Extraction et représentation des constructions à verbe support en espagnol

Sandra Milena Castellanos Páez
LIG – GETALP. Grenoble, France.
sandra.castellanos@imag.fr

Résumé. Le traitement informatique de constructions à verbe support (prendre une photo, faire une présentation) est une tâche difficile en TAL. Cela est également vrai en espagnol, où ces constructions sont fréquentes dans les textes, mais ne font pas souvent partie des lexiques exploitables par une machine. Notre objectif est d'extraire des constructions à verbe support à partir d'un très grand corpus de l'espagnol. Nous peaufinons un ensemble de motifs morpho-syntaxiques fondés sur un grand nombre de verbe support possibles. Ensuite, nous filtrons cette liste en utilisant des seuils et des mesures d'association. Bien que tout à fait classique, cette méthode permet l'extraction de nombreuses expressions de bonne qualité. À l'avenir, nous souhaitons étudier les représentations sémantiques de ces constructions dans des lexiques multilingues.

Abstract. The computational treatment of support verb constructions (take a picture, make a presentation) is a challenging task in NLP. This is also true in Spanish, where these constructions are frequent in texts, but not frequently included in machine-readable lexicons. Our goal is to extract support verb constructions from a very large corpus of Spanish. We fine-tune a set of morpho-syntactic patterns based on a large set of possible support verbs. Then, we filter this list using thresholds and association measures. While quite standard, this methodology allows the extraction of many good-quality expressions. As future work, we would like to investigate semantic representations for these constructions in multilingual lexicons.

Mots-clés : Expressions à verbe support, extraction, corpus, expressions polylexicales.
Keywords: Support verb expressions, extraction, corpus, multiword expressions.

1 Introduction

Les locuteurs natifs d'une langue ne se rendent pas compte que l'utilisation d'un certain mot provoque souvent l'utilisation d'un autre, et que ce processus permet de produire une expression correcte et naturelle. Après la lecture de la vaste littérature sur ce phénomène (Firth, 1957; Mel'čuk, 1981; Choueka, 1988; Smadja, 1993), nous adopterons la définition de Manning et Schutze (1999). Une « expression polylexicale » est une expression constituée de deux ou plusieurs mots qui correspondent à une façon conventionnelle de dire les choses et qui est caractérisée par une compositionnalité sémantique limitée, c'est-à-dire que le sens de l'expression ne peut pas être prédit à partir des sens des mots qui la composent.

Nous nous intéressons à une sous-classe spéciale de collocations, les *constructions à verbe support* (CVS)[1]. Les constructions de ce type correspondent à une structure linguistique formée par un verbe et un nom prédicatif. Les verbes *donner*, *faire* et *prendre* font partie de la liste des verbes support inclus dans la grammaire CVS[2] du français. Ils y sont intégrés en raison du peu de contenu sémantique qu'ils apportent à des expressions comme *faire un pas*, *donner une gifle* et *prendre la fuite*. Il y a peu d'éléments dans les sens des verbes *donner*, *faire* ou *prendre* qui nous indiquent la raison pour laquelle nous avons à dire *faire un pas* à la place de **donner un pas*.

Les CVS jouent un rôle important en ce qui concerne les applications concrètes telles que la traduction automatique, la recherche et l'extraction d'informations, les systèmes de questions-réponses, la synthèse, etc. (Laporte et al., 2008). Les CVS ont une fréquence élevée d'apparition dans la langue espagnole (Alvariño, 1999). Par conséquent, le choix du verbe pour un nom est complexe et donc présente des problèmes pour les apprenants de cette langue étrangère. Par exemple, la traduction automatique des CVS vers une autre langue peut donner comme résultat : (1) CVS → CVS ; (2) CVS → Verbe ; (3) Verbe → CVS (Zarco, 1997).

[1] La terminologie de ces verbes est variée : *light verb* (Jespersen, 1965), *funktionsverb* (Von Polenz, 1963), *predicado complejo* (Zarco, 1998).

[2] Cf. Liste des verbes support inclus dans la grammaire CVS (Laporte et al., 2008)

Les principaux objectifs de ce travail sont (1) estimer la présence et l'ubiquité des CVS dans les dictionnaires et les corpus, (2) appliquer des techniques d'extraction de CVS à partir de corpus. Pour des raisons liées à la facilité de l'évaluation des résultats par des locuteurs natifs, nous nous concentrons sur l'espagnol.

2 État de l'art

2.1 Constructions à verbe support

Les CVS sont des expressions lexicalisées, et plus précisément des expressions syntaxiquement flexibles (Sag et al., 2002). Il s'agit d'un syntagme verbal résultant, le plus souvent, d'une combinaison entre un verbe sémantiquement vide et un nom déverbal. Cette structure est soumise à une variabilité syntaxique complète et à un certain degré de compositionnalité sémantique.

1.	(a)	Kim	*gives advice*	to	first year students	2.	(a)	Paul	takes	a walk
	(b)	Kim	*donne un conseil*	aux	étudiants de première année		(b)	Paul	*prend	une promenade
	(c)	Kim	*da un consejo*	a los	estudiantes de primer año		>	*Paul fait une promenade*		
							(c)	Paul	*toma	un paseo
							>	*Paul da un paseo*		

L'exemple 1 décrit la même CVS en anglais, en français et en espagnol. Ici, c'est le verbe *donner* qui se comporte comme verbe support du nom *conseil*. Une traduction mot à mot semble pertinente pour ce cas précis, mais ce phénomène n'est pas toujours tout à fait présent. Dans la CVS de l'exemple 2, le sujet *Paul* peut *faire une promenade* (en français) ou *donner une promenade* (en espagnol) mais il ne pourra jamais la *prendre* comme c'est le cas en anglais.

D'autre part, alors que les CVS acceptent une variabilité syntaxique totale (exemple 3), elles présentent un degré de compositionnalité sémantique qui empêche la formation des CVS alternatives, comme le montre l'exemple 4.

3. John dio una explicación – John a donné une explication

(a)	Una explicación	fue dada	por John (*Passivation)*	
>	Une explication	a été donnée	par John	
(b)	¿Qué	tipo	de explicación	dio John? (*Extraction)*
>	Quelle	type	d'explication	John a donné ?
(c)	John	dio	una reveladora explicación (*Modification interne*)	
>	John	a donné	une explication révélatrice	

4.

(a)	Susan	{ofreció/*dio/*entregó/*regaló} disculpas	a	su padre
(b)	Susan	a {présenté/*montré/*exposé/*proposé} ses excuses	à	son père
>	Susan	s'est excusée	auprès de	son père

Dans ce dernier cas (4), même quand les verbes qui précèdent le nom *disculpas* (*ses excuses*) sont des synonymes dans d'autres contextes, là, seulement le verbe *ofrecer* (*présenter*) correspond à la combinaison correcte, qui permet de donner la signification de « présenter ses excuses à quelqu'un ».

2.2 Présence dans les lexiques

En ce qui concerne les dictionnaires[3], même quand nous pouvons estimer qu'il y a environ 25 000 CVS, il manque de l'information qui permet d'éliminer notamment les ambiguïtés possibles. Pour l'expression *tomar medidas* (*prendre des mesures*), nous pouvons distinguer plusieurs sens possibles, liés aux différents sens du mot mesure (une taille ou une action). Ajouter la fonction lexicale dans laquelle le nom prédicatif est la base de la collocation semble alors une solution intéressante

Cette solution fût d'ailleurs abordée par Mel'čuk (1996, 2003, 2004). Dans le cadre de la Théorie Sens-Texte, il introduit alors le concept de Fonctions Lexicales (FL). Une FL associe une unité lexicale **L** à une autre unité lexicale **L'** laquelle a une relation sémantique envers **L**. En ce qui nous concerne, la FL $Oper_i$ est la plus intéressante car elle est en charge d'associer un nom prédicatif avec un verbe sémantique vide.

Le tableau 1 contient une compilation de quelques CVS trouvés dans les dictionnaires[4]. Les verbes se situent dans la

[3] (DRAE, 2001 ; WordReference, 2008 ; Gran diccionario de la lengua española, 2005)

[4] (DRAE, 2001 ; WordReference, 2008 ; Gran diccionario de la lengua española, 2005)

première colonne. Les significations accompagnées du numéro assigné à chaque entrée (Gran diccionario de la lengua española, 2005) se trouvent dans la deuxième colonne. Et enfin, des exemples sont placés dans la troisième colonne.

Verbe	Signification	Exemples
Hacer	19 Réduire une chose à la signification du nom auquel elle est attachée	Hizo pedazos la carta
	20 Avec quelques noms, exprime l'action des verbes formés à partir de la racine de ces noms.	Le hizo burla
Dar	5 Effectuer l'action indiquée par le nom.	Nos dimos un abrazo
Tener	11 Avec des noms signifient le temps, exprime la durée ou l'âge.	Tiene seis años
Tomar	9 Suivi par certains noms déverbaux, il indique l'action de faire ce que le verbe (d'où les noms dérivent) exprime.	Tomar un baño Tomar una decisión.
	10 Recevoir ou acquérir ce que les noms qui accompagnent signifient.	Tomar fuerza Tomar aliento
Echar	32 Suivi par certains noms, il indique de faire l'action qui est signalé par ceux-ci.	Echarse una siesta.
	34 Suivi par des noms ou des expressions indiquant peine ou sanction, il s'agit de condamner une personne à celles-ci.	Le echaron 5 años de cárcel.

TABLEAU 1 Compilation de quelques CVS trouvés dans les dictionnaires

2.3 Extraction automatique à partir du corpus

Pour ce qui est de l'extraction automatique, divers travaux existent pour identifier les CVS. Il existe des approches qui peuvent soit tenir compte du contexte pour choisir si un candidat est une CVS ou non, soit extraire des paires verbe-objet et après l'utilisation de certaines méthodes indépendantes du contexte, faire son choix. Certaines méthodes sont statistiques, fondées sur la fréquence des mots, tandis que d'autres se fondent sur des règles linguistiques.

Lin (1998) utilise ainsi des techniques statistiques combinées à des traitements syntaxiques. Son travail consiste à collecter des triplets de dépendance, corriger leurs erreurs de comptage et enfin, les filtrer avec leur information mutuelle. Dans la même ligne de conduite, Stevenson et al. (2004), utilisent la même mesure pour détecter les collocations et capturer les propriétés linguistiques de la construction.

Par contraste, Vincze (2013) fait appel à l'utilisation des caractéristiques contextuelles et au modèle des champs aléatoires conditionnels (cf. Lafferty, 2001). Diab et Bhutada (2009) se servent d'un système supervisé pour classifier les combinaisons « verbe + nom » comme des expressions littérales ou idiomatiques, en fonction du contexte.

Finalement, Dias (2003) présente un système hybride qui permet d'extraire des candidats partir d'un corpus étiqueté avec des séquences de partie du discours. Dans son travail, il identifie automatiquement des patrons syntaxiques à partir du corpus, et ensuite, des statistiques de mots sont combinées avec de l'information linguistique pour extraire les séquences de mots les plus intéressantes.

3 Méthodologie

3.1 Constitution du corpus

L'extraction des expressions polylexicales demande le traitement des ressources de grande taille, notre choix a donc ainsi été porté sur l'imposant corpus émanant d'un projet intitulé « GrAF version of Spanish portions of Wikipedia Corpus » (Boleda et Vivaldi, 2012). Cette ressource comporte un corpus en espagnol de 257 019 articles provenant de Wikipédia, qui contiennent environ 150,1 millions de mots au format texte brut. Lors du projet, les auteurs ont nettoyé cette dernière par l'effacement des pages d'homonymie, la suppression de certaines étiquettes XML et l'homogénéisation des étiquettes de terminaison de listes. Ensuite, ils ont ajouté du marquage structurel (tête, paragraphe, phrase, liste, etc.) et de l'information morphosyntaxique.

Cependant, il a été nécessaire de faire du traitement et des transformations de format sur la ressource, en envisageant une utilisation ultérieure de l'outil mwetoolkit (qui sera introduit dans la section ci-dessous) sur le corpus résultant. Dans un premier temps, nous avons créé des scripts pour obtenir des indices de segments à partir du traitement des fichiers fournis. Ensuite, la lemmatisation a été corrigée pour certains segments. Finalement, le format a été transformé et adapté à celui que mwetoolkit reconnaît comme entrée.

À l'issue, nous obtenons un corpus en format XML d'un volume de 4 838 937 segments dont en moyenne 18 mots par segment (d'un total de 88 683 071 mots).

3.2 Extraction des CVS

Nous suivrons la méthodologie décrite dans la Figure 1. Elle est basée sur l'environnement d'extraction d'expressions polylexicales à partir du corpus de l'outil mwetoolkit (Ramisch, 2010).

Tout d'abord, l'outil reçoit comme entrée un corpus prétraité avec le format XML résultant du prétraitement. Ensuite, il emploie une méthodologie qui consiste en une phase d'extraction de candidats, suivie d'une phase de filtrage des candidats. Le système extrait les candidats, en utilisant des patrons morphosyntaxiques spécifiques. Une fois la liste de candidats extraite, il est possible de la filtrer avec des critères simples de seuil, ou des critères plus complexes, tels que leurs mesures d'association. Nous avons traité le corpus dans une phase précédente (sans utiliser l'outil mwetoolkit) pour lui donner le format correct. La mise au point des patrons morphosyntaxiques, employés dans la phase d'extraction des candidats, correspond au résultat de la combinaison entre certains patrons existant dans la littérature (De Miguel, 2008 ; Moncó, 2013) et d'autres extraits à partir de l'analyse d'un pourcentage du texte brut du corpus. La Figure 2 montre un exemple qui permet d'extraire les candidats à être des CVS (ci-après dénommés « candidats CVS ») de la forme « V: hacer »+ « DET : l'Art.Indef. »+ « NC », comme par exemple *hacer una pregunta, hacer una cita, hacer una reserva, hacer una confesión, hacer una llamada, hacer una pausa,* etc.

Dans la phase suivante, nous avons mis en œuvre un filtre heuristique pour garder seulement les candidats qui apparaissent plus de deux fois dans le corpus. La mise en place d'un filtre basé sur les mesures d'association est actuellement en cours.

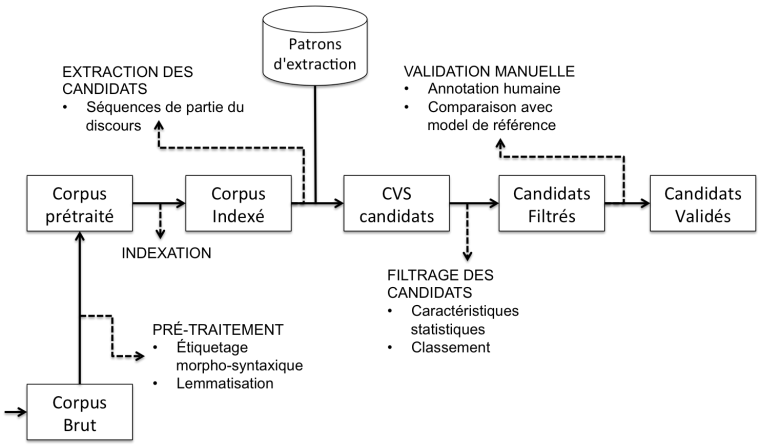

FIGURE 1 Méthodologie d'extraction et de validation des CVS

```
<pat>
        <w pos="V*" lemma="hacer"/>
        <!-- una -->
        <pat repeat="?"><w pos="T.FS*"/></pat>
        <w pos="NC*" />
</pat>
```

FIGURE 2 Patron de la forme V+DET +NC

4 Résultats

La méthodologie et les patrons ont été appliqués au corpus provenant de Wikipédia (Tableau 2). Le tableau 2 présente dans la première colonne la classification des verbes par ordre de pertinence. Les colonnes suivantes présentent la liste des verbes définis comme verbes support et les nombres d'extraction de candidats correspondant à la relation morphosyntaxique prédite par le patron. Ainsi, l'extraction génère un total de 81 274 candidats CVS donc il y a environ 1,7% des segments qui contiennent des candidats CVS. On tire de ce résultat douze verbes dont les plus représentatifs sont *tener* (*avoir*), *hacer* (*faire*) et *dar* (*donner*).

Rang	VS	# cand. à CVS	Rang	VS	# cand. à CVS
12	*Echar*	357	6	*Presentar*	6 220
11	Cometer	510	5	*Tomar*	5 286
10	*Guardar*	1 014	4	*Recibir*	6 135
9	*Sufrir*	2 823	3	*Dar*	11 272
8	*Perder*	2 905	2	*Hacer*	13 673
7	*Ofrecer*	3 603	1	*Tener*	27 476
				Total	**81 274**

TABLEAU 2 Verbes support du corpus

Le tableau 3 présente les 15 meilleurs candidats CVS du corpus, triées par mesure d'association. La première colonne est dédiée à une mesure qui tient compte des candidats composés seulement de 2 mots et les colonnes suivantes tiennent compte des candidats composés de n mots (n > 2). À travers ces trois mesures[5], on peut constater la fréquence élevée des verbes classifiés, nommés ci-dessus. Bien que les deux dernières colonnes contiennent les mêmes candidats, la combinaison avec les candidats extraits pour la mesure `ll` semble importante pour pouvoir amplifier la plage de couverture des candidats CVS. Finalement, on peut remarquer que la nominalisation de quelques candidats est aussi possible, par exemple, *hacer referencia* (faire une référence) peut être remplacé par *referenciar* (se référer), *tomar parte* (prendre part) par *participar* (participer), *dar nombre* (donner un nom) par *nombrar* (nommer), etc. Cette analyse nous permet de voir que l'extraction à partir de corpus est une voie prometteuse pour la constitution d'un lexique électronique de CVS de l'espagnol. La validation manuelle des candidats CVS est actuellement en cours et pourra confirmer cette hypothèse.

`ll`	`t_score`	`mle`
tener lugar	**tener lugar**	**tener lugar**
hacer referencia	**hacer referencia**	**hacer referencia**
dar cuenta	**dar cuenta**	**dar cuenta**
dar lugar	tener una superficie	tener una superficie
hacer cargo	tener un área	tener un área
tomar posesión	tener una población	tener una población
tomar parte	**recibir el nombre**	**recibir el nombre**
hacer prisionero	**dar lugar**	**dar lugar**
tener éxito	**tomar parte**	**tomar parte**
dar origen	**hacer cargo**	**hacer cargo**
dar nombre	tener una longitud	**dar nombre**
dar inicio	**dar nombre**	tener una longitud
tener constancia	**tener éxito**	**tener éxito**
tomar I	**dar origen**	tener forma
dar empereurs	tener forma	**dar origen**

TABLEAU 3 Top-15 candidats CVS triés par leurs mesures d'association (Candidats positifs en gras).

Même si, parmi les premiers candidats, il semble y avoir une proportion d'environ 69% de CVS correctement identifiés, nous voulons vérifier comment cette proportion évolue dans la suite de la liste.

5 Conclusions et travaux futurs

Nous avons montré comment on peut extraire des CVS en espagnol à partir d'un grand corpus. La prochaine étape de ce projet en cours est, dans un premier temps, la validation des candidats. Dans un deuxième temps, nous nous pencherons sur la modélisation de la variabilité des CVS pour représenter ces informations dans le lexique. Puis, dans un troisième temps, nous chercherons à obtenir des représentations sémantiques en utilisant soit des voisins distributionnels, soit des méthodes de paraphrase soit par l'identification du degré de figement ou de variabilité des CVS. Ensuite, dans un quatrième temps, il serait profitable d'évaluer et d'analyser le corpus parallèle pour tenter de trouver des représentations multilingues des CVS.

Références

ALVARIÑO P. (1999). *Sistematización léxico-sintáctica de los predicados complejos*. Tomás Jiménez Juliá, M. Carmen Losada Aldrey, José F. 505-510.

BOLEDA G., VIVALDI J. (2012). *GrAF version of Spanish portions of Wikipedia Corpus.* Universitat Politècnica de Catalunya. Research Group on Natural Language Processing; Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada (IULA). http://hdl.handle.net/10230/20047

CHOUEKA, Y. (1988). *Looking for needles in a haystack or locating interesting collocational expressions in large textual databases*. In RIAO'88, 609–624.

DE MIGUEL, E. (2008). Construcciones con verbos de apoyo en español. De cómo entran los nombres en la órbita de los verbos. En Actas del *XXXVII Simposio Internacional de la SEL.* [http://www.unav.es/linguis/simposiosel/actas/]

[5] `ll`: log likelihood; `t_score`: Student's t score; `mle`: Maximum likelihood estimator.

DIAB, M. AND BHUTADA, P. (2009). Verb noun construction MWE token supervised classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications.* 17-22.

DIAS, G. (2003). Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18 (MWE '03), Vol. 18.* 41-48.

FIRTH, J. R. (1957). *Papers in Linguistics* 1934-1951. Oxford, UK : Oxford UP. 233.

GRAN DICCIONARIO DE LA LENGUA ESPAÑOLA (2005). Barcelona, España: Larousse.

JESPERSEN, O. (1965). *A Modern English Grammar on Historical Principles*, Part VI, Morphology. London: George Allen and Unwin Ltd.

LAFFERTY J., MCCALLUM A., PEREIRA F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning.* 282- 289.

LAPORTE E., RANCHHOD E., AND YANNACOPOULOU A. (2008). *Syntactic variation of support verb constructions*. Lingvisticae Investigationes, 31(2):173–185. DOI: 10.1075/li.31.2.04lap.

LIN, D. (1998). Extracting collocations from text to corpora. In *Proceedings of the First Workshop on Computational Terminology.* 57-63.

MANNING C. AND SHUTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, USA : MIT Press. 620p.

MEL'ČUK, I. A. (1981). *Meaning-text models: a recent trend in Soviet linguistics*. The Annual Review of Anthropology.

MEL'ČUK, I. A. (1996). Lexical functions : A tool for the Description of Lexical Relations in the Lexicon. In : Wanner. Leo (ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/Philadelphia : Benjamins, 37-102.

MEL'ČUK, I. A. (2003). Collocations dans le dictionnaire. In : Thomas Szende (réd.), *Les écarts culturels dans les Dictionnaries bilingues*, Paris : Honoré Champion, 19-64.

MEL'ČUK, I. A. (2004). *Verbes suppports sans peine*. Lingvisticae Investigationes, 27 :2, 203-217.

MONCO S. (2013). Adquisición de las construcciones con el verbo «hacer», enfoque plurilingüe. *Revista Nebrija de Lingüística Aplicada* 13 (numéro spécial).

RAMISCH C., VILLAVICENCIO A., BOITET C. (2010). mwetoolkit: a Framework for Multiword Expression Identification. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valetta, Malta.

REAL ACADEMIA ESPAÑOLA. (2001). Diccionario de la lengua española (22.a ed.). [http://www.rae.es/rae.html]

SAG, I. A., BALDWIN, T., BOND, F., COPESTAKE, A., & FLICKINGER, D. (2002). Multiword expressions: A pain in the neck for NLP. *Proceedings of Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002*, Lecture Notes in Computer Science, 2276, 1-15.

SMADJA, F. A. (1993). *Retrieving collocations from text: Xtract.* Comp. Ling., 19(1):143– 177.

STEVENSON, S., FAZLY, A., AND NORTH, R. (2004). Statistical Measures of the Semi-Productivity of Light Verb Constructions. In *Second ACL Workshop on Multiword Expressions : Integrating Processing*. 1-8.

VINCZE V., NAGY I., ZSIBRITA J. (2013). Learning to detect english and hungarian light verb constructions. *ACM Trans. Speech Lang. Process.* 10, 2, Article 6 (June 2013), 25 pages.

VON POLENZ, P. (1963). *Funktionsverben im heutigen Deutsch.* Düsseldorf : Wirken-des Wort, Beiheft 5.

WORDREFERENCE.COM ONLINE LANGUAGE DICTIONARIES. (2008). Avalaible on : http://www.wordreference.com/

ZARCO TEJADA, Mª A. (1997). Codificación en el lexicón de las relaciones de concurrencia. *Philologia Hispalensis* 11, 83-93.

ZARCO TEJADA, Mª A. (1998). *Predicados complejos y Traducción automática.* Cádiz : Servicio de Publicaciones de la Universidad de Cádiz, 285.