Addressing article errors In Spanish Learners of English using a learner corpus

Fiorella Dotti and Mick O'Donnell
Universidad Autónoma de Madrid

Abstract: Two of the most common errors in the language of English learners involve the incorrect presence or absence of the article in noun phrases. This paper presents a corpus-based study of this phenomenon within the language of Spanish University learners of English. By exploring exactly which referential contexts cause the problem for our learners, we hope to present a clarified pedagogical model for teaching correct article production without requiring the teaching of all aspects of the complex area of article use. In regards to wrongful inclusion of articles, the study reveals that nearly all errors occur in cases where Spanish would include a definite article but English does not, and that two particular referential contexts are responsible for almost 90% of these errors: generic reference with a plural noun, and generic reference with a noncount noun. The remaining errors can be accounted for by three special contexts. The paper briefly looks at the other main article error type: where no article is produced but should be, and makes suggestions as to possible cause. Finally, some suggestions for teaching of this area are made.

Fiorella Dotti is a PhD student at the Universidad Autónoma de Madrid. Her research interests are computational and corpus linguistics, Second Language Acquisition. Her PhD thesis is on the automatic recognition of learner errors in written text.

Mick O'Donnell is a lecturer and researcher at the Universidad Autónoma de Madrid. He has worked in areas of corpus linguistics, computational linguistics and systemic functional linguistics, and is perhaps best known for his corpus annotation tools, including UAM CorpusTool, Systemic Coder and RSTTool. His current interests concern studying the grammatical development of learners of English using computational tools.

Introduction

Many of the foreign language teaching resources available to teachers have been developed for an international market, not targeting any particular mother-tongue group, but addressing the most common needs of the general learner community. As such, these resources are not perfectly tuned to the needs of learners of particular mother tongues, providing too much emphasis on issues which are not problematic for the mother tongue, and not enough emphasis on issues which are.

Examining a learner corpus from a particular mother-tongue community allows us to explore the problems faced by that particular community, discovering their most frequent errors, the kinds of structures they under-use, and those they over-use, in comparison with proficient speakers. Teaching can then be fined-tuned for the particular community.

The TREACLE project (O'Donnell, et al., 2009) has applied this methodology to Spanish University learners of English, studying a 700,000 word corpus of learner essays, both in terms of error annotation, and also automatic syntactic annotation. The English teaching curriculum in the partner universities has already been modified as a result of the study.

The singularly most frequent error identified by the study concerned the incorrect presence or absence of the article in noun phrases, with around 10% of all errors produced by the students falling into this area.

The current paper will explore this phenomena in more depth, showing why the students make the error, and in what particular referential contexts. The paper will also make suggestions as to how teaching of this phenomenon can be improved.

Determiner Errors

Two of the most common errors in English learner productions involve the incorrect presence or absence of the article in noun phrases. For instance,

- **Inappropriate inclusion of article:** *The drugs are substances who provocate dependence in the person who take it*
- **Inappropriate non-inclusion of article:** *the beginning of ( ) year*

These errors have been shown by many corpus studies to be very common, and not only for Spanish learners. Learners whose mother tongue lacks an article have particular problems, such as Chinese (Robertson, 2000), Japanese (Butler, 2002) and Russian. Spanish does have an article system, however, the rules for use of the definite article differ from those of English, resulting in the high degree of errors in this area.

These problems have been explored extensively in the past, both theoretically (Quirk and Greenbaum, 1973; Hawkins, 1978; Bickerton, 1981), and applied to ESL/EFL (e.g., Master, 1997, 2002; Pica, 1984; Huebner & Bickerton, 1983; Thomas, 1989). More recently, corpus-based quantitative studies have been carried out (Butler, 2002; Robertson, 2000). Some studies have explored the problem from the perspective of Spanish learners of English, in particular (Díez-Bedmar and Papp , 2008; Díez-Bedmar and Pérez-Paredes, 2012; García Mayo, 2008).

The Corpus

The TREACLE project annotated the errors in a 116,000 word corpus of essays written by Spanish University learners of English, identifying 16,000 errors, of which 1087 of these involved students producing an article where not appropriate for English (MacDonald et al., 2011). The original study made use of essays both from students in an English Studies degree (the WriCLE corpus: Rollinson & Mendikoetxea, 2010), and students in other degrees (e.g., Engineering) who were studying English (the UPV Learner Corpus: Andreu Andrés et al., 2010). Each essay is associated with a proficiency score provided by the Oxford Quick Placement Test (UCLES, 2001).

For the current study, we restricted ourselves to only the WriCLE component. The UPV component consists of far shorter texts, generally at a lower proficiency level, and as such, patterns are less clear using this corpus. This subcorpus consists of 78 essays, with 67,600 words, and 656 instances of incorrect absence or presence of the article. Note that this subcorpus lacks A1 learners.

For the more delicate study of contexts of use, reported later, we had insufficient time to code the contexts of use for each article error, and thus a further reduction of corpus size was used. This subcorpus consisted of 50 documents (44,500 words).

A Preliminary View of article errors in Spanish learners of English

The error coding in TREACLE was not very fine-grained in relation to article errors, tagging only whether the article was wrongly absent or present. However, this data does allow us to explore the evolution of the two error types with increasing proficiency. Overall, from our sample of 67,600 words, we observed article errors to the following degree:

Error Type	Count	Per 1000 Words
article-present-not-required	439	6.5
article-absent-required	217	3.2

Figure 1 shows the number of errors per 1000 words at each of the 5 proficiency levels in our corpus. The graph shows that learners do improve tremendously as they progress in proficiency, but that even upper advanced learners still make mistakes occasionally.

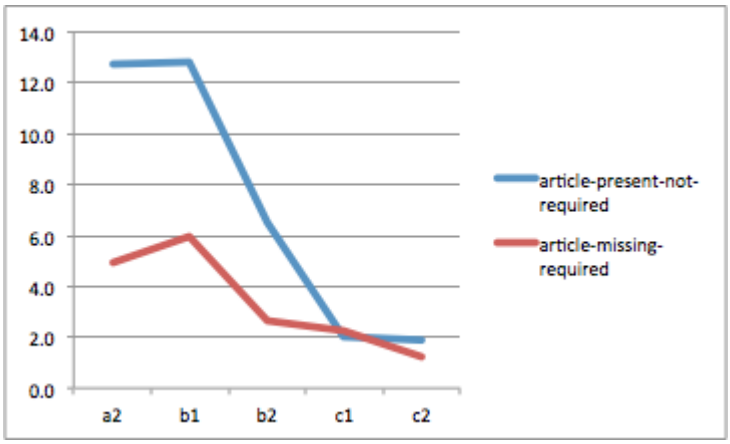

Figure 1: Falling rates of article-inclusion errors against rising proficiency (per 1000 words)

This graph hides the fact that the rules for determining if a definite article is necessary are quite complex. One goal of the present study is to provide a more detailed setting out of the referential contexts in which these noun phrases are produced, and to see if particular contexts are mastered earlier than others.

The other reason for performing this exploration is to identify which of these referential contexts are most problematic for Spanish learners of English, which will tell us where we need to focus our teaching effort, and at what levels of proficiency should students be presented with each of the contexts.

Referential Contexts determining Article use

Three referential factors seem to be adequate to explain article optionality: identifiable vs. not identifiability, generic vs. specific and count vs. noncount. These factors will be outlined below.

Identifiability (Definitiveness)

Quirk (1985) defined the usage rule for the definite article as follows:

> "The definite article *the* is used to mark the phrase it introduces as definite, *ie* as 'referring to something which can be identified uniquely in the contextual or general knowledge shared by speaker and hearer'". (p. 265).

Indefinite reference (referring to something which cannot be uniquely identified from context) on the other hand is marked by the use of "a" (singular nouns), or the zero article (indicated by 'Ø') or "some" (plural nouns).

Quirk uses the terms 'definite' and 'indefinite' to refer both to formal items (e.g., the definite article) and to a context of use (e.g., definite reference). To clarify, we will restrict our use of 'definite' to form: it can refer to the definite article ('the') or to a definite noun phrase (a noun phrase with a definite article). We will use the terms 'identifiable' and 'not identifiable' for the contextual use: whether or not the speaker intended the referenced entity to be taken as identified uniquely in textual or extra-textual context.

We note here that one of the main authorities in this area, Bickerton, uses the terms +Hearer-Knowledge (+HK) and -Hearer-Knowledge (-HK) for this distinction. Because we have found this terminology less clear than the identifiability terms, we will not use Bickerton's terms here.

Generic vs. Specific

A second important distinction in relation to the use of the determiner concerns generic vs specific reference.

- *Specific reference*: reference to a particular entity or group of entities, e.g., *These cats make too much noise.*

- *Generic reference*: reference to a class of entities, e.g., *Cats are night creatures.*

Generic reference can make use of 'the', 'a' or 'Ø': *The domestic cat is a small domestic animal. A cat is a nocturnal animal. Cats have poor color vision.* We note though that the use of 'the' in generic reference is limited to fairly formal definitions, and does not play much of a role in learner writing. So, with this exception, generic reference is realised as indefinite reference, using 'a' or the zero article.

Bickerton (1981) uses the terms +Speaker-Reference (+SR) and -Speaker-Reference (-SR) in place of specific and generic. We will stick to the more traditional terms for clarity.

Count vs. Noncount

The third relevant factor affecting determiner optionality concerns whether the noun is count or noncount. Count nouns can be singular ('a cat') or plural ('some cats'). Noncount nouns fall into two classes: mass nouns (e.g., 'water') or abstract nouns (e.g., 'terrorism').

Determining Article use

In summary, determiners are used in English as follows, according to the conditioning factors described above. We note that where 'the' is possible for specific reference, another specific determiner is possible (e.g., 'this cat', 'my cats', etc.):

Specific Reference	**Identifiable**	**Nonidentifiable**
	"the"	**"a" / some / Ø**
count:sing	the cat	a cat
count:plur	the cats	cats / some cats
noncount	the water, the love	water / some water love / some love

Generic Reference	*Definition use*: the: "the cat is…" *General case*:
count:sing	a cat
count:plur	cats / some cats
noncount	water / love

Here we follow Quirk and Greenbaum (1973) in applying the identifiable/nonidentifiable distinction only to specific reference (p. 68). They argue that since generic reference is used to denote what is normal or typical for members of a class, the speaker is not identifying specific entities.

Bickerton (1981) however does extend the identifiable/nonidentifiable distinction to his -SR category, producing two subgroups, generics (+HK), and nonreferentials (-HK). The last set would handle cases like "I need a new car" where the speaker is not referring to a specific item (thus -SR), and the hearer is not expected to identify the entity from context (-HK). This last set is realised identically to the specific nonidentifiables (+SR/-HK). Since our primary distinction is between generic and specific, and these nonreferentials are not generic, we place them along with the specific nonidentifiables. In this regard, the term 'nongeneric' would be more appropriate than 'specific', but we will stick to our original terms.

Differences between English and Spanish

Many learner errors result from differences between the mother tongue and the second language. Because we believe this to be particularly true in the case of article errors, we set out here the different rules conditioning article use for Spanish in contrast to English.

Table 2: English and Spanish Realisations of articles

Context of Reference	English	Spanish	Example
Specific: identifiable	the	el/la	the water
Specific:nonidentif:single	a/an	un/una	a dog
Specific:nonidentif:plural	some/Ø	unos/unas	some dogs/dogs
Specific:nonidentif:noncount	some/Ø	Ø	some water/water some doubt/doubt
Generic: singular	a/an	un/una	a cat
Generic: plural (i)	Ø	los/las	cats/los gatos
(ii)	some	unos/unas	some cats/unos gatos
Generic: non-countable	Ø	el/la	society/la sociedad

Those cases where there is distinct realisation in Spanish than English are marked in gray. In English, generic reference often does not use an article, and, except for limited use in definitions (e.g., 'the cat is...') does not use the definite article. Spanish on the other hand requires an article for these cases: in those contexts where english uses the zero article, Spanish uses the definite article. We would thus expect these two cases (generic plurals and generic noncounts) to be the major source of error in our learners.

In English, there are (at least) three exceptions to the requirement of an article for specific identifiable reference:

1. Places for their primary use: In English, specific reference to some places ('work', 'university', 'school', 'church', 'home', etc.) is given without the preceding article, e.g., "I am going home", "He went to church", etc. In Spanish, these would mostly be realised with the definite article, with the exception of "casa", e.g. *voy a casa* (*I am going home*).
2. Meal names: In English, specific reference to meals usually occur without the article, e.g., *See you after breakfast/lunch/dinner.* Spanish would realise these as with other specific references, with an article: *despues de la cena.*
3. Proportions: In English, referring to proportions of a whole is usually realised without an article, e.g., *20% of the respondents, half of my friends, most of your problems.* In Spanish, the article would appear: *el 20% de los encuestados, la mitad de mis amigos, la mayoría de sus problemas.* One exception to this exception is "the majority of".

Results from the Corpus

In this section, we present the results from the study of our corpus.

Article present not required

Our corpus included 440 cases of article-present-not-required. To see which of the referential contexts were most problematic for our learners, we extended the coding scheme to capture also the context of reference of each error, e.g., specific-identifiable, specific-non-identifiable, generic-singular, generic-plural and generic-noncount. To this we added also 3 tags for the 3 exception cases: *workplace-home-*

etc., *specific-meal* and *proportion*. Results were as shown in Table 3. 11 cases were tagged as 'uncodable' where the student writing was impenetrable).

Table 3: Referential contexts for article-present-not-required errors

Context of Reference	Instances	% of errors
Specific:identifiable	0	
Specific:nonidentifiable	0	
Generic:count:singular	0	
Generic:count:plural	82	31%
Generic:noncount	144	55%
workplace-home-etc	5	2%
specific-meal	0	
proportion	19	7%
uncodable	11	4%

All cases correspond to an interlingual difference in article use between English and Spanish: cases where Spanish would use an article, but English does not. Most noticeable here is that generic reference accounts for 86% of these errors, split between the plurals and the non-count type.

Of the three exception cases, the proportions type error (*the 20% of...*) was the most common, with 19 cases. There were no errors with meal references in our restricted corpus, although a word search did reveal 2 cases in the full corpus. The *workplace-home-etc*. category revealed 5 cases in our restricted corpus.

Looking at the nouns for which determiners are wrongly supplied, by far the most common are 'immigration', 'people' and 'society'. While the first is no doubt due to the high incidence of the topic in the essays, 'people' is more general, and any teaching material on article use would need to make special mention of items such as these.

We also explored whether the various referential contexts for article-missing error rose or fell in prominence with increasing proficiency. Figure 2 below shows the proportion of article-present errors due to each referential context at each proficiency level. While patterns are not clear, there is a suggestion that the noncount-generic context tends to be mastered by learners from C2 level, while the noncount-generic context continues to be a problem for these learners.

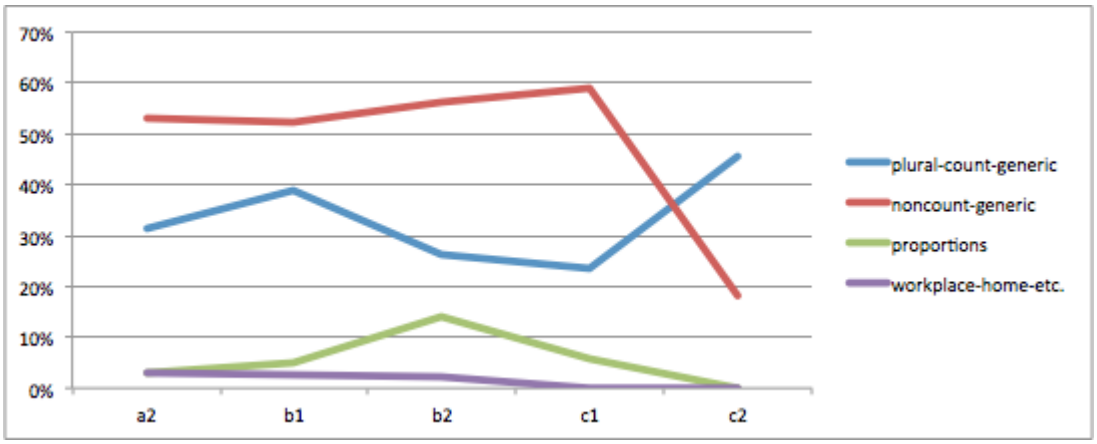

Figure 2: Proportions of referential contexts involved in article-wrongly-included errors with rising proficiency

Article absent required

Our efforts have focused on the article-wrongly-present error, and we have given less attention to the article-absent-required problem. However, we can make some comments. Firstly, this error makes less sense, since in all contexts where English requires an article, Spanish also will require an article.

One possibility is over-correction: the learner is corrected for transferring the Spanish article in generic reference, so wrongly applies their new-learnt rule to other cases as well. More study would be needed to verify this hypothesis.

Another (partial) explanation concerns cases where the learner may be transferring from Spanish involving a contracted preposition and article. Spanish will contract preposition and article in two cases: *del* (de+el), *al* (a + el). We hypothesise that in some cases, the learner translates the contraction as a single word, and in most cases, as the preposition. In a small sample of our corpus, in 29 out of 217 cases, the missing article is preceded by "of". In approximately half of these, the original Spanish would have used "del", e.g.,

- … of American president (del presidente americano)
- … people of same sex (gente del mismo sexo)
- … part of time (parte del tiempo)

However, the other cases would not have had a contracted preposition-article, e.g., *one of principle issues* (*una de las principales cuestiones*). So this is not the full answer for these cases, but may explain some of them.

Pedagogical Applications

Liu & Gleason (2002) conclude that:

> "Because of its high complexity and frequent use, the English article system … is one of the most difficult structural elements for ESL learners. In fact, it has often been considered hard grammar, very difficult if not impossible to teach".

They also cite a range of researchers who explore different approaches and techniques for teaching article usage, and for assessing the effectiveness of these techniques.

Our study focuses on Spanish learners of English. It has made clear that knowing when to produce an article or not is one of the most critical skills learners need to master. Most critical of the two errors, they need to learn when **not** to produce an article, and the learners need to be made aware of two main referential contexts where their Spanish practice should not be mapped onto English: the production of generic forms using plural and noncount nouns. For this, explicit teaching as to how to identify generic reference would be useful. In our own classes, we tell them to try placing "*en general*" (in general) after the noun phrase in Spanish, and if the sense is not changed, then the reference is generic.

The study also revealed three particular referential contexts which are problematic for Spanish learners of English, and where specific teaching materials could be provided:

- The production of proportions, such as '20% of...' or 'most of...'
- References to places for their primary use (e.g., *going to university*)
- References to mealtimes, such as *breakfast* and *lunch*.

The mass/count can also present problems for Spanish learners of English, as some words change from mass to count (or visa versa) in translation: e.g., *una información/some information*. However, the set of such words is small, and these words can be explicitly taught in class.

Conclusions

This study has aimed to study the major grammatical problem for Spanish learners of English, that of whether or not to include an article in a noun phrase. We found that the major problem is the overproduction of articles, and this occurs in referential contexts where Spanish would produce the article, but English would not. We identified the five referential contexts where this occurs: in all cases of generic reference involving plural or noncount nouns, and in three specific contexts: references to meals, references to places for their primary use, and references to proportions of a whole.

We explored whether these referential contexts are more or less problematic at each proficiency level, but did not find any clear pattern here.

A brief view on the other article problem, absence of an article where required, revealed no clear pattern. In general, Spanish will require an article whenever English does. One possible explanation is that the learner is over-correcting: they are corrected for including an article in generic references, so they wrongly extend the principle to other cases. We also noted one context which may lie behind some of these errors: translating a Spanish contraction of preposition plus pronoun (e.g., *del*) as simply the preposition.

Finally, we discussed how the findings of the paper could be incorporated into language classes, particularly that a means of teaching Spanish learners to distinguish generic and specific reference would be most valuable.

While the current study has focused on the learning of English by Spanish speakers, the methodology used here for using learner corpus study to inform language teaching could be applied to other mother-tongues as well, producing distinct pedagogical recommendations from those made here, but applicable to the learners of that mother tongue.

References

Andreu, M., Astor, A., Boquera, M., MacDonald, P., Montero, B., & Pérez, C. (2010). Analysing EFL learner output in the MiLC project: An error it's*, but which tag?. In M.C. Campoy, B. Belles-Fortuno & M.L. Gea-Valor (Eds.), *Corpus-Based Approaches to English Language Teaching* (pp. 167-179). London: Continuum.

Bickerton, D. (1981). *Roots of language*. Ann Arbor: Karoma.

Butler, Y. G. (2002). Second Language Learners' Theories on the Use of English Articles. *Studies in Second Language Acquisition*, *24*(03), 451–480.

Díez-Bedmar, M. B., & Papp, S. (2008). The use of the English article system by Chinese and Spanish learners. In G. Gilquin, S. Papp, & M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 147-195). Amsterdam: Rodopi.

Díez-Bedmar, M. B., & Pérez-Paredes, P. (2012). A cross-sectional analysis of the use of the English articles in Spanish learner writing. In Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research* (pp. 139-157). John Benjamins

García Mayo, M.P. (2008). The acquisition of four nongeneric uses of the article the by Spanish EFL learners. *System* 36, 550-565.

Hawkins, J.A. (1978). *Definiteness and Indefiniteness: A Study in Reference and Grammaticality Prediction*. London: Croom Helm & San Diego, CA: Academic Press.

Huebner, T., & Bickerton, D. (1983). *A longitudinal analysis of the acquisition of English*. Ann Arbor, MI: Karoma.

Liu, D., & Gleason, J. L. (2002). Acquisition of the article the by nonnative speakers of English. *Studies in Second Language Acquisition*, 24(01), 1–26.

MacDonald, P., Murcia, S., Boquera, M., Botella, A., Cardona, L., García, R. Mediero, E., O'Donnell, M., Robles, A., & Stuart, K. (2011). Error Coding in the TREACLE project. In M. L. Carrió Pastor, & M.A. Candel Mora (Eds.), *Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de córpora. Actas del III Congreso Internacional de Lingüística de Corpus* (pp. 725-740. Valencia: Universitat Politècnica de València.

Master, P. (1997). The English article system: Acquisition, function, and pedagogy. *System* 25: 2, 215-232.

Master, P. (2002). Information structure and English article pedagogy. *System* 30. 331-348.

O'Donnell, M., Murcia, S., García, R., Molina, C., Rollinson, P., MacDonald, P., Stuart, K., Boquera, M. (2009). Exploring the proficiency of English learners: The TREACLE project. Proceedings of the Fifth Corpus Linguistics, Liverpool.

Pica, T. (1984). Methods of morpheme quantifications: their effect on the interpretation of second language data. *Studies in Second Language Acquisition*, 6: 69-78.

Quirk, R. (1985). *A Comprehensive grammar of the English language*. London: Longman.

Quirk, R., & Greenbaum, S. (1973). *A university grammar of English*. London: Longman.

Robertson, D. (2000). Variability in the use of the English article system by Chinese learners of English. *Second Language Research* 16, 135-172.

Rollinson, P. & Mendikoetxea, A. (2010). Learner corpora and second language acquisition: Introducing WriCLE. In J. L. Bueno Alonso, D. Gonzáliz Álvarez, U. Kirsten Torrado, A. E. Martínez Insua, J. Pérez-Guerra, E. Rama Martínez & R. Rodríguez Vázquez (Eds.), *Analizar datos: Describir variación/Analysing data: Describing variation* (pp. 1-12). Vigo: Universidade de Vigo.

Thomas, M. (1989). The acquisition of English articles by first- and second- language learners. *Applied Psycholinguistics* 10, 335-355.

UCLES (2001). *Quick Placement Test (Paper and pencil version)*. Oxford: Oxford University Press.