Procesamiento del Lenguaje Natural, Revista nº 46 marzo de 2011, pp 107-112 recibido 17-01-11 revisado 10-02-11 aceptado 24-02-11

False Paraphrase Pairs in Spanish for Verbs and Verb+Noun Collocations*

Falsas Paráfrasis de Verbos Plenos y Colocaciones de Verbo Soporte en Español

María Auxiliadora Barrios
Dpto. Lengua Española
Universidad Complutense de Madrid
Paraninfo Ciudad Universitaria s/n
28008, Madrid, Spain
auxiba@filol.ucm.es

Luz Rello
NLP and Web Research Groups
Universitat Pompeu Fabra
C/Tànger, 122-134
08018, Barcelona, Spain
luzrello@gmail.com

Resumen: En este trabajo hemos estudiado algunas paráfrasis contenidas en un recurso lingüístico llamado BADELE.3000, una base de datos que agrupa los 3.600 sustantivos más frecuentes del español, así como más de 2.800 verbos de uso habitual. Su combinación restringida da lugar a más de 23.000 colocaciones, descritas en este recurso mediante el lenguaje formal de las Funciones Léxicas de la Teoría Texto-Sentido. La aplicación de la Regla 18 de dicha Teoría permitió la extracción manual de 777 pares verbos plenos y sus correspondientes paráfrasis de verbos soporte y sustantivo. El presente trabajo deja de lado estos casos, y describe los tres tipos de situaciones para las que no hay paráfrasis: a) colocaciones sustantivo-verbos que no tienen verbo equivalente (*tener gripe, *gripear*); b) verbos que no tienen colocaciones sustantivo-verbos equivalentes (*respirar, *hacer la respiración*); c) colocaciones sustantivo-verbos que aunque tienen verbo equivalente, difieren en el significado (*expedir, hacer una expedición*).
Palabras clave: paráfrasis, colocaciones, Funciones Léxicas, Teoría Texto Sentido

Abstract: In this paper we have studied some pairs of paraphrases which are present in a linguistic resource called BADELE.3000, a data base that contains more than 3,600 high frequency Spanish nouns and 2,800 high frequency Spanish verbs. The restricted combinatory of both kinds of words means more that 23,000 collocations, which are expressed by Lexical Functions, a tool of Meaning-Text Theory. Through the application of Rule 18 of this framework, paraphrase pairs consisting of a verb and a verb-noun collocation were manually extracted from BADELE.3000. The paper focuses on the three sets of pairs that are false paraphrases: a) verb-noun collocations that have no verbal counterpart (*to have flu,*to flu*); b) verbs that have no verb-noun collocation counterpart (*to breath, *to do the breathing*); c) verbs and verb-noun collocation counterpart that differ in meaning (*to count, to do a count*).
Keywords: paraphrases, collocations, Lexical Functions, Meaning-Text Theory.

1 Introduction

The aim of this work is to obtain a set of paraphrases composed of verbs which can be paraphrased with collocations using a support verb plus a noun, e.g., *acoger* ('to welcome'), *dar acogida* (*'to give welcome'). This task requires the discarding of false paraphases such as pairs or words with a morphological but not semantic relationship e.g., *expedir* ('to issue') and *hacer una expedición* ('to explore'). In this study each of the verbs and nouns are lexical unites (Mel'čuk, 1996), that is, each of them are disambiguated and correspond to only one meaning.

The motivation of this study lies in the fact that there is no Meaning-Text Theory rule from which one could fully predict a list

* We thank Socorro Bernardos who helped throughout the implementation process of BADELE.3000; Prof. Leo Wanner for his helpful comments; and Déborah Paton and Borja Menéndez for his help with English collocations.

of paraphrases composed of verbs and their matching collocations as well as the filtering of false paraphrases. This list of pairs can be used for paraphrasing which is found to be useful for many NLP tasks such as machine translation, question answering or summarization (Androutsopoulos and Malakasiotis, 2010).

Moreover, this work completes previous work (Barrios, 2010) in which two lists of false paraphrases were obtained from the same corpus. These two lists include verbs which do not correspond with a collocation, e.g., *respirar* ('to breath'), **hacer respiración*[1] (*'to do the breathing') and nouns which participate in a collocation but have no verbal counterpart e.g., **gripear* (*'to flu'), *tener gripe* ('to have flu').

Although the use of automatic methods for generating paraphrases is successful (Madnani and Dorr, 2010), the semantic precision of the lists presented required to performed this task manually. For this purpose BADELE.3000 was created. BADELE.3000 is a database that contains the 3,300 most frequently used Spanish nouns. The information about each noun includes the definition and the combinatorial possibilities amongst other linguistic information (Barrios, Aguado de Cea, and Ramos, 2009b).

For the formalization of BADELE.3000 Lexical Functions taken from the Meaning Text-Theory (MTT) were applied. The syntagmatic and paradigmatic relations formalized in the Lexical Functions have been successfully applied to lexicography (Mel'čuk, 1996), machine translation (Apresjan et al., 2003), and text generation (Laureau and Wanner, 2007).

In particular, BADELE.3000 paraphrases find their starting point in the MTT number 18 rule for paraphrases, named *fissions à verbe support* (Mel'čuk, 1992) together with other language dependent strategies used specially for Spanish paraphrases. The next section states the methodology used for compiling BADELE.3000, whilst section 3 is devoted to the explanation of the MTT number 18 rule for paraphrases. The remainder of the paper describes the filtering processes used to discriminate false paraphrase patterns found in BADELE.3000: the case of lack of paraphrases (Section 4) and the cases where there is a morphological but not a semantic correlation (Section 5). Finally, the conclusion section states the suitability of BADELE.3000 as a resource.

2 Data: BADELE.3000

The database BADELE.3000 contains 3,600 nouns, which are the most frequent nouns in Peninsular Spanish, and 2,800 high frequency Spanish verbs. This set of words was extracted from a Spanish frequency list selected from the Cumbre Corpus of 20 million words (Almela et al., 2005).

For each selected noun (e.g., *paseo*, 'walk') two types of correlations were found. First, a set of collocations in which the noun occurs (e.g., *dar un paseo*, 'to have a walk') and second, the verbs which are morphologically related to those nouns (e.g., *pasear*, 'to walk').

The restricted combinatory of both kinds of words means more than 20,000 collocations. Actually, BADELE.3000 includes a total of 23,000 relations which are formalized by means of Lexical Functions taken from the MTT: a total of 395 Lexical Functions were applied, which means that nouns participate in 395 kinds different of relation.

Approximately 9,000 lexical relations were obtained fully automatically whilst the remainder required manual work such as the formalization of lexical relations found in combinatorial dictionaries of Spanish (Bosque, 2004; Bosque, 2006).

3 MTT Paraphrase Rule 18: Fissions à verbe support

The baseline chosen to obtain the paraphrase pairs was the application of the MTT Paraphrase Rule 18 *fissions à verbe support* (Mel'čuk, 1992). Each paraphrase pair is composed by a verb and collocations formed by a support verb and a noun.

The baseline is transcribed as: given a verb (C_0) such as *recibir* ('to receive') is paradigmatically related to a noun (S_0 *recibimiento*, 'reception'), if the noun S_0 appears in a collocation together with a support verb (*ofrecer un recibimiento*, *'to give a reception'[2]), then, the verb C_0 corresponds to the collocation and therefore both expressions (*recibir* and *ofrecer un recibimiento*) are interchangeable.

[1] In this paper, the asterisk symbol is used to identify ungrammatical expressions.

[2] Note that in English this verb/collocation pair is not a case of paraphrases as they differ in meaning.

Paraphrase Rule 18: Fissions à verbe support:

C_0	S_0	—II—	$\mathbf{Oper}_1(S_0(C_0))$
recibir	*recibimiento*		*ofrecer un recibimiento*
(‘to receive’)	(‘reception’)		(*‘to give a reception’)

This equivalence rule entails a relationship in which each element of the pair —the verb and the collocation —are interchangeable. For instance (Kahane, 2001).

C_0	S_0	—II—	$\mathbf{Oper}_1(S_0(C_0))$
alegrarse (‘to be happy’)	*alegría* (‘happiness’)		*sentir alegría* (‘to feel ‘happiness’)
morir (‘to die’)	*muerte* (‘death’)		*sucumbir a la muerte* (‘to succumb to death’)

Notice that the relationship between verbs (C_0) and nouns (S_0) with the same meaning is covered by the structural Lexical Funcions S_0 ’derived noun’ and V_0 ’derived verbs’ being both, noun and verb equivalent:

$S_0(V) = V_0(S)$
$S_0(\text{alegrarse}) = V_0(\text{alegría})$
$S_0(\text{‘to be happy’}) = V_0(\text{‘happiness’})$

Following the rule 18, a total of 777 pairs of phrasal paraphrases whose collocation is composed by an abstract noun was found in BADELE.3000.

As presented in section 4.2, there is no means in the language to predict which abstract nouns have a corresponding collocation, e.g: it is said *hacer un guiño* (*‘to make a wink’) together with *guiñar* (‘to wink’) but not **hacer un parpadeo* (*‘to make a blink’) from *parpadear* (‘to blink’). In Table 1 some examples of abstract nouns in a collocation with their corresponding verb are stated. All these examples satisfy the Lexical Function **Oper** (Mel’čuk, 1996).

4 Lack of Paraphrases

In BADELE.3000, there were observed cases in which it was impossible to assign a paraphrase collocation to a verb (Section 4.1) or assign a verb to a collocation formed by a frequent Spanish noun (Section 4.2).

4.1 Non-existing Verbs

One of the problems found when applying the MTT paraphrase rule 18 to BADELE.3000 was the lack of a paraphrase verb corresponding to a collocation in language.

V_0	Collocation = $\mathbf{Oper}_1(S_0)$
seleccionar (‘select’)	*hacer una selección* (‘to make a selection’)
rechazar (‘reject’)	*mostrar rechazo* (‘to show rejection’)
asistir (‘assist’)	*prestar asistencia* (‘to give assistance’)
apoyo (‘support’)	*dar apoyo* (‘to give support’)
investigar (‘research’)	*realizar una investigación* (‘to do a research’)
controlar (‘control’)	*someter a control* (‘to subject to control’)
propagar (‘spread’)	*hacer propaganda* (‘to make propaganda’)
definir (‘define’)	*formular una definición* (‘to formulate a definition’)
difundir (‘broadcast’)	*dar difusión* (‘to broadcast’)
recordar (‘remember’)	*tener un recuerdo* (‘to have a memory’)

Table 1: Paraphrase pairs.

4.1.1 Names of Illnesses and Feelings

The lack of a paraphrase verb corresponding to a collocation in language was found to be particularly frequent in nouns which denote illnesses and feelings which occur in collocations.

For instance, in table 2, regarding the nouns occurring in Spanish collocations, there is no verb in the language which paraphrases the collocation.

V_0	Collocation = $\mathbf{Oper}_1(S_0)$
**gripear* (*‘to flu’)	*tener gripe* (‘to have flu’)
**diabetear* (*‘to diabetes’)	*tener diabetes* (‘to have diabetes’)
**soledear* (*‘to lonely’)	*sentir soledad* (‘to feel loneliness’)
**felicidadear* (*‘to happy’)	*sentir felicidad* (‘to feel happiness’)

Table 2: Lack of Paraphrases: Non-existing Verbs for Nouns Denoting Illnesses and Feelings.

Notice that the Spanish verb *acalorarse* (‘to get hot’) has a more restricted meaning

as it requires the semantic features of human and inchoateness.

4.1.2 Names of Physical and Non-physical Facts

The second group of collocations do not have a paraphrase verb. The collocations are composed of nouns which denote physical facts such as *vistazo* ('look'), *bocanada* ('mouthful'), *esencia* ('essence'), *fragancia* ('fragrance'), *aroma* ('aroma'), *calor* ('heat'), *tufo* ('stench'), *coito* ('coitus'); as well as nouns which denote non-physical facts such as *injusticia* ('injustice'), *estrategia* ('strategy'), *chiste* ('joke'), *bien* ('good'), *culto* ('cult'), *incidente* ('incident'), *conducta* ('behaviour'), *hazaña* ('feat'), *iniciativa* ('initiative'), *milagro* ('miracle'), *eficacia* ('efficiency'); or activities such as *senderismo* ('hiking'), *turismo* ('tourism') or *montañismo* ('mountaineering').

Fifty nouns of this kind occuring in Spanish collocations but without any verb which express such meaning in language were extracted.

S_0	V_0	Collocation = **Oper**$_1$(S_0)
vistazo ('look')	**vistacear* ('to look')	*echar un vistazo* ('to look at sth')
bocanada ('mouthful')	**bocanear* (*'to mouthful')	*dar una bocanada* (*'to do a mouthful')
calor ('heat')	**calorar* ('to heat')	*hacer calor* ('to be hot')
injusticia ('injustice')	**injusticiar* (*'to injustice')	*hacer una injusticia* ('to make an injustice')
estrategia ('strategy')	**estrategiar* (*'to strategy')	*tener una estrategia* ('to have a strategy')
chiste ('joke')	**chistear* ('to joke')	*contar un chiste* ('to tell a joke')
incidente ('incident')	**incidentar* (*'to incident')	*vivir un incidente* (*'to live an incident')
senderismo ('hiking')	**senderear* ('to hike')	*hacer senderismo* ('to go hiking')
turismo ('tourism')	**turistear* (*'to tourism')	*hacer turismo* ('to travel around')
conducta ('behaviour')	**conductear* ('behave')	*tener una conducta* ('to have a behaviour')
realidad ('reality')	**realidacear* (*'to reality')	*hacerse realidad* ('to make sth real')

Table 3: Lack of Paraphrases: Non-existing Verbs for Nouns Denoting Physical and Non-physical Facts.

From a linguistic point of view, no explanation was found for the existence of collocations with a paraphrase verb (Table 1) and collocations with no paraphrase verb (Table 2 and 3).

4.2 Non-existing Collocations

4.2.1 Abstract Nouns

Previous work (Barrios, 2010) pointed out that abstract nouns tend to participate in collocations using support verbs. Nevertheless, some exceptions (Table 4) to this generalization were extracted from BADELE.3000. These abstract nouns denote physical actions —*respiración* ('breathing'), *ovulación* ('ovulation'), *inspiración* ('inspiration'), *parpadeo* ('blink') —, human noises —*clamor* ('clamour'), *tarareo* ('humming'), *canto* ('singing') —, or economic actions —*financiación* ('financing') and *cotización* ('quotation').

S_0	V_0	Collocation = **Oper**$_1$(S_0)
respiración ('breath')	*respirar* ('to breath')	**hacer la respiración* (*'to do the breathing')
ovulación ('ovulation')	*ovular* ('to ovulate')	**hacer la ovulación* (*'to do an ovulation')
parpadeo ('blink')	*parpadear* ('to blink')	**hacer un parpadeo* (*'to do a blink')
clamor ('clamour')	*clamar* ('to clamour')	**soltar un clamor* (*'to do a clamour')
tarareo ('humming')	*tararear* ('to hum')	**lanzar un tarareo* (*'to do a humming')
financiación ('financing')	*financiar* ('to finance')	**hacer una financiación* (*'to do a finance')
cotización ('quotation')	*cotizar* ('to quote')	*hacer una cotización* (*'to do a quotation')

Table 4: Lack of Paraphrases: Non-existing Collocations for Abstract Nouns.

5 False Paraphrases

The application of the MTT Paraphrase Rule 18 can also give false paraphrases which need to be discarded from BADELE.3000.

There are cases where the verb and the noun occurring in the collocation have a different meaning. For instance (de Miguel, 2006) *tener frío* ('to be cold') or *tener cansancio* ('to be tired') do not mean the same as *enfriarse* ('to get cold') or *cansarse* ('to

get tired') as the first word denotes a state whilst the second one a process. Whilst the collocations are perfective the verbs are inchoative.

Although the examples of false paraphrases fulfil the paradigm of: noun (S_0), verb (V_0) and collocation, the requirement of (S_0) = (V_0) is not fulfilled, that is, there is no semantic equivalence between the noun and the verb, even if there is a morphological relationship between the two words.

For instance, the collocation *hacer una expedición* ('to do an expedition') does not mean the same as *expedir* ('to issue'), or the noun *vicio* ('vice') is not ingressive as the verb *enviciarse* ('to become addicted'). The examples of false paraphrases are shown in Table 5.

V_0	Collocation = $\mathbf{Oper}_1(S_0)$
expedir ('to issue')	*hacer una expedición* ('to explore')
atarearse ('to get busy')	*hacer una tarea* ('to do a task')
balancear ('to swing')	*hacer un balance* ('to do a balance')
faenar ('to fish')	*hacer una faena* ('to play a dirty trick')
malear ('to spoil')	*hacer el mal* ('to make evil')
observar ('to observe')	*hacer una observación* ('to make a remark')
contar ('to count')	*hacer cuentas* ('to do a calculation')
aproximar ('to approximate')	*hacer una aproximación* ('to make an estimation')
escribir ('to write')	*hacer una escritura* ('to write a deed')
dejar ('to leave')	*hacer una dejación* ('to be irresponsible')
holgar ('to be idle')	*hacer una huelga* ('to go on strike')

Table 5: False Paraphrases.

Although the criteria chosen to consider a full paradigm as a false paraphrase was a difference in meaning, there are cases in which this meaning variation is subtle. For example, *fracaso* ('failure') and *fracasar* ('to fail'), when it is stated *fracasó en su vida* ('he failed in his life') it means his whole life was a failure and not that he made one failure in his life. Moreover, *lanzar un par de sonrisas* (*'to dedicate a couple of smiles') does not mean the same as to smile twice in Spanish.

6 Conclusions

In this paper we have presented a data base called BADELE.3000. BADELE.3000 is a linguistic resource useful for natural language processing applications as it has already proven to be beneficial for ontologies in previous research (Barrios, Aguado de Cea, and Ramos, 2009a), (Barrios, Aguado de Cea, and Ramos, 2009b), (Barrios and Vilches, 2010). The lists extracted containing verb/collocation paraphrases and false paraphrases can complement future research on paraphrasing.

Acknowledgements

We thank Socorro Bernardos who helped throughout the implementation process of BADELE.3000, Déborah Paton and Borja Menéndez for his help with English collocations.

References

Almela, R., M. Almela, P. Cantos, A. Sánchez, and R. Sarmiento. 2005. *Frecuencias del español. Diccionario y estudios léxicos y morfológicos*. Universitas, Madrid.

Androutsopoulos, I. and P. Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Apresjan, J., I. Boguslavsky, L. Iomdin, L. Leonid, L. Tsinman, A. Lazursky, V. Sannikov, and V. Sizov. 2003. Lexical functions as a tool of etap-3. In *First International Conference on Meaning-Text Theory*, pages 279–288, Paris, France. Lattice, CNRS.

Barrios, M. A. 2010. El dominio de la funciones léxicas en el marco de la Teoría Sentido-Texto. *Estudios de Lingüística del Español (ELiEs)*, 30.

Barrios, M. A., G. Aguado de Cea, and J. A. Ramos. 2009a. Badele: Applying MTT to specialized domain definitions in a knowledge-based tool. In M. C. L'Homme, editor, *8th International Conference on Terminology and Artificial Intelligence*, Toulouse, France. Université Toulouse.

Barrios, M. A., G. Aguado de Cea, and J. A. Ramos. 2009b. Enriching a lexicographic tool with domain definition problems and solutions. In G. Sierra, M. Pozzi, and J.M. Torres, editors, *1st International Workshop on Definition Extraction, RANLP 09*, Shoumen, Bulgaria. INCOMA Ltd.

Barrios, M. A. and L. Vilches. 2010. Is it possible to enrich ontologies with a specialized domain linguistic resource? In *Establishing and using ontologies as a basis for terminological and knowledge engineering resources*, Dublin, Ireland.

Bosque, I. 2004. *Redes. Diccionario combinatorio del español actual contemporáneo.* S. M., Madrid.

Bosque, I. 2006. *Diccionario combinatorio práctico del español actual contemporáneo.* S. M., Madrid.

de Miguel, E. 2006. Tensión y equilibrio semántico entre nombres y verbos: el reparto de la tarea de predicar. In *Actas del XXXV Simposio Internacional de la Sociedad de Lingüística Española*, pages 1289–1313, León, España. Universidad de León.

Kahane, S. 2001. Formal foundations of lexical functions. In *Actes du colloque collocation Computational Extraction*, pages 8–15, Tolouse, France.

Laureau, F. and L. Wanner. 2007. Towards a generic multilingual dependency grammar for text generation. In *In Proceedings of Grammatical Engineering Across Frameworks (GEAF)*, Stanford, CA.

Madnani, N. and B. J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Mel'čuk, I. 1992. *Dictionnaire Explicatif et Combinatoire du Français Contemporain. Recherches lexico sémantiques III.* Les Presses de l'Universitè de Montreal, Montreal.

Mel'čuk, I. 1996. Lexical functions: A tool for the description of lexical relations in a lexicon. In *Lexical functions in lexicography and natural language processing.* John Benjamin, Amsterdam, Philadelphia, pages 37–102.