, C_v^{s_j})$, with $C_n^{s_i}$ being the componential description of the ith sense of the noun and $C_v^{s_j}$ the description of the jth sense of the verb. As indicated above, our approach can also be applied to non-disambiguated elements of candidate bigrams – at a considerably higher cost. Experiments with non-disambiguated candidate bigrams have not yet been completed. Therefore, it is still unclear at this point, how our approach performs in practice with non-disambiguated candidate bigrams. Our goal is to be able to work with non-disambiguated data.
- So far, our study covers a major subset of verb-noun LFs, and we are confident that with verb-noun LFs that have not yet been tested (such as *Son*: *dog barks*, *teeth chatter*, *hurricane roars*, etc. and *Degrad*: *teeth decay*, *temper frays*, *discipline decays*, etc.) we will achieve similar quality figures. However, our goal is to be able to classify not only verb-noun bigrams, but, rather, all types of bigrams. We expect that our methodology can also be used to classify adjective-noun and adverb-verb combinations, provided that a resource with a lexico-semantic description of adjectives and adverbs is available. For the time being, only rudimentary descriptions of adjectives and adverbs can be obtained from SpanWN. Therefore, we plan experiments on English with the Princeton WN. Since, as mentioned in Section 4, the Princeton WN does not use semantic labels of the BC and TC type, it remains to be seen whether the descriptions of lexical items as provided by the Princeton WN are sufficiently discriminatory for our task.

  For closed-class words such as prepositions no lexico-semantic descriptions are available. Therefore, our methodology cannot be used to identify and classify preposition-noun and preposition-verb collocations. A different approach is needed to account for these types of collocations.
- In our experiments, we tested the classification of verb-noun combinations in terms of LFs as a "stand-alone" application: the candidate combinations are taken from a manually compiled list. In a more general scenario of the acquisition of collocations from the corpus, the classification will be preceded by a partial parsing stage, in which candidate combinations will be retrieved from the corpus, and possibly also by a WSD stage. We are working on such

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

a scenario in cooperation with the University of Barcelona. The corpus on which this work is based is the POS-tagged corpus LEXESP (Sebastián, 2000). For the parsing stage, we use the parser described in Castellón *et al.* (1998).

Another issue that needs to be tackled at some point in the future is the use of unannotated corpora. So far, we assume that only those candidate verb-noun bigrams that possess the syntactic structure of at least one of the LFs are considered for classification. In other words, our approach requires that either the candidate bigrams be retrieved in a partial parsing stage or that the corpus be syntactically annotated so as to allow for the retrieval of bigrams by browsing the corpus. None of these options can always be presupposed. One way to acquire candidates from unannotated corpora would be to use heuristics for the identification of specific head-dependency relations between elements of a sentence, filtering out verb-noun combinations that would serve as input to the classification procedure. This approach has been pursued in Riccardi *et al.* (1999). Another option would be to start from a list of given nouns (verbs) in order to look for verbal governors (nominal dependents) again using heuristics.

The focus of our work has so far been on Spanish. This work will continue with the objective of building up a Spanish verb-noun LF-dictionary. At the same time, we plan to extend our work to English and French. English is well-suited for our experiments. First, 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. And second, 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.

Appendix

Deviance of the candidates with respect to the centroid in Experiment 1

Table 11 shows the deviance of the score of some candidate bigrams from the LF-centroids. The centroids are given in parantheses.

When a candidate bigram is an instance of an LF, the deviance of its score from the centroid of this LF is in bold. As already done in the course of the article, we add, when it appears advantageous, an article and a preposition. Also, we indicate when the base requires an adjectival modification (as in *producir una* [ADJ] *decepción*). *N* stands for 'noun'.

Examples of candidate score deviance in Experiment 2

The following tables show the thirty smallest deviances of candidate scores from the centroids of the LFs used in Experiment 2 during the first iteration of the classification procedure. The deviance of positive candidate bigram scores is in bold.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 11. *Deviance of the scores of some candidates from the centroids in Experiment 1 ($C_2F_1 = Caus_2Func_1$, $O_1 = Oper_1$, $CO_1 = ContOper_1$, $FF_0 = FinFunc_0$, $IF_1 = IncepFunc_1$)*

	C_2F_1 (15.131)	O_1 (12.743)	CO_1 (3.481)	FF_0 (10.863)	IF_1 (1.337)
[la] admiración cesa	1.170	1.684	5.626	**0.029**	4.599
[la] admiración se desvanece	0.985	1.512	6.153	**0.139**	6.730
guardar admiración	1.297	1.289	**0.225**	1.344	1.306
sentir admiración	1.027	**0.004**	3.909	1.925	8.964
experimentar alegria	1.021	**0.013**	3.935	1.941	8.984
tener alegria	0.980	**0.108**	8.679	1.282	6.940
producir [una ADJ] decepción	**0.133**	2.175	5.218	1.780	9.987
[la] desesperación se apodera [de *N*]	1.089	1.448	1.519	1.180	**0.539**
provocar [una ADJ] desesperación	**0.470**	2.596	6.994	2.144	15.403
causar enfado	**0.910**	1.254	1.753	1.162	2.532
causar entusiasmo	**0.169**	2.208	5.188	1.676	9.503
conservar [la] esperanza	1.355	1.346	**0.322**	1.406	1.319
[la] esperanza se desvanece	0.968	1.556	6.753	**0.018**	7.407
dar horror	**0.074**	2.283	5.714	1.819	11.933
experimentar horror	1.038	**0.307**	3.236	1.733	7.000
provocar indignación	**0.322**	2.433	6.403	2.034	14.022
ira cesa	1.146	1.670	5.690	**0.020**	4.667
[la] ira se apodera	1.092	1.422	1.507	1.175	**0.600**
[la] ira desaparece	0.943	1.812	9.108	**0.193**	7.827
sentir ira	1.010	**0.044**	3.967	1.931	9.168
dar miedo	**0.601**	2.811	7.801	2.137	16.911
[el] miedo desaparece	0.937	1.868	9.623	**0.268**	8.308
[el] miedo entra [en *N*]	1.185	1.193	2.534	1.163	**0.195**
[el] miedo invade [*N*]	1.009	1.177	4.298	0.923	**0.234**
experimentar odio	1.018	**0.092**	4.131	1.989	9.601
causar panico	**0.151**	2.228	5.257	1.689	9.677
guardar respeto	1.274	1.278	**0.323**	1.305	1.325
tener [una] sorpresa	0.966	**0.070**	8.123	1.246	6.524
[el] temor cesa	1.156	1.636	5.252	**0.118**	4.340
conservar temor	1.282	1.275	**0.089**	1.324	1.280
[el] temor entra [en *N*]	1.166	1.174	2.304	1.147	**0.3218**
producir temor	**0.070**	2.316	5.831	1.855	11.395
tener verguenza	0.961	**0.271**	9.578	1.309	7.660

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 12. *CausFunc$_0$ classification (with the initial centroid of 4.631626); D_{CF_0} stands for 'deviance from the centroid of CausFunc$_0$'*

Candidate	D_{CF_0}	Candidate	D_{CF_0}
dar [una] conferencia	**0.010**	levantar [el] entredicho [a]	**0.770**
hacer [una] tentativa [de V_{inf}]	**0.040**	hacer [un] favor	**0.817**
publicar [un] libro	**0.092**	hablar [una] lengua	0.818
cumplir [una] promesa	0.101	hacer [un] papel	**0.834**
organizar [una] operación	**0.136**	causar [una] sensación	**0.856**
hacer [un] esfuerzo	**0.276**	iniciar [la] negociación	**0.878**
poner orden	**0.277**	dictar [la] sentencia	**0.880**
hacer obras	**0.315**	hacer [una] llamada	**0.881**
hacer [una] corrección	**0.322**	iniciar [una] conversación	**0.887**
presentar [una] excusa	**0.394**	llevar [un] diario	0.925
causar alarma	**0.544**	iniciar [una] discusión	**0.962**
reconocer [una] acusación	0.687	acatar [la] ley	0.979
publicar [un] diario	**0.694**	iniciar [un] debate	**1.161**
atender [una] solicitud	0.725	tener simpatia	1.169
celebrar [una] entrevista	**0.757**	responder [a una] objeción	1.211

Table 13. *Oper$_1$-classification (with the initial centroid of 3.041775); D_{O_1} stands for 'deviance from the centroid of Oper$_1$'*

Candidate	D_{O_1}	Candidate	D_{O_1}
poner reparos	**0.074**	presentar [una] demanda	**0.503**
soltar [una] indirecta	**0.100**	preparar [una] lista	0.507
sellar [un] compromiso	0.103	establecer [un] record	0.508
sacar apuntes	**0.113**	cometer [una] falta	**0.534**
celebrar elecciones	**0.116**	cumplir [un] deber	0.602
poner [una] queja	**0.210**	recibir [una] visita	0.633
reconocer acusación	0.305	hacer [un] examen	0.696
lanzar [un] grito	**0.378**	sacar [una] canción	**0.764**
acudir [a una] cita	0.399	imponer [el] orden	0.808
hacer [un] comentario	**0.405**	llevar [un] diario	0.837
pronunciar dictar [la] sentencia	**0.434**	organizar [una] manifestación	0.845
sacar [una] novela	0.447	provocar [una] discusión	0.882
responder [a una] objeción	0.449	dar alarma	**0.923**
promover [una] campaña	0.454	rendirse [al] chantaje	0.928
dar [una] explicación	**0.492**	dar [un] grito	**0.945**

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 14. *$Oper_2$-classification (with the initial centroid of 2.3011153); D_{O_2} stands for 'deviance from the centroid of $Oper_2$'*

Candidate	D_{O_2}	Candidate	D_{O_2}
estar [bajo] arresto	**0.198**	encontrar resistencia	**1.236**
prestar ayuda	0.213	caer [en una] emboscada	1.255
recibir atención	**0.221**	recibir [un] abrazo	**1.290**
encontrar [una] aversión	**0.265**	celebrar elecciones	1.550
rendirse [al] chantaje	0.293	saldar [una] deuda	1.609
recibir [una] disculpa	**0.433**	liquidar [una] deuda	1.609
someterse [a un] análisis	**0.573**	sufrir [un] ataque	**1.748**
recibir aprobación	**0.732**	prestar atención	1.843
estar [bajo el] fuego	**0.766**	tener [un] compromiso	1.884
encontrar [un] rechazo	**0.803**	tomar medidas	1.919
atender [una] llamada	0.968	seguir [un] curso	2.020
estar [en] observación	0.988	obedecer [una] orden	2.025
confirmar [el] pronóstico	1.050	tener simpatia	**2.050**
confirmar [una] predicción	1.050	satisfacer [un] requisito	2.083
corroborar [una] predicción	1.050	tener autoridad	2.089

Table 15. *$Real_1$-classification (with the initial centroid of 1.6136968); D_{R_1} stands for 'deviance from the centroid of $Real_1$'*

Candidate	D_{R_1}	Candidate	D_{R_1}
pagar [una] deuda	**0.092**	estar [en un] hospital	1.672
lanzar [un] misil	**0.095**	estar [bajo la] influencia	1.808
lanzar [una] bomba	**0.141**	dictar [la] sentencia	1.813
publicar [un] periodico	0.350	utilizar [el] teléfono	**1.872**
publicar [una] revista	0.400	despachar [una] prescripción	1.929
apretar [un] gatillo	**0.638**	surtir efecto	1.950
poner [el] freno de mano	**0.827**	atender [la] solicitud	1.957
echar [el] ancla	**0.852**	celebrar elecciones	1.959
esgrimir [una] espada	**0.921**	corroborar [una] predicción	1.987
pisar [una] mina	1.038	confirmar [una] predicción	1.987
pisar [el] freno	**1.063**	confirmar [el] pronostico	1.987
tener autoridad	1.298	sacar [una] canción	2.006
estar [al] ancla	**1.428**	acatar [la] ley	2.032
tener simpatia	1.510	erigir [una] estatua	2.117
estar [en] cama	**1.616**	sufragar gastos	2.129

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Table 16. *$Real_2$-classification (with the initial centroid of 3.2169716); D_{R_2} stands for 'deviance from the centroid of $Real_2$'*

Candidate	D_{R_2}	Candidate	D_{R_2}
obedecer [la] regla	**0.053**	liquidar [una] deuda	1.526
estar [sobre] la agenda	0.103	saldar [una] deuda	1.526
aceptar [un] argumento	**0.304**	echar [la] culpa	1.640
reconocer [una] acusación	**0.326**	seguir [la] etiqueta	**1.674**
sellar [un] convenio	0.482	hacer tentativa	1.687
tener [un] pasaporte	0.591	satisfacer [una] condición	**1.742**
seguir [un] consejo	**0.594**	aceptar [una] excusa	**1.763**
estar [bajo un] bombardeo	0.805	recibir [un] homenaje	1.813
cumplir [una] promesa	1.040	confirmar [una] profesia	**1.829**
tener simpatia	1.076	sacar [una] canción	1.830
dictar [una] sentencia	1.120	encontrar [una] contradicción	1.832
cumplir [un] criterio	**1.22**	prestar ayuda	1.843
guardar [un] secreto	1.241	tener [el] apoyo [de *N*]	1.882
tener ayuda	1.427	aceptar [una] invitatión	**1.898**
estar [bajo la] influencia	1.450	aceptar [una] disculpa	**1.911**

Acknowledgements

The work described in this paper has been partially supported by a Xunta de Galicia grant to the project *Optimización de la indexación semántica por medio de información colocacional* under the number PGIDIT02PXIB30501PR.

The results of the first preliminary tests on the automatic classification of collocations with respect to LFs were presented together with Margarita Alonso Ramos at the *Journée d'études de l'ATALA: La Collocation* organized by Béatrice Daille and Geoffrey Williams in Paris in January 2001. The discussions with Margarita helped me clarify my ideas and to get to the point where they became implementable. Many thanks to Marcelo Magallón and Begoña Sanromán for providing the major part of the Spanish material.

The paper has been read and commented upon by Margarita Alonso, Miguel Alonso, Sylvain Kahane, Stefan Klatt, Igor Mel'čuk, Alain Polguère and Helmut Schmid. Terry Alvarez corrected my English. I would like to express my heartfelt gratitude to all of them. Many thanks also to the two anonymous reviewers of the *NLE Journal*, whose detailed comments and suggestions have not only significantly improved the final form of this paper, but also prompted me to run a series of new experiments which have led to a more optimal parametrization of the classification procedure and thus to better results.

References

Allerton, D. J (1984) Three or Four Levels of Co-occurrence Relations. *Lingua* **63**: 17–40.

Alonso Ramos, M. (2002) Hacia un Diccionario de colocaciones del español y su codificación. In: M. A. Martí, editor, *Estudios sobre el léxico*. Edicions de la Universitat de Barcelona. presented at the *I Symposium internacional de lexicografía, Universitat Pompeu Fabra, Barcelona.*

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Beckwith, R., Fellbaum, C., Gross, D. and Miller, G. A. (1991) WordNet: A Lexical Database Organized on Psycholinguistic Principles. In: U. Zernik, editor, *Lexical Acquisition: Exploiting on-line resources to build a lexicon*, pp. 211–232. Erlbaum, Hillsdale, NJ.

Benson, M., Benson, E. and Ilson, R. (1997) *The BBI Combinatory Dictionary of English*. John Benjamins, Amsterdam.

Benson, M. (1989) The Structure of the Collocational Dictionary. *International Journal of Lexicography*, **2**(1): 1–13.

Castellón, I., Civit, M. and Atserias, J. (1998) Syntactic Parsing of Unrestricted Spanish Texts. *Proceedings First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain.

Choueka, Y., Klein, T. and 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.

Christ, O., Schulze, B. M., Hofmann, A. and König, E. (1999) The IMS Corpus Workbench: Corpus Query Processor (CQP) – User's Manual. Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

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

Clark, S. and Weir, D. (2001) Class-Based Probability Estimation Using a Semantic Hierarchy. *Proceedings 2nd Conference of the North American Chapter of the Association for Computational Linguistics*.

Cowie, A. P. (1993) Phraseology. In: R. E. Asher and J. M. Y. Simpson, editors, *The Encyclopedia of Language and Linguistics, Vol. 6*, pp. 3168–3171. Pergamon, Oxford.

Fellbaum, Ch. (1998) *WordNet. An Electronic Lexical Database*. MIT Presss, Cambridge, MA.

Firth, J. R. (1957) Modes of Meaning (1951). In: J. R. Firth, editor, *Papers in Linguistics 1934–1951*, pp. 190–215. Oxford University Press, Oxford.

Fontenelle, T. (1997) *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Niemayer, Tübingen.

Giguet, E. and Vergne, J. (1997) Syntactic Analysis of Unrestricted French. *Proceedings International Conference on Recent Advances in Natural Languages Processing (RANLP'97)*, pp. 11–13, Tzigov Chark, Bulgaria.

Grefenstette, G. and Teufel, S. (1995) Corpus-Based Method for Automatic Identification of Support Verbs for Nominalizations. *Proceedings Biannual Meeting of the European Chapter of the Association for Computational Linguistics*, pp. 27–31.

Halliday, M. A. K. (1961) Categories of the Theory of Grammar. *Word* **17**: 241–292.

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

Hausmann, F.-J. (1984) Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortwendungen. *Praxis des neusprachlichen Unterrichts* **31**(4): 395–406.

Heid, U. (1992) Décrire les collocations: deux approches lexicographiques et leur application dans un outil informatisé. *Traduction et terminologie* **(2/3)**.

Heid, U. (1998) Towards a Corpus-Based Dictionary of German Noun-Verb Collocations. *Proceedings of the EURALEX International Congress*, Liège, Belgium.

Hindle, D. (1983) Deterministic Parsing of Syntactic Noun-Fluencies. *Proceedings 21st Annual Meeting of the Association for Computational Linguistics*.

Justeson, J. S. and Katz, S. M. (1995) Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering* **1**: 9–27.

Kahane, S. and Polguère, A. (2001) Formal Foundation of Lexical Functions. *Collocation: Computational Extraction, Analysis and Exploitation, Workshop at ACL 2001*, pp. 8–15. Toulouse.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Klatt, S. (1997) Pattern-matching Easy-first Planning. In: A. Drewery, G.-J. M. Kruijff and R. Zuber, editors, *Proceedings Second ESSLLI Student Session*, Aix-en-Provence.

Leacock, C. and Chodorow, M. (1998) Combining Local Context and WordNet Similarity for Word Sense Identification. In: C. Fellbaum, editor, *WordNet. An Electronic Lexical Database*, pp. 265–283. MIT Press, Cambridge, MA.

Li, H. and Abe, N. (1998) Generalizing Case Frames Using a Thesaurus and the MDL Principle. *Computational Linguistics* **24**(2): 217–244.

Lowe, H. J. and Barnett, G. O. (1994) Understanding and Using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches. *Journal of the American Medical Association* **271**(4): 1103–1108.

Manning, C. D. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. MIT Press.

McCarthy, D. (1997) Word Sense Disambiguation for Acquisition of Selectional Preferences. *Proceedings 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. *Proceedings 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 256–263.

Mel'čuk, I. A. and Wanner, L. (1996) Lexical Functions and Lexical Inheritance for Emotion Lexemes in German. In: L. Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pp. 209–278. Benjamins, Amsterdam.

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

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

Mel'čuk, I. A. (1996) Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In: L. Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pp. 37–102. Benjamins Academic, Amsterdam.

Merlo, P. and Stevenson, S. (2001) Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics* **27**(3): 373–408.

Miller, G. A. (1990) WordNet: An On-Line Lexical Database. *International Journal of Lexicography* **3**(4).

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

Ogawa, Y., Morita, T. and Kobayashi, K. (1991) A Fuzzy. Document Retrieval System Using the Keyword Connection Matrix and a Learning Method. *Fuzzy Sets and Systems* **39**: 163–179.

Polguère, A. (2000) Towards a Theoretically-Motivated General Public Dictionary of Semantic Derivations and Collocations for French. *Proceedings EURALEX International Congress*, pp. 517–527. Stuttgart.

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

Quinlan, R. (1992) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, CA.

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

Resnik, P. (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings International Joint Conference on Artificial Intelligence*.

Resnik, P. (1999) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *JAIR* **11**: 95–130.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press

Ribas, F. (1995) On Learning More Appropriate Selectional Restrictions. *Proceedings Biannual Meeting of the European Chapter of the Association for Computational Linguistics*, pp. 112–118. Dublin.

Riccardi, G., Bangalore, S. and Sarin, P. D. (1999) Learning Head-Dependency Relations from Unannotated Corpora. *Automatic Speech Recognition and Understanding Workshop*, Keystone, CO.

Rosario, B. and Hearst, M. (2001) Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy. *Proceedings 6th Conference on Empirical Methods in NLP*, pp. 82–90.

Sanfilippo, A. (1997) Using Semantic Similarity to Acquire Cooccurrence Restrictions from Corpora. *Proceedings Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources at the 35th. Annual Meeting of the Association for Computational Linguistics*, pp. 82–89.

Sanromán, B. (1998) Contribución lexicográfica al estudio de los nombres de emoción. Master's thesis, Universidade da Coruña.

Sebastián, N. (2000) *LEXESP, léxico informatizado del español*. Edicions de la Universitat de Barcelona, Barcelona.

Siegel, E. V. (1999) Corpus-Based Linguistic Indicators for Aspectual Classification. *Proceedings 37th Annual Meeting of the Association for Computational Linguistics*, pp. 112–119.

Sinclair, J. (1994) *Cobuild: English Learner's Dictionary*. Collins, London.

Smadja, F. and 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.

Tapanainen, P., Piitulainen, J. and Järvinen, T. (1998) Idiomatic Object Usage and Support Verbs. *Proceedings COLING/ACL '98*, pp. 1289–1293. Montréal.

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

Vossen, P. (1999) EuroWordNet. General Document. Technical report, University of Amsterdam, Amsterdam.

Williams, G. (1999) *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique*. PhD thesis, Université de Nantes.

https://doi.org/10.1017/S1351324904003328 Published online by Cambridge University Press