Pragmatic annotation of a domain-restricted English-Spanish comparable corpus

Rosa Rabadán[1], Noelia Ramón[1], Hugo Sanjurjo-González[2]

[1]University of León (Spain), [2]University of Deusto (Spain)

This paper explores the multi-layer annotation of a written domain-restricted English-Spanish comparable corpus (CLANES – Controlled LANguage English Spanish), focusing on pragmatic annotation. The annotation scheme draws on part of speech tagging and a semantic annotation scheme, i.e. the UCREL Semantic Analysis System, with some added categories to fit the food-and-drink domain represented in CLANES. These are used to build significant (pragmatic) metapatterns. Seven different pragmatic functions have been identified in our corpus, namely <STATE>, <DIRECT>, <SUGGEST>, <RECOMMEND>, <PRAISE>, <EVIDENCE> and <RELATE TO READER>. Computer scripts translate this linguistic information into regular expressions to be used in unsupervised annotation. Partial results indicate that applying lexical restrictors boosts the success rate considerably. However, metadata is preferred because of increased replicability and generality. Replicability issues and limitations encountered during testing are also addressed.

Keywords: semantic annotation, pragmatic annotation, comparable corpus, regular expressions, English/Spanish

1. Introduction

Richly annotated corpora are essential for the retrieval of usable information in different applied environments. In bilingual corpora, multi-layered annotation becomes vital to carry out detailed contrastive studies and as a basis for applications in the ever-increasing hybrid, human-machine text production flows. Most bilingual corpora feature at least part-of-speech (PoS) annotation, and some include other types of lexico-grammatical annotation, e.g. cohesive devices in Kunz and Lapshinova-Koltunski (2018), or genre-specific multiword combinations in Pizarro Sánchez (2017), among others. However, semantically and pragmatically annotated bilingual corpora are still rare and very much in demand.

This paper explores the multi-layer annotation of the domain-restricted English-Spanish comparable CLANES corpus (Controlled LANguage English Spanish), focusing primarily on pragmatic annotation. CLANES includes over 1.5 million words in the two working languages distributed across six subcorpora corresponding to two different written genres: informational-promotional texts of gourmet foods and drinks, on the one hand, and instructive texts in the

CROSSING THE BORDERS: ANALYSING COMPLEX CONTRASTIVE DATA. Edited by Anna Čermáková, Signe Oksefjell Ebeling, Magnus Levin and Jenny Ström Herold. *BeLLS* Vol 11, No 1 (2021), DOI: 10.15845/bells.v11i1.3445.

form of culinary recipes, on the other. The aim is to automatically identify (semantic and) pragmatic meanings in such texts. As the initial tagging of the corpus was limited to PoS and rhetorical moves (Labrador *et al.*, 2014), an annotation scheme beyond the boundaries of the initial one needed to be developed. Thus, attempts at integrating semantic and pragmatic information had to be implemented to facilitate such analyses. Another limitation was the need to conduct unsupervised annotation at these levels on large amounts of texts.

In a traditional bottom-up approach to annotating textual material, we can find the first step in PoS tagging, namely assigning a particular grammatical category to each separate linguistic item. The next level in linguistic annotation, semantic annotation, will post a specific meaning to each item and multiword expression (MWE). And finally, pragmatic labels will refer to the speaker/writer's intentions when using the language for communicating with the intended audience, and contextual features play a role in the choice of linguistic elements. Considering the wide range of functions that may be performed employing language and all the extralinguistic factors involved, pragmatic annotation reveals itself as a complex task, more so if we try to carry it out automatically. As a result of these challenges, pragmatic annotation is still much less advanced than other linguistic annotation types.

Besides, most pragmatic studies are relatively small-scale qualitative analyses concentrating on spoken language data samples (Archer *et al.*, 2008: 613; Milá-García, 2018). Pragmatically annotated corpora of written texts are still rare but see Marín-Arrese's CESJD tagset (2017, 2019) or Weisser's in-progress TART dataset proposal (2018: 280ff).

A multi-layered linguistic analysis of CLANES may reveal a significant amount of relevant data for many applied purposes in the language industries, including teaching technical writing or developing semi-automatic applications for guided writing. With such applications in mind, the present paper describes the procedure followed for constructing a pragmatic annotation scheme to be applied to the CLANES corpus.

2. Data and method

2.1 Corpus description

This study's starting point is the CLANES corpus compiled at the University of León, Spain, in 2014-2019. It is a comparable corpus including 772,953 words in English and 776,100 in Spanish distributed across six subcorpora in each language. It comprises informational-promotional texts of gourmet foods and drinks and instructive texts in the same domain. The informational-promotional subcorpora include texts on wine (López Arroyo and Roberts, 2016), cheese (Labrador and Ramón, 2020), biscuits, herbal teas (Izquierdo and Pérez-Blanco, 2020) and dried meats (Ortego Antón, 2020). The instructive subcorpus is made up of culinary recipes (Rabadán *et al.*, 2016). All the texts were retrieved from online company web pages, open-access blogs and producer/retailer-facilitated materials. Table 1 shows the number of words per language in each subcorpus.

Table 1. Number of words in the CLANES corpus.

CLANES CORPUS		
Name of subcorpus	**Number of words English**	**Number of words Spanish**
RECIPES	290,498	257,184
CHEESE	128,347	139,017
WINE	117,874	140,694
BISCUITS	98,994	81,456
DRIED MEATS	85,419	42,161
HERBAL TEAS	51,821	115,588
TOTAL	**772,953**	**776,100**

2.2 Method and working procedure

Initially, the CLANES corpus was PoS tagged using TreeTagger (Schmid, 1995) and rhetorically annotated using an *ad hoc* tool, the ACTRES Tagger.[1] The long-term aim underlying the annotation project was to develop support for authors of promotional texts in the food and drinks industry (Labrador and Ramón, 2020). It soon became evident that the use of the annotated materials initially was limited to contrastive rhetoric and grammatical analyses and that attempts to go beyond these boundaries required higher-level semantic and pragmatic information (see this section and section 3 below).

The semantic annotation scheme employed to tag all the words in this corpus was based on USAS (UCREL Semantic Analysis System, Rayson *et al.*, 2004) developed at the Lancaster University from 2013, covering several different languages, including English and Spanish. The USAS scheme is based on the Longman Lexicon of Contemporary English (McArthur, 1986), composed of 21 major discourse fields and 232 labels, shown in Table 2 (based on Archer *et al.*, 2002).

Table 2. Twenty-one major discourse fields of the USAS scheme.

A – General and abstract terms	**B** – The body and the individual	**C** – Art and crafts
E - Emotion	**F** – Food and farming	**G** – Government and public
H – Architecture, housing and the home	**I** – Money and commerce in industry	**K** – Entertainment, sports and games
L – Life and living things	**M** – Movement, location, travel and transport	**N** – Numbers and measurement
O – Substances, materials, objects and equipment	**P** – Education	**Q** – Language and communication
S – Social actions, states and processes	**T** – Time	**W** – Word and environment
X - Psychological actions, states and processes	**Y** – Science and technology	**Z** – Names and grammar

Due to the general nature of the USAS categories, the semantic annotation had to be implemented manually by adding more specific subcategories from the F domain, which is highly relevant to the CLANES material, e.g. F1: Food has been expanded into F1.1, accounting for ‘variety/class of food x,’ such as *jamón ibérico* (< Iberico ham). F1.2 marks ‘meal organization’, i.e., when the food is typically eaten, as this is an important cross-cultural difference: *breakfast, dinner, snack*. F1.3 indicates ‘cuts,’ i.e., meat/ fish commercial cuts such as *chop, steak, fillet*, etc. The resulting semantic dataset includes over 5,000 domain-specific entries in both languages, plus an additional 10,000 general language entries in Spanish.

The amount of data spiked the need to conduct unsupervised annotation at these levels on larger amounts of text. Manual tagging was effected on a section of the corpus using first regular expressions and symbolic analysis. Then, a semantic word labelling tool (Sanjurjo-González, 2020) that includes different NLP (Natural Language Processing) processes together with word2vec (Mikolov *et al.*, 2013) and fastText (Bojanowski *et al.*, 2017) algorithms were used for unsupervised annotation. Current semantic annotation results show an overall degree of success of 89% in Spanish, including MWEs. For English, the success rate is close to 90%. These results refer to the food-and-drinks domain dataset in CLANES.

Pragmatic annotation starts from identifying combined PoS and semantic patterns that indicate one particular pragmatic function (see section 3 below). Our initial scheme includes

[1] Rhetorical move tagger® Available at http://contraste2.unileon.es/web/en/tagger.html. ACTRES stands for Contrastive Analysis and Translation English-Spanish in its Spanish acronym (Análisis Contrastivo y Traducción English-Spanish) https://actres.unileon.es/.

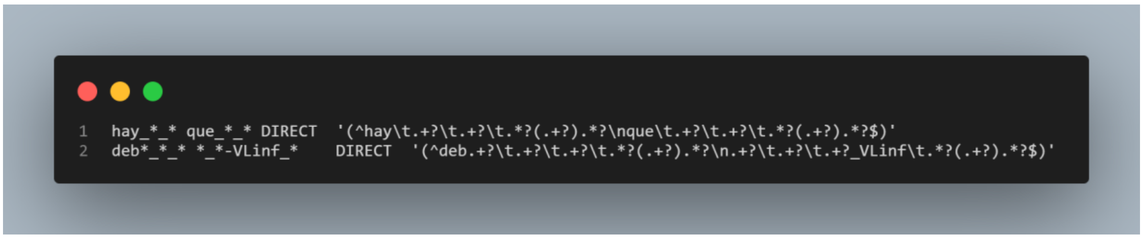
six categories, namely <STATE>, <DIRECT>, <SUGGEST>, <RECOMMEND>, <PRAISE> and <EVIDENCE>. An additional category, <RELATE TO READER>, was identified when testing replicability on popular science texts (see section 4 below).

<STATE> simply marks the delivery of referential information and applies to names of products, dishes, etc., as in *Buxton Blue*, *Hafner Vineyards 2009 Chardonnay*. <DIRECT> indicates an action to be carried out to fulfil a goal, as in *stir into batter* or *remove from oven*. <RECOMMEND> singles out the best course of action for the task at hand, as in *best eaten at room temperature*. <SUGGEST> signals that options offered may or may not be put into practice, as in *it can be enjoyed all year round* or *food pairing suggestions*. <PRAISE> refers to the product's good properties, as perceived intersubjectively, as in *perfectly balanced flavour combination*. <EVIDENCE> adds positive factual information about the product, as in *this cheese has won many awards*. <RELATE TO READER> promotes and marks the reader's involvement in the text, as in *I will save you, reader, the detailed account of …; but what does 'thermal equilibrium' really mean? I refer the reader to 1.15, where …*

Once the tagset had been defined, one of the first issues we had to address was segmentation. How could we decide where to set the boundaries of our pragmatic annotation scheme? Most previous attempts at (automatic) pragmatic annotation are applied to spoken data. However, the CLANES corpus contains written texts, although with very specific contextual settings: promotional and instructive texts from the food and drink industry. Bearing in mind that we were dealing with written texts, segmentation based on turns was not an option. It was decided to employ full stops and other punctuation marks indicating sentence boundaries as the 'pragmatic unit' to be tagged, as "all 'semantically complete' units, even if they consist of syntactic fragments (e.g. single noun phrases (NPs) that answer questions), should have a meaning and pragmatic function that is largely independent of the surrounding meanings and is thus also worth labelling individually" (Weisser, 2015: 89).

2.3 Developing the CLANES Annotation Scheme

The CLANES pragmatic annotation scheme uses the Python programming language and consists of two independent scripts that perform two primary tasks. The first script is responsible for converting the patterns into valid regular expressions using re-Python package syntax. Patterns are rule-like and are used to locate a particular combination of meta-items within sentence-based strings. The second script's role is to match those regular expressions with tokens of a specific, pragmatic function. Roughly, it works as follows: Texts are segmented into sentences using the NLTK sentence tokenizer (Bird *et al*., 2009). The script runs the regular expressions through the segmented units and checks whether and where a match can be found. If so, it applies the corresponding pragmatic tag (Figure 1).

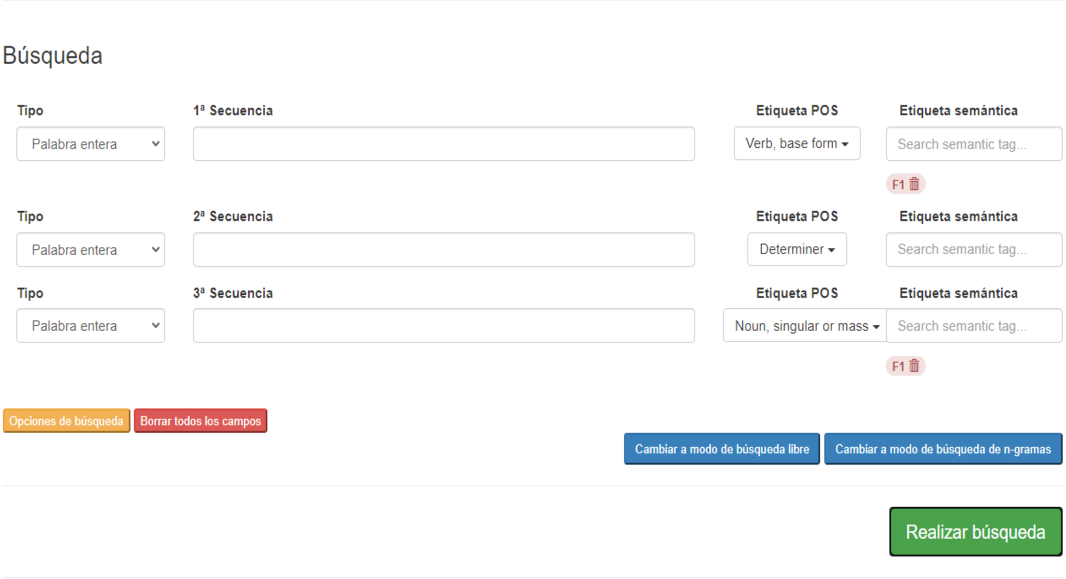


```
hay_*_* que_*_* DIRECT  '(^hay\t.+?\t.+?\t.*?(.+?).*?\nque\t.+?\t.+?\t.*?(.+?).*?$)'
deb*_*_* *_*-VLinf_*    DIRECT  '(^deb.+?\t.+?\t.+?\t.*?(.+?).*?\n.+?\t.+?\t.+?_VLinf\t.*?(.+?).*?$)'
```

Figure 1. Python regular expressions sample.

Pattern querying is done through the ACTRES Corpus Manager (Sanjurjo-González, 2017). This custom-made user platform allows for retrieving information for three layers of metadata, at present grammatical and semantic, and pragmatic (this last one still under construction).

Figure 2 shows the browser interface, where the searches can be carried out by lemmas or by PoS and semantic categories. More than one semantic category can be listed. In Figure 2, we have selected a search for any verb in the base form with the semantic category F1 food, followed by a determiner and followed by any noun in the singular with the semantic category F1 food.



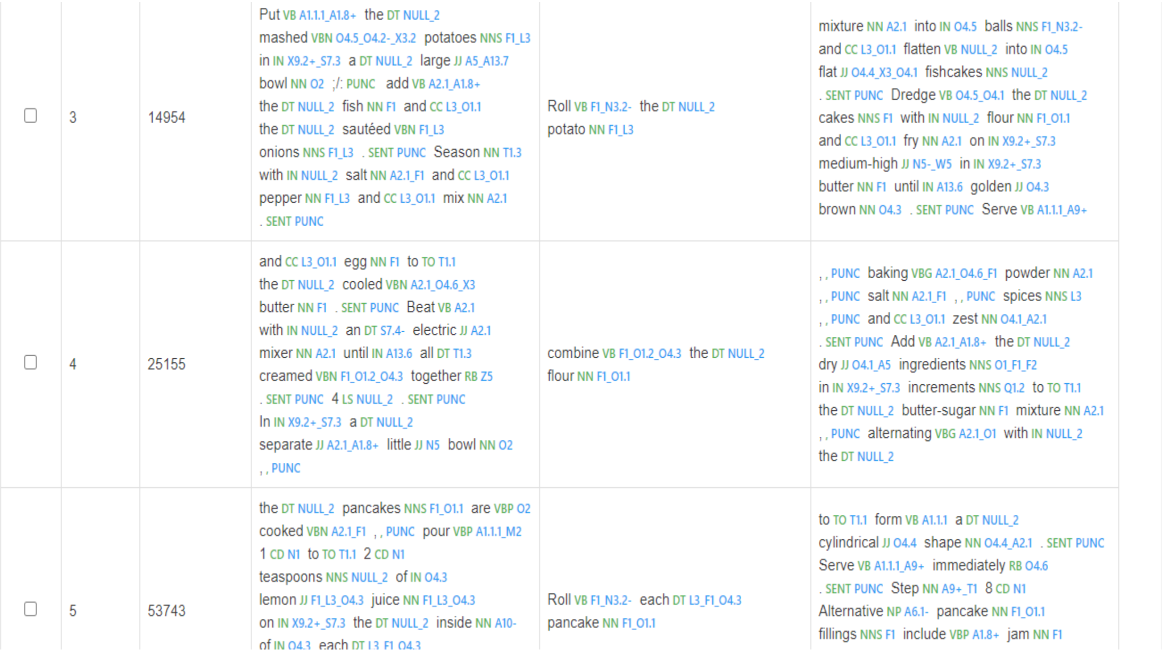
Búsqueda

Tipo	Secuencia	Etiqueta POS	Etiqueta semántica
Palabra entera	1ª Secuencia	Verb, base form	Search semantic tag... F1
Palabra entera	2ª Secuencia	Determiner	Search semantic tag...
Palabra entera	3ª Secuencia	Noun, singular or mass	Search semantic tag... F1

Opciones de búsqueda
Borrar todos los campos
Cambiar a modo de búsqueda libre
Cambiar a modo de búsqueda de n-gramas
Realizar búsqueda

Figure 2. Interface of the ACTRES Corpus Manager.

Figure 3 shows some of the hits for the query described in Figure 2 above: *roll the potato mixture*, *combine the flour*, *roll each pancake*, etc. We can see that all the concordance lines are displayed with all their PoS and semantic tags for further analysis.



☐	3	14954	Put VB A1.1.1_A1.8+ the DT NULL_2 mashed VBN O4.5_O4.2-_X3.2 potatoes NNS F1_L3 in IN X9.2+_S7.3 a DT NULL_2 large JJ A5_A13.7 bowl NN O2 ;/: PUNC add VB A2.1_A1.8+ the DT NULL_2 fish NN F1 and CC L3_O1.1 the DT NULL_2 sautéed VBN F1_L3 onions NNS F1_L3 . SENT PUNC Season NN T1.3 with IN NULL_2 salt NN A2.1_F1 and CC L3_O1.1 pepper NN F1_L3 and CC L3_O1.1 mix NN A2.1 . SENT PUNC	Roll VB F1_N3.2- the DT NULL_2 potato NN F1_L3	mixture NN A2.1 into IN O4.5 balls NNS F1_N3.2- and CC L3_O1.1 flatten VB NULL_2 into IN O4.5 flat JJ O4.4_X3_O4.1 fishcakes NNS NULL_2 . SENT PUNC Dredge VB O4.5_O4.1 the DT NULL_2 cakes NNS F1 with IN NULL_2 flour NN F1_O1.1 and CC L3_O1.1 fry NN A2.1 on IN X9.2+_S7.3 medium-high JJ N5-_W5 in IN X9.2+_S7.3 butter NN F1 until IN A13.6 golden JJ O4.3 brown NN O4.3 . SENT PUNC Serve VB A1.1.1_A9+
☐	4	25155	and CC L3_O1.1 egg NN F1 to TO T1.1 the DT NULL_2 cooled VBN A2.1_O4.6_X3 butter NN F1 . SENT PUNC Beat VB A2.1 with IN NULL_2 an DT S7.4- electric JJ A2.1 mixer NN A2.1 until IN A13.6 all DT T1.3 creamed VBN F1_O1.2_O4.3 together RB Z5 . SENT PUNC 4 LS NULL_2 . SENT PUNC In IN X9.2+_S7.3 a DT NULL_2 separate JJ A2.1_A1.8+ little JJ N5 bowl NN O2 , , PUNC	combine VB F1_O1.2_O4.3 the DT NULL_2 flour NN F1_O1.1	, , PUNC baking VBG A2.1_O4.6_F1 powder NN A2.1 , , PUNC salt NN A2.1_F1 , , PUNC spices NNS L3 , , PUNC and CC L3_O1.1 zest NN O4.1_A2.1 . SENT PUNC Add VB A2.1_A1.8+ the DT NULL_2 dry JJ O4.1_A5 ingredients NNS O1_F1_F2 in IN X9.2+_S7.3 increments NNS Q1.2 to TO T1.1 the DT NULL_2 butter-sugar NN F1 mixture NN A2.1 , , PUNC alternating VBG A2.1_O1 with IN NULL_2 the DT NULL_2
☐	5	53743	the DT NULL_2 pancakes NNS F1_O1.1 are VBP O2 cooked VBN A2.1_F1 , , PUNC pour VBP A1.1.1_M2 1 CD N1 to TO T1.1 2 CD N1 teaspoons NNS NULL_2 of IN O4.3 lemon JJ F1_L3_O4.3 juice NN F1_L3_O4.3 on IN X9.2+_S7.3 the DT NULL_2 inside NN A10- of IN O4.3 each DT L3_F1_O4.3	Roll VB F1_N3.2- each DT L3_F1_O4.3 pancake NN F1_O1.1	to TO T1.1 form VB A1.1.1 a DT NULL_2 cylindrical JJ O4.4 shape NN O4.4_A2.1 . SENT PUNC Serve VB A1.1.1_A9+ immediately RB O4.6 . SENT PUNC Step NN A9+_T1 8 CD N1 Alternative NP A6.1- pancake NN F1_O1.1 fillings NNS F1 include VBP A1.8+ jam NN F1

Figure 3. Sample hits of the ACTRES Corpus Manager.

3. Results: CLANES pragmatic annotation

This section describes the pragmatic patterns designed using combined strings of the PoS and semantic tags in the CLANES corpus of food and drinks, as illustrated in Figures 2 and 3. Preliminary results of unsupervised annotation testing are also included (section 3.6). The aim was to obtain prototypical patterns that could be used to identify each of the pragmatic functions described above. Computer scripts would be successfully applied to annotate the corpus with pragmatic tags.

3.1 Pragmatic function <DIRECT>

The pragmatic function <DIRECT> refers to an action to be carried out to fulfil a goal. To build the patterns that will lead to *regular expressions*, we identified both obligatory and optional pattern components combining PoS and semantic tags. Starting manually from PoS tags, it was found that the use of a verb in the imperative form (VB) was the most common mark of <DIRECT> in English. It was commonly followed by some noun (NN, NNS) in the field of food or drink (*add onions*), a field-related object (*remove the pan from the stove*) or a time expression with cardinal numbers (*simmer 6 to 7 minutes*). Optionally, determiners, prepositions, adverbs, adjectives or past participles may appear in between the obligatory items of verb and noun, as in *stir in salt and pepper; serve slightly chilled; enjoy with milk; serve on cheeseboard, etc.*

We listed all the semantic labels attached to the obligatory building blocks in this pattern to boost pattern identification further (see Table 2 for an overview of the semantic/discourse fields represented by A, F, O, etc.). Verbs: A1.1.1, general actions; A1.8, inclusion/exclusion; A2.1, affect: modify, change; A9, getting and giving: possession; A10, open/ closed, hiding/hidden, finding, showing; F1, food; F2, drinks; M2, putting, taking, etc.; O4.6, temperature, and X3, sensory; and nouns occurring after the verb: F1, food; F2, drinks; F4, farming and horticulture; L3, plants; O2, objects generally. The optional components have been shown to work better if only their PoS tags were considered determiners (DT), prepositions (IN), adjectives (JJ) or adverbs (RB), as well as cardinal numbers (CD) or past participles (VBN).

An example of a text chunk correctly tagged as <DIRECT> reads as follows:

(1) <DIRECT> *Stir* VB A1.1.1_M1 *in* INX 9.2 *salt* NNA2.1_F1 *and* CC *pepper* NNF1_L3 </DIRECT>.

Table 3 shows the <DIRECT> pattern in English with the obligatory components highlighted.

Table 3. <DIRECT> pattern in English.

PATTERN <DIRECT>					
OBLIGATORY		[OPTIONAL 1]	[OPTIONAL 2]	OBLIGATORY	
PoS	USAS			PoS	USAS
VB	A1.1.1	DT	CD	*NN*	F1
	A1.8	IN	JJ		F2
	A2.1	RB	VBN		F4
	A9				L3
	A10				T1.3
	F1				O2
	F2				

	M2				
	O4.6				
	X3				
	X3.3				

In English, additional restrictions need to be implemented in the search to distinguish imperatives from the infinitives, as they are not inflected, and both tagged as VB, verb base form. These restrictions include leaving out instances of VB preceded by an item labelled semantically as *Z5* (*grammatical bin*), or with PoS tags *pronoun PP*, or *noun NN*, as in (2) where 'you PP I2.2' signals that 'remove' is not an imperative.

(2) When WRB O4.4 you PP I2.2 remove VB A2.1_A1.8 the DT Z5 beans NNS F1_L3

In Spanish, however, these restrictions were not required, as the PoS annotation marks verbal inflections. Alternatives in the verbal slot are a *se*-passive (*se añaden la sal y la pimiento* (add (the) salt and pepper)) or a first-person plural present (*añadimos la sal y la pimienta* (we add (the) salt and pepper).

The same procedure was replicated for all other pragmatic patterns in English and Spanish to construct a pragmatic annotation scheme. The patterns have been translated into regular expression rules and subsequently used to carry out an unsupervised pragmatic annotation.

3.2 Pragmatic function <RECOMMEND>

The pragmatic function <RECOMMEND> indicates the best course of action to savour, prepare or present a food or drink item among those available and represented across all CLANES subcorpora. We have identified two typical patterns. Tables 4 and 5 show the <RECOMMEND> pattern in English with the obligatory and the optional components. The first (a) includes an adjective semantically tagged as A5.1+ Evaluation: good/bad; O4.2 Judgement of appearance; X3.1+ Sensory: taste, or X3.5 Sensory: smell followed by a preposition or, less frequently, by a condition marker, as in *excellent with grilled red meats*.

Table 4. <RECOMMEND a> pattern in English.

<RECOMMEND a> PATTERN			
OBLIGATORY		OBLIGATORY	
PoS	USAS	PoS	USAS
JJ	A5.1+	*IN*	Z5
	O4.2	IF	Z7
	X3.1+		
	X3.5		

The second pattern (b) features a superlative (RBS) followed by a preposition (IN) or an adverb (RB) is also possible in this function, as in *best at six months of ageing*. The adverb must semantically belong to X9.2 Ability: Success and failure or S7.3 Competition as *in best served slightly cool at 12-15ºC*. Optionally, this pattern may include a past participle (VBN) belonging to one of the following semantic categories: A1.1.1 General actions, making; A4.1 Generally kinds, groups, examples; E2 Liking; S3.1 Relationship: General, as in *best enjoyed with milk*.

Table 5. <RECOMMEND b> pattern in English.

<RECOMMEND b> PATTERN					
OBLIGATORY		[OPTIONAL]		OBLIGATORY	
PoS	USAS	PoS	USAS	PoS	USAS
RBS	A5.1	VBN	A1.1.1	*IN*	Z5
	A13.2		A4.1	*RB*	X9.2
	X3.1		E2		S7.3
	X3.5		S3.1		
			Z7		

Examples of text chunks tagged as <RECOMMEND> read as follows:

(3) <RECOMMEND> excellent A5.1_O4.2+ with Z5 grilled F1 red O4.3 meats F1 </RECOMMEND>

(4) <RECOMMEND> best RBS A13.2_A5.1+++ served VBN A1.1.1_A9 at IN Z5 room NN A1.1.1_A10 temperature NN O4.6 </RECOMMEND>.

In the case of Spanish, we find the same <RECOMMEND a> pattern: *ideal para ensaladas* (ideal in salads); *perfecto para tus picoteos* (perfect as a snack). Additionally, in Spanish, we also have a reflexive passive (A6.2, Q2, F1 and F2) followed by an infinitive or an adjective without a specific semantic profile, as in *se recomienda acompañar de un vino blanco Generoso* (we recommend pairing it with a Generous white wine).

3.3 Pragmatic function <SUGGEST>

The pragmatic function <SUGGEST> offers alternatives to carry out the task that may or may not be put into practice, and it appears across all subcorpora. We identified a primary pattern (Table 6) consisting of a pronoun (PP) or a noun (NN), whose meaning falls in the domain food (F1), drinks (F2) or L3 (plants), followed by a modal (MD) indicating possibility, a verbal base form and a past participle. The latter needs to be semantically tagged as A1.1.1 General actions, making; A2.1 Affect: Modify, change or E2 Liking, as in *it can be served hot* or *onions may be cooked in advance*.

Table 6. <SUGGEST a> pattern in English.

OBLIGATORY		OBLIGATORY		OBLIGATORY		OBLIGATORY	
PoS	USAS	PoS	USAS	PoS	USAS	PoS	USAS
PP	Z8	*MD*	A7	*VB*	A3	*VBN*	A1.1.1
NN	F1						A2.1
	F2						E2
	L3						

We were also able to single out a secondary pattern (Table 7) using an *–ing* form (VBG) of verbs meaning A1.1.1 general actions; A9 getting and giving: possession, or F1 food, combined with a noun (NNS) meaning Q2.1 Speech: Communicative or Q2.2 Speech acts, as in *serving suggestions* or *(food) pairing suggestions*.

Table 7. <SUGGEST b> pattern in English.

<SUGGEST b> PATTERN			
OBLIGATORY		OBLIGATORY	
PoS	USAS	PoS	USAS
VBG	A1.1.1	*NNS*	Q2.2
	A9-		Q1.1
	F1		

Examples of text chunks tagged as <SUGGEST> read as follows:

(5) <SUGGEST>mango NN L3_F1 can MD A7 be VB A3 replaced VBN A2.1 with IN Z5 sugar NN F1</SUGGEST>

(6) <SUGGEST> Food NN F1 pairing NN F1 suggestions NNS Q2.2 </SUGGEST>.

In Spanish, this pragmatic function's main pattern involves a modal verbal periphrasis with *se: se puede sustituir por leche de almendras* (it can be replaced with almond milk). An alternative is using the 1st person plural in the modal verb followed by an infinitive: *podemos utilizar jengibre en polvo* (we may use ginger powder). Additionally, we found a pattern similar to English <SUGGEST b>: a noun (Q2.2) in the plural optionally followed by a preposition: *Sugerencias de degustación* (serving suggestions); *maridaje* (food pairing).

3.4 Pragmatic function <PRAISE>

The pragmatic function of <PRAISE> refers to the product's good properties, as perceived intersubjectively. This function is widespread in the promotional texts of our corpus. In this case, we identified a pattern with three elements with the following PoS tags: one obligatory (a positive adjective JJ) and two optional elements: a pre-modifying adverb (RB) and a noun (NN-NNS) placed after the adjective. Both optional items may occur at the same time: *wonderfully creamy texture; truly lovely cheese*. At least one of the optional elements must co-occur with the adjective: either a pre-modifying adverb (*intensely fruity; absolutely delicious*) or the noun being pre-modified (*delicious milk; toasty aroma*).

The obligatory adjective in this pattern must belong to one of the following semantic categories: A.12 Easy/difficult; A5.1 Evaluation: good/bad; O4.2 Judgement of appearance; O4.3 Colour and colour patterns; O4.5 Texture; T2 Time: beginning and ending; T3 Time: old, new and young; age; X3.1 Sensory: taste; X3.3 Sensory: touch; X3.5 Sensory: smell. Moreover, the noun being modified must belong to one of the following semantic categories: A1.8 Inclusion/exclusion; A5.1 Evaluation: good/bad; F1 food; F2 drinks; X3.1 Sensory: taste; X3.5 Sensory: smell. The pre-modifying adverb may belong to any semantic category.

An example of a text chunk correctly tagged as <PRAISE> reads as follows:

(7) <PRAISE>wonderfully RB creamy JJ X3.1_O1.1 texture NN O4.5 </PRAISE>.

Table 8 shows the <PRAISE> pattern in English with the obligatory components highlighted.

Table 8. <PRAISE> pattern in English.

PATTERN <PRAISE>				
[OPTIONAL 1]	**OBLIGATORY**		[OPTIONAL 2]	USAS
	PoS	**USAS**		
RB	**JJ**	A12	NN	A1.8
		A5.1		A5.1
		O4.2		F1
		O4.3		F2
		O4.5		X3.1
		T2		X3.5
		T3		
		X3.1		
		X3.3		
		X3.5		

In Spanish, the pragmatic function of <PRAISE> is very similar and also pivots around an obligatory adjective with significantly positive semantic tags (A5.1: Evaluation: good/bad; X3.1: Sensory: taste; O4.2: Judgement of appearance), preceding or following a common noun (NC in the Spanish PoS notation system) with one of the following semantic tags: A5.1, F1, F2, X3.1 and X3.5. An example tagged for <PRAISE> reads as follows:

(8) <PRAISE>Es VSfin A3+_L1_Z5_X2.4 un ART Z5_N1_T3_T1.2_Z8 bizcocho NC F1 muy ADV A13.3 esponjoso ADJ O4.5_O4.1</PRAISE> (it is a fluffy, delicious sponge cake).

3.5 Pragmatic function <EVIDENCE>

The pragmatic function of <EVIDENCE> adds positive factual information about the product, e. g. comments about medals or awards won by the product, Protected Designation of Origin or other quality certifications in the food and drinks domain: *a gold medal winner at the World Cheese Awards*. Together with <PRAISE>, the pragmatic function of <EVIDENCE> is widespread in promotional discourse. The intended audience will be more inclined to buy a particular product if it has certified proof of quality.

In this case, we have noticed that several different patterns pivot around one single semantic label, namely S7.3: Competition, whether this semantic label is attached to an adjective (JJ): *an award-winning cheese*; to a noun (NN/NNS) such as *medal* or *winner*, or to a verb (VB), such as *award* or *win*, as in *this celebrated cheese has won many medals*; *this tea has won many awards*. The various PoS strings in which these items engage are typical unmarked syntactic patterns of English, such as adjective + noun or verb + determiner + noun. These common patterns are abundant and singled out as evidence by the sole presence of the semantic tag S7.3.

An example of the function <EVIDENCE> reads as follows:

(9) <EVIDENCE> A DT Z5 gold JJ O4.3 medal NN O2_S7.3 winner NN X9.2_S7.3 at IN Z5 the DT Z5 World NN W1 Cheese NN F1 Awards NN S7.3 </EVIDENCE>.

Similar patterns with the dominance of the S7.3 semantic category were observed in Spanish:

(10) <EVIDENCE> Ganador NC X9.2_S7.3 del PDEL Z5 premio NC S7.3 Cincho NP O2_A6.2 de PREP Z5 Oro NC O1 2006</EVIDENCE>. (Winner of the Cincho de Oro 2006 award).

We also tested these patterns on texts from different domains. In the case of popular science materials, we found the same pattern. However, the noun's semantics refers to someone in an authorial position: S7.2 Respect, P1 Education, X9.1 Ability, intelligence, as in *Professor at Princeton University; winner of the Nobel Prize.*

3.6 Other pragmatic functions: <RELATE TO READER> and <STATE>

The pragmatic function of <RELATE TO READER> signals the author's willingness to seek and maintain the reader's involvement. It is a phatic function, also found in academic lectures (Hyland 2005: 182–189), aiming to ensure the reader's engagement in text flow and progression (see section 2.2). Relating to the reader is done by addressing the reader directly, using rhetorical questions to mark advancement in presenting facts or concepts. The typical pattern consists of an interrogative such as *how* or a *wh*-pronoun/ adverb (WRB/WP) followed by an interrogative clause, which is uninformative in the context.

Another strategy, which can be combined with the one just mentioned, addresses the reader directly. An example of the function <RELATE TO READER> reads as follows:

(11) <RELATE TO READER> But CCB Z5 how WRB Z5 does DZ Z5_ A1.1.1 dark JJ W2 energy NN1 Y1_W1_X5.2+ work VB I3.1_A1.1.1? SENTPUNC </RELATE TO READER>.

A parallel pattern has been noted in Spanish, consisting of an interrogative clause featuring the standard word order required by Spanish syntax, i.e., *¿Qué <u>tienen que ver</u>, podría el lector preguntar, estas cuestiones de biología y de química con la uniformidad del universo primitivo*? (What, the reader might ask, <u>do</u> these questions of biology and chemistry <u>have to do</u> with the early universe's uniformity?).

The pragmatic function of <STATE> can be considered a default category that describes products or narrates sequences of actions and may adopt an almost unlimited combination of elements. This situation results in great difficulty in operationalizing patterns for this particular function. Any pragmatic segment not assigned to any of the other functions will be considered <STATE> as in *All the milk is unpasteurized.*

3.7 Preliminary testing results

The ACTRES pragmatic annotation scheme has been repeatedly tested at different stages of its development. Preliminary results showed a score of around 75% in Spanish and 62.5% in English. If taken by subcorpus, the informational-promotional subcorpora's success rate exceeded 70% in Spanish and was near 60% in English. For the instructive genre (recipes), the overall results hit 84% in Spanish and just below 65% in English. If taken by pragmatic category, in Spanish, the success rate ranged from 92% for <RECOMMEND> to 43.44% for <SUGGEST>. In English, the accuracy ranged from 88% for <STATE> to 5% for <SUGGEST>. These trials were all carried out using lexical (content word) restrictors in addition to PoS and semantic categories. For example, a sentence including the Spanish verbal periphrasis *hay que* (have to) would be identified automatically as having the pragmatic function <DIRECT>, or the noun *award* would prompt the tag <EVIDENCE>. Annotation testing exclusively using metadata (with no lexical restrictors) is currently underway.

4. Limitations and replicability

Results so far demonstrate that pragmatic tagging of written texts faces several challenges. Deciding on the unit for pragmatic annotation and the relevant segmentation of the text was one of them. Using punctuation marks such as full stops as sentence boundaries seemed to be the right choice initially. However, texts in our corpus contain promotional language, including titles, headings, and subheadings, most of which are not followed by any punctuation mark that could be used to define unit boundaries. As a result, some of our segments include more than one illocution, as headings tend to be grouped with the first sentence after the header. This means that the script only assigns one pragmatic function where manual annotation would assign two. Apart from manual revision, the possible solution is to consider other typographical features to discriminate segments adequately, e.g., paragraph indents.

Other limitations spring from mistagging in the PoS or semantic layers, which is misleading when identifying the patterns that form the basis for the regular expressions script. Minor but time-intensive manual corrections have been necessary at both levels to ensure that any minor mistake or null tag in a particular linguistic unit will not interfere with accurate pragmatic tagging. In our case, this problem has been an issue, as the PoS tagset makes use of different tags for English and Spanish, our two working languages. Both the searches and the pattern formulation are carried out using language-specific tags, which considerably slows down the process. Upgraded versions of the browser will try to achieve homogeneity in this respect.

Our tagset suffers from underspecification, particularly in the pragmatic function <STATE>, requiring a more detailed design to stop being the "default" function.

Replicability is central in annotation schemes. We have run informal pattern recognition tests in popular science (Rabadán and Gutiérrez-Lanza, 2020) and business texts (Pizarro 2017; Rabadán *et al.* in press). The goal was to check whether the patterns triggering the regular expressions hold in different domains and genres. Business texts revealed that <EVIDENCE> and <STATE> are typically found in reports; <RECOMMEND> was also found although marginally. Our test on popular science materials yielded a massive output of <STATE>, occasional but regular cases of <EVIDENCE>, and an additional pragmatic function that had not materialized in CLANES, but was added as a result of this test, <RELATE TO READER>.

5. Conclusions and further work

The long-term aim of setting up a pragmatic annotation scheme is to offer essential support to authors/communicators in the food and drinks industry. Previous attempts in this and other environments (Labrador and Ramón, 2020; Rabadán *et al.*, in press) showed the need for better, more informative corpora. Annotation has been a staple feature of written corpora for decades now but is still mainly confined to part-of-speech tagging. Although indisputably useful, additional information types become essential when facing tasks more sophisticated than basic grammatical contrast. Semantic annotation mostly follows USAS with some domain-specific additions. We aimed to set up a pragmatic annotation scheme based on previous PoS and semantic information. Using both layers, PoS and semantic, we identified prototypical patterns that have been used to characterize seven pragmatic functions. A computer script transformed the patterns into regular expressions, whose role is to detect tokens of a particular pattern within a sentence. Another script executes the unsupervised annotation using the regular expressions.

Our tests have shown that using lexical restrictors in the patterns boosts the success rate considerably. However, it detracts significantly from cross-linguistic replicability since sets of lexical restrictors would have to be changed according to language, genre, and domain. For example, verb forms tagged grammatically as infinitives will tend to be functioning pragmatically as <directive>, and semantic categories like A5.1+ (Evaluation: good) will always be tagged as <PRAISE> in any text type/domain, not only in promotional discourse in the food and drink industry. With nouns, however, semantic categories may need a different dataset according to particular domains, i.e., F1: Food nouns (e.g. salad, *ensalada*) would be replaced by Y1: Science and technology in a popular science corpus (e.g. spiral galaxy, *galaxia espiral*) or I2: Business in a business reports corpus (e.g. assets, *activos*).

The tests also suggest that pragmatic function frequency is linked to text type and overall text function rather than the domain. Results highlight the need for robust metapatterns rather than lexical items as pattern restrictors. They further suggest that adding an additional layer of annotation with more detailed information on grammatical functions would improve the usefulness of "supporting metadata" for pragmatic annotation. An example is verbal periphrases, which contribute meanings unrepresented in current PoS tags, for example, aspect types, such as inchoative in *poner a hervir* (start boiling) or continuative and gradual in *ir añadiendo* (roughly, keep adding) (Yllera, 1999: 3412–3420).

Replicability depends on the metapatterns underlying the regular expressions that allow the computer scripts to "extract rules" and apply them successfully to corpus annotation. So far, our pragmatic categories seem to work outside their home domain of food and drink.

Work in progress focuses on streamlining the regular expressions to improve script performance and success rate in pragmatic tagging and upgrading browser capabilities. Results will enable a wealth of studies and contribute to developing new applications, such as building a pre-editing workbench for bilingual text production of instructive and promotional genres in the food-and-drink domain. They will also improve existing author support tools (Rabadán *et al.*, in press) and be an essential component in designing a "drafter controlled language" for specific domains.[2]

Acknowledgements

This research has been funded by grant FFI2016- 75672-R awarded by the Spanish Ministry of Science and Innovation and ERDF (European Regional Development Fund).

References

Archer, D., Wilson, A and Rayson, P. 2002. Introduction to the USAS Category System. Retrieved from: http://ucrel.lancs.ac.uk/usas/usas_guide.pdf [Last accessed 21 March 2021].

Archer, D., Culpeper, J. and Davies, M. 2008. Pragmatic Annotation. In *Corpus Linguistics: An International Handbook*, M. Kytö and A. Lüdeling (eds), 613–642. Berlin: Mouton de Gruyter.

Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media http://www.nltk.org/book_1ed/ [Last accessed 26 November 2020].

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5: 135–146.

Hyland, K. 2005. Stance and Engagement: A Model for Interaction in Academic Discourse. *Discourse Studies* 6 (2): 173–191. DOI: 10.1177/1461445605050365.

[2] A "drafter controlled language" is a restricted interactive language that presents stub sentences and guides the user to gradually complete these sentences as required using the lexicogrammatical, semantic and pragmatic information on offer (Kuhn 2014: 138).

Izquierdo, M. and Pérez Blanco, M. 2020. A Multi-level Contrastive Analysis of Promotional Strategies in Specialised Discourse. *English for Specific Purposes* 58: 43-57. DOI: 10.1016/j.esp.2019.12.002.

Kuhn, T. 2014. A Survey and Classification of Controlled Natural Languages. *Computational Linguistics* 40(1): 121–170. DOI: 10.1162/COLI_a_00168.

Kunz, K. and Lapshinova-Koltunski, E. 2018. English vs. German from a Textual Perspective: Looking inside Chain Intersection. *Bergen Language and Linguistics Studies* 9(1): 21–42. DOI: 10.15845/bells.v9i1.1520.

Labrador, B., Ramón, N., Alaiz-Moretón, H. and Sanjurjo-González, H. 2014. Rhetorical Structure and Persuasive Language in the Subgenre of Online Advertisements. *English for Specific Purposes* 34: 38–47. DOI: 10.1016/j.esp.2013.10.002.

Labrador, B. and Ramón, N. 2020. Building a Second-language Writing Aid for Specific Purposes: Promotional Cheese Descriptions. *English for Specific Purpose*s 60: 40–52. DOI: 10.1016/j.esp.2020.03.003 9.

López Arroyo, B. and Roberts, R.P. 2016. Differences in Wine Tasting Notes in English and Spanish. *Babel* 62 (3): 370–401. DOI: 10.1075/babel.62.3.02lop.

Marín Arrese, J. 2017. Multifunctionality of Evidential Expressions in Discourse Domains and Genres. Evidence from Cross-linguistic Case Studies. In *Evidentiality Revisited: Cognitive Grammar, Functional and Discourse-Pragmatic Perspectives*, J. Marín Arrese, G. Hassler and M. Carretero (eds), 195–224. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/pbns.271.

Marín Arrese, J. 2019. CESJD-JMA Tagset for Annotation of Epistemic and Effective Stance Markers [Data set]. http://corpusnet.unileon.es/assets/uploads/tools/CESJD-TAGSET.pdf [Last accessed 1 June 2021].

McArthur, T. 1986. *Longman Lexicon of Contemporary English*. London: Longman.

Mikolov, T., Chen, K., Corrado, G., and Dean, J.. 2013. Efficient Estimation of Word Representations in Vector Space. In arXiv preprint arXiv:1301.3781v3. Ithaca: Cornell University.

Milá-García, A. 2018. Pragmatic Annotation for a Multi-Layered Analysis of Speech Acts: A Methodological Proposal. *Corpus Pragmatics* 2: 265–287.

Ortego Antón, M.T. 2020. Las fichas descriptivas de embutidos en español y en inglés: un análisis contrastivo de la estructura retórica basado en corpus. *Revista Signos* 53 (102): 170–194.

Pizarro Sánchez, I. 2017. A Corpus-based Analysis of Genre-specific Multiword Combinations. Minutes in English and Spanish. In *Cross-linguistic Correspondences: From Lexis to Genre*, T. Egan and H. Dirdal (eds), 221–252. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/slcs.191.09san.

Rabadán, R., Colwell, V. and Sanjurjo-González, H. 2016. BiTeXting Your Food: Helping the Gastro Industry Reach the Global Market. In CILC2016 (EPiC Series in *Language and Linguistics*, vol. 1), A. Moreno Ortiz and C. Pérez-Hernández (eds), 361–371. https://easychair.org/publications/open/Wv4r; https://doi.org/10.29007/4xtp [Last accessed 1 June 2021].

Rabadán, R. and Gutiérrez-Lanza, C.. 2020. Developing Awareness of Interference Errors in Translation: An English-Spanish Pilot Study in Popular Science and Audiovisual Transcripts. In *Specialised Languages and Multimedia. Linguistic and Cross-Cultural Issues*, E. Manca and F. Bianchi. Special issue of *Lingue e Linguaggi*, 40: 379–404. http://siba-ese.unisalento.it/index.php/linguelinguaggi [Last accessed 1 June 2021].

Rabadán, R., Pizarro, I. and Sanjurjo-González, H. In press. Authoring Support for Spanish language Writers: A Genre-restricted Case Study. *Revista Española de Lingüística Aplicada*, RESLA.

Rayson, P., Archer, D., Piao, S. and McEnery, T. 2004. The UCREL Semantic Analysis System. In *Proceedings of the Workshop Beyond Named Entity Recognition, Semantic Labelling NLP Tasks* (LREC 2004), 7–12. Lisbon: European Language Resources Association.

Sanjurjo-González, H. 2017. *Creación de un Framework para el tratamiento de corpus lingüísticos* [Development of a Framework for corpus linguistic analysis]. Doctoral dissertation, University of León, Spain.

Sanjurjo-González, H. 2020. Increasing Accuracy of a Semantic Word Labelling Tool Based on a Small Lexicon. In *Proceedings of the 17th International Conference on Natural Language Processing* (ICON-2020), Patna, India.

Schmid, H. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.

Weisser, M. 2015. Speech Act Annotation. In *Corpus Pragmatics: A Handbook*, K. Aijmer and C. Rühlemann (eds), 84–113. Cambridge: Cambridge University Press.

Weisser, Martin. 2018. *How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond*. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/scl.84.

Yllera, A. 1999. Las perífrasis verbales de gerundio y participio. In *Gramática descriptiva de la lengua española*, I. Bosque and V. Demonte (eds), 3391–3441. Madrid: Espasa.

Authors' addresses

Rosa Rabadán
Department of Modern Languages
Campus de Vegazana
University of León
ES-24071 León
Spain
rraba@unileon.es

Noelia Ramón
Department of Modern Languages
Campus de Vegazana
University of León
ES-24071 León
Spain
noelia.ramon@unileon.es

Hugo Sanjurjo-González
Department of Computing, Electronics and Communication Technologies
University of Deusto
Faculty of Engineering
Unibertsitate Etorbidea 24006
ES-48014 Bilbao
Spain
hugo.sanjurjo@deusto.es