Annotating article errors in Spanish learner texts: design and evaluation of an annotation scheme

María del Pilar Valverde Ibañez
Faculty of Foreign Languages
Aichi Prefectural University
1522-3 Ibaragasama, Nagakute-shi
Aichi, 480-1198, Japan
valverde@for.aichi-pu.ac.jp

Akira Ohtani
Faculty of Informatics
Osaka Gakuin University
2-36-1 Kishibe-minami, Suita-shi
Osaka, 564-8511, Japan
ohtani@ogu.ac.jp

Abstract

Annotating a corpus with error information is a challenging task. This paper describes the design, evaluation and refinement of an annotation scheme for Spanish article errors in learner data, so that future work on corpus annotation and automatic article error detection can progress. To evaluate reliability, 300 noun phrases with definite, indefinite and zero article have been tagged by four annotators. We analysed different types of disagreement, presented suggestions to increase reliability and applied the refined annotation scheme to create a gold-standard annotation.

1 Introduction

The annotation of learner texts with error information is necessary for linguistic research as well as for the development of language learning applications using natural language processing (NLP) techniques. While much efforts have concentrated on English, it is necessary to develop learner corpora and tools for other foreign languages like Spanish. This is the most commonly studied foreign language in the United States and the second most studied foreign language -after English- in many other countries. Overall, it is estimated that nearly 20 million people are studying Spanish as a foreign language (Instituto Cervantes, 2013). However, learner corpora and tools for this language are scarce (Lozano and Mendikoetxea, 2013; Nazar and Renau, 2012; del Pilar Valverde and Ohtani, 2012; Wanner et al., 2013). The goal of this paper is to define an annotation scheme that is suitable for reliable Spanish article error annotation, so that future work on corpus annotation and automatic article error detection can progress.

Automatic detection of errors has focused on function words such as articles (Izumi et al., 2004; Han et al., 2006; Felice and Pulman, 2008b; Gamon et al., 2008; Yi et al., 2008), prepositions (Felice and Pulman, 2008a) and particles (Dickinson, 2008; Oyama and Matsumoto, 2010). Function words are the most frequent words in any language, and they are also a very common source of mistakes for learners.

As for error annotation, one of the main difficulties is reliability. For some learner errors, like number and gender agreement, rules are clearly defined. Other kind of errors, like article or preposition presence and choice, are harder to annotate because native speakers differ widely with respect to what is acceptable usage. For article and noun number selection, for example, in Lee et al. (2009) raters found more than one valid construction for more than 18% of noun phrases.

To address this problem, we experiment with a preliminary annotation scheme for article errors, analyse the form disagreement among annotators takes, and refine the annotation scheme according to it. The paper is organized as follows. In section 2 we briefly summarize the linguistic properties of Spanish articles. In section 3 we explain an experiment carried out with a preliminary annotation scheme on article error annotation. In section 4 we examine the sources of disagreement among the annotators and in 5 we summarize the recommendations for reliable annotation. Section 6 presents the conclusions.

28th Pacific Asia Conference on Language, Information and Computation pages 234–243

2 Spanish articles

2.1 General overview

In Spanish, articles can be *definite* (as in English *the*) or *indefinite* (in English *a/an*), and their form changes according to the gender and number of the noun they complement, as shown in Table 1. [1]

	Definite		Indefinite	
	masc.	fem.	masc.	fem.
singular	*el*	*la*	*un*	*una*
plural	*los*	*las*	*unos*	*unas*

Table 1: Spanish articles

Article usage is complex because it is the result of the interaction of pragmatic, semantic, syntactic and lexical factors. Taxonomies of article use are abundant in the literature, targeted towards learners (Butt and Benjamin, 2014) or linguists (Bosque and Demonte, 1999; RAE, 2009). Basically, the main function of articles is to indicate the relationship between the nominal expressions and the entities to which the speakers refer by means of such expressions (Bosque and Demonte, 1999). For example, among other usages, we use the definite to generalize, that is, to refer to a whole class of things or people, as in (1) and to refer to something that is identifiable to the listener, as in (2). In (2), Maria's son must be identifiable for the listener either because a) Maria has only one son, or b) we have talked about him before. We use the indefinite to refer to any object of a particular class, as in (3), and we use no article when we are talking about an indefinite amount of something, as in (4) (examples from Alonso et al. (2013)).

(1) Los hijos dan muchos disgustos.
'Children cause a great deal of trouble.'

(2) El hijo de María tiene dos años.
'Maria's son is two years old.'

(3) Tener un hijo es lo mejor que te puede pasar en esta vida.
'Having a child is the best thing that can happen in life.'

(4) No tengo hijos pero tengo sobrinos.
'I do not have children but I have nephews.'

With regard to syntactic factors, for example two or more coordinated nouns should have their own article if they refer to different things: *un gato y un perro*, "a cat and dog" (*un gato y perro* suggests a cross between a cat and a dog) (Butt and Benjamin, 2014).

As for semantic factors, there are many rules which require specific knowledge. For example, place names usually have no article (*México*). For some of them the article is optional (*(el) Perú*) or depends on the context (*el México de los mexicanos*, "Mexicans' Mexico"), while the definite is obligatory for rivers, mountains, seas and oceans (*el Mediterráneo*). Other rules exist for numbers, proper nouns, names of languages, days of the week, etc.

Finally, there exist many set phrases and idioms which require definite (e.g. *con el objetivo de* 'with the objective of'), indefinite (*por una parte*, 'on the one hand') or zero article (e.g. *a corto plazo*, 'in the short run').

2.2 Difficulties for learners

Definite articles are the most frequent word in Spanish. In Davies (2005) frequency list the definite article is the most frequent *type* and the indefinite article is the 7th most frequent. In 9 billion words Spanish TenTen corpus (Jakubíček et al., 2013) the definite is also the most frequent type and the indefinite is the 6th. Approximately one out of every ten words in this corpus are articles.

Articles are also one of the most frequent grammatical errors, specially for speakers of languages that do not have articles like Chinese, Japanese, Korean or Russian. For speakers of Japanese, Fernández (1997) found 2.2 article errors per 100 words in a 4433 words sample.[2] In addition to that, this type of error diminishes as proficiency increases, but it tends to fossilize. The difficulty of the article system of Spanish may be comparable to English. McEnery et al. (2006) found that articles were the most difficult to acquire for Japanese learn-

[1] Spanish also has a definite article with neuter gender (*lo*), but its usage is quite different from the rest, so it will not be considered in this paper.

[2] The most frequent grammatical error in her sample concerns the verb (3.2 verb tense errors per 100 words), followed by prepositions (2.8 per 100 words) and articles.

ers of English, since even proficient learners had not achieved the acquisition rate of 90%. Therefore, we decided to use Japanese learners' texts to develop our annotation scheme.

3 Experiment

Annotation of learner errors is a challenging task for several reasons. First, learner sentences often contain interacting surrounding errors which can make the understanding of the meaning of the sentence quite difficult. Second, for some errors like number and gender agreement there are clear-cut rules about what is grammatical. But for other kind of errors, like article or preposition presence and choice, rules are usually not clearly defined, so in some cases more than one article choice may be acceptable. And third, in some cases more textual context or world knowledge may be needed to be able to determine the correct article usage.

As a result, inter-annotator agreement for error annotations can be relatively low. This issue has been put forward by the NLP community, that has found difficulties for evaluating error detection systems (Chodorow et al., 2012), but it has not received much attention in the learner corpus linguistic field. Several measures can be taken to address the varying number of corrections and levels of acceptability a sentence can have.

With regard to the number of possible analysis a sentence can receive, most error-annotated learner corpora permit only one tag per error. However, the "single correct construction" approach has been questioned and in recent annotation efforts there is a tendency to allow the inclusion of several alternative codes for the same item (Lüdeling et al., 2005; Boyd, 2010; Lee et al., 2012; Rozovskaya and Roth, 2010). However, it is unattainable to list all possible interpretations for every error, so this is done only "when there is doubt".

With regard to the level of confidence in the annotators' judgments, some projects include global measures of inter-annotator agreement (Rozovskaya and Roth, 2010; Lee et al., 2012) but annotated corpora do not explicitly provide confidence levels for every error. Only in some annotation experiments the annotators are asked to indicate their level of confidence for every item (as "low" or "high") (Tetreault and Chodorow, 2008).

We carry out an experiment on annotation of article errors with the following objectives:

1. Calculate inter-annotator agreement.
2. Analyse the types and sources of disagreement, to find out which are the main difficulties the annotators face when annotating article errors in learner texts.
3. Based on this experience, refine the guidelines and annotation scheme for error annotation.

3.1 Data collection

We used learners' texts written by 4th grade Japanese students of Spanish with an intermediate level of proficiency, at Aichi Prefectural University. A teacher of Spanish as a Foreign Language extracted sentences containing at least one article error from these texts, 50 sentences for each kind of article (definite, indefinite and zero article). The same number of sentences, but with at least one correct article usage, was then collected from the same texts. In every sentence only one highlighted noun phrase had to be annotated. The distribution of the resulting 300 sentences is as Table 2 shows.

	Definite	Indefinite	0 article	Total
Correct	50	50	50	150
Incorrect	50	50	50	150
Total	100	100	100	300

Table 2: Number of noun phrases and articles they contain

3.2 Preliminary annotation scheme

The 300 noun phrases were tagged by four annotators. The annotators were two experts (teachers of Spanish as a Foreign Language, who correct learners' texts on a regular basis), which we will call E1 and E2, and two non-experts (native speakers of Spanish with higher education but without experience in corpus annotation), which we will call NE1 and NE2.

They all annotated the same noun phrase in the same sentences, but presented in different orders, using a Microsoft Excel spreadsheet. Annotators were

provided with the target sentence plus the preceding and the following sentence, which they could resort to if they needed more context. If the target sentence was at the beginning or end of paragraph or text in the original text, no context was provided (a "beginning or end of paragraph or text" mark was inserted instead).

They were asked to classify the noun phrase into one of the categories shown in Table 3. We are only concerned with article presence and choice, so we did not tag malformation (e.g. spelling or agreement) and order errors.

Missing (definite)	AD
Missing (indefinite)	AI
Extraneous	E
Confusion	C
Article is correct	OK
Difficult to judge	NC

Table 3: Tags

Missing article (AD, AI) A missing error occurs when the learner does not use any article but the sentence should contain one: definite, as in (5) (AD|AD|AD|AD||AD)[3] or indefinite as in (6) (AI|NC|AI|AI||AI).

(5) Originalmente el español y el portugués son categorizados en mismo grupo lingüístico, la lengua románica.
'Originally Spanish and Portuguese are categorized in the same linguistic group, the romance language.'

(6) Osu está cerca del barrio de Sakae que es centro comercial muy animado y moderno.
'Osu is near Sakae area which is a very lively and modern commercial district.'

Extraneous article (E) An extraneous article error occurs when the article used by the learner is not necessary (zero article should be used instead), as in (7) (E|E|E|E||E).

(7) El objetivo de este trabajo es conocer cómo propagó el tomate como la verdura comestible desde el continente americano.
'The goal of this paper is to know how tomato spreaded as an edible vegetable from the American continent.'

Confusion error (C) A confusion error occurs when the learner used a definite article instead of an indefinite, or vice versa. In (8) (C|C|C|C||CA) the article should be definite because "victoria" refers to the last -unique and therefore identifiable- victory which ended the war.

(8) Franco consiguió una victoria en la Guerra Civil en 1939 y su dictadura comenzó.
'Franco pursued the victory in the Civil War in 1939 and his dictatorship began.'

Difficult to judge (NC) It was expected that the annotators would some times be unsure about the acceptability of article usage in a given sentence, or unable to determine the most likely correction.

We opted for allowing only one tag per sentence, but not forcing the annotators to mark the article usage as "right" or "wrong" and instead gave the possibility of marking sentences as "difficult to judge", as Han et al. (2006). We we wanted the annotators to correct the sentences only when they were sure about their decision, and not forcing them to make a best guess, which could lower inter-annotator agreement. Later we could look at the sentences marked as problematic, as (14), and analyse what they have in common.

4 Inter-annotator agreement

Tables 4 and 5 show the confusion matrices for expert and non-expert annotations. Observed agreement[4] is 0.79 for expert annotators and 0.76 for non-experts.

However, using observed agreement to measure reliability does not take into account agreement that is due to chance and hence is not a good measure of reliability. Therefore, an analysis using Cohen's Kappa statistic (Cohen, 1960) was performed. Perfect agreement would equate to a kappa of 1, and

[3] For every example from the learner data, in parenthesis we indicate the tags by the four annotators, in the following order: E1|E2|NE1|NE2||gold standard. For more details about the gold standard version, see section 5.

[4] Defined as the number of items on which annotators agree divided by the total number of items

E1:↓ E2: →	AD	AI	C	E	NC	OK	Tot
AD	**37**	0	0	0	2	2	41
AI	0	**5**	0	0	2	0	7
C	0	0	**30**	3	2	1	36
E	0	0	3	**39**	7	1	50
NC	1	0	1	4	**5**	8	19
OK	4	0	4	7	10	**122**	147
Total	42	5	38	53	28	134	300

Table 4: Confusion matrix for E1 and E2 annotators.

NE1:↓ NE2: →	AD	AI	C	E	NC	OK	Tot
AD	**31**	2	0	1	0	10	44
AI	2	**5**	0	0	0	2	9
C	1	0	**23**	2	2	6	34
E	0	0	4	**57**	2	10	73
NC	0	0	0	1	**0**	0	1
OK	5	1	5	7	2	**119**	139
Tot	39	8	32	68	6	147	300

Table 5: Confusion matrix for NE1 and NE2 annotators.

chance agreement would equate to 0. For the whole set of sentences (300, correct or incorrect), inter-annotator agreement for experts was found to be Kappa = 0.71 ($p < 0.001$), 95% CI (0.65, 0.77), and for non-experts it was 0.68 ($p < 0.001$), 95% CI (0.62, 0.75), which indicates substantial agreement. If we exclude the 45 sentences marked as"difficult to judge" by at least one annotator, kappa is 0.85 and 0.73 respectively. If we exclude 97 sentences tagged as "correct" by the four of them, remaining only sentences where at least one annotator considers there is an error, kappa is 0.62 and 0.58. If we exclude both sentences marked as NC by at least one annotator and sentences marked as OK by four annotators (remaining only 159 sentences) kappa is 0.79 and 0.61.

In the following sections we examine different types of disagreement: disagreement due to the annotators (4.1), due to the annotation scheme (4.2) and genuine disagreement (4.3), and propose some measures to reduce it.

4.1 Disagreement due to the annotators expertise: experts vs non-experts

The difference between experts and non-experts' reliability is due to the fact that non-experts make more slips than experts, and they are also less conservative when they correct texts.

In the data we find at least five mistakes (there can be more which we cannot detect), all by non-expert annotators: in four sentences they tag for a missing article a noun phrase which already contains one article, as (9) (C|C|AD|OK||OK), and in another one they tag for an extraneous article error a noun phrase without article.

(9) En Guatemala, la gente que tiene alta enseñanza piensa que "voseo" es una norma culta.
'In Guatemala, people who have higher education think that "voseo" is an educated norm.'

To prevent this kind of mistakes, any annotation project should automatically constrain the tags the annotators can use depending on the input (e.g. if there is already an article preceding a noun phrase, do not allow the "missing" error tag). Table 6 shows the error tags a noun phrase can receive depending on the article it contains.

Error tag	Definite	Indefinite	0 article
AD			x
AI			x
C	x	x	
E	x	x	

Table 6: Error tags a noun phrase can receive depending on the type of article it contains

In addition to that, even though non-experts are supposed to be less confident about the acceptability of sentences because pointing out errors in a text is a task for which they have no previous experience, in fact they are less cautious when they correct texts. For example, in (10) (OK|OK|E|E||OK)) experts consider the article is acceptable, while non-experts classify it as an extraneous article.

(10) Segundo,ahora ya no es imprescindible usar la coca para los objetivos antiguos,como para alivia de dolor o anestesia [...].
'Second, now it is no longer necessary to use the coca for the ancient purposes, like pain relieve or anaesthetic [...].'

This bias explains why, for example, NE1 uses the tag "difficult to judge" only one time (0.3%), while E2 uses it almost once every 10 sentences (9.3%), and non-experts use the tag "extraneous article" (specially for definite articles) more frequently than experts (23.5% vs. 12.2% of times).

Principle of minimal change Part of the variability on annotators' rigour could be reduced by giving clear guidelines about the optimum level of intervention in the texts. In this regard, we advocate for following a principle of minimal change: so we should not mark as errors the sentences where the learner choice is acceptable, even if the learner choice is not the best choice, that is, the goal of the annotator should be to produce an acceptable rather than a perfect result (e.g. Hana et al. (2010)), When the input is incomprehensible and the annotator cannot make a decision, it should be left without annotation.

In relation to that, annotators should be informed about the halo effect, by which the judgement of a sentence as acceptable or unacceptable is influenced by our overall impression of previous sentences. In other words, one is more likely to find errors in a text if this text already contains other errors. Experts (teachers of a foreign language) are trained on evaluation methods and they are aware of the importance of reliability in students' evaluation. They know how external factors (e.g. the halo effect and contrast effect) can have a negative impact and what can be done to reduce it. However, non-experts lack this training and are not aware of the challenges faced to perform a fair evaluation -annotation.

4.2 Disagreement due to the annotation scheme

We find some disagreements are due to the design of the preliminary annotation scheme, specially concerning the tags "difficult to judge" (NC) and "confusion error" (C).

The tag "difficult to judge" With regard to the reliability of the 6 tags used for annotation (Table 3), "difficult to judge" is the one that causes more disagreement: most of the times (67.7%) it is used by only one of the four annotators, and it is never used by three or four annotators in the same sentence. On the contrary, the rest of tags have a much higher agreement: on average, they are used by the four annotators 63.2% of the times, by three 19.9%, by two 9.2% and by one 7.7% of times.

Therefore, this tag should at most be used to filter out problematic sentences, which annotators cannot comprehend, and not for proper annotation of sentences.

We advocate for not using this tag and instead set clear principles in the annotation guidelines specifying what the annotators should do when they are not confident about the error analysis of a sentence.

The tag "confusion error" We found there was ambiguity in the guidelines about the meaning of this tag: in principle, it refers to the confusion between definite and indefinite articles but annotators also use it to indicate the confusion between an article and another type of determiner.

Indeed, learners frequently confuse the indefinite article with the indefinite determiner *alguno* 'some', when they refer to an indefinite amount of things, as in (11) (C|C|OK|OK||CD).

(11) Los hispanos están aumentando rápidamente y la población está centrada en unos estados.
'Hispanics are increasing rapidly and the population is concentrated in some states.'

To include this kind of error in the annotation, we should break down the tag into two: confusion between definite and indefinite article (CA) and confusion between article and another type of determiner (CD).

4.3 Genuine disagreement

As explained in section 2, article presence and choice can be determined by several factors. In our data, it mainly depends on pragmatic factors (69.0% of noun phrases), followed by lexico-semantic (20.7%) and syntactic factors (10.3%).

Leaving aside sentences tagged as acceptable by four annotators, agreement is higher when the article choice depends on lexico-semantic factors ($k = 0.835$ for experts and 0.780 for non-experts) and lower with pragmatic factors (($k = 0.514$ for experts and 0.496 for non-experts). Syntactic factors seem to be in between ($k = 0.750$ for experts and 0.523 for non-experts), although their low frequency

makes the figures less reliable. Therefore, more care should be paid to pragmatic distinctions.

Specifically, disagreement is more likely in noun phrases where two pragmatic interpretations (and article choices) are possible, and annotators choose one of the alternatives in an inconsistent manner (§ 4.3.1 and 4.3.2). Disagreement can also be due to a lack of the world knowledge that is needed to be able to determine the correct article usage (§ 4.3.3). As for syntactic and lexico-semantic factors (§ 4.3.4), disagreement occurs because annotators do not have a good knowledge about the existing prescriptive rules about article usage.

4.3.1 Definite article or zero article

Frequently both definite and zero article are acceptable for the same noun phrase. This happens when the noun phrase can refer to *a whole class of things or people in general* (definite article) or to *an indefinite amount of something* (zero article), as explained in 2. This distinction frequently does not change the meaning of the sentence significantly and in fact some languages with articles like English usually use the zero article in both cases.

When both pragmatic interpretations are possible for a given sentence, annotators unevenly choose one of them: some annotators tag the noun phrase for a missing article in (12) (OK|AD|AD|OK||OK) while they tag it for extraneous article in (13) (E |NC|OK|E||OK), even though in both sentences both the definite article and the zero article are acceptable, so the learner's choice should be left unchanged.

(12) Los políticos hablan en público y manifiestan sus opiniones con el objeto de conseguir votos de ciudadanos [...]
'Politicians talk in public and show their opinion with a view to get votes from the citizens [...].'

(13) Concretamente los cursos que consiguieron participantes japoneses y que ofrecen los certificados oficiales como IMEC(Instituto de Medicina China) continuarán existiendo [...].
'Specifically the courses which obtained Japanese participants and offer official certificates like IMEC (Chinese Medicine Institute) will continue existing [...].'

This distinction is specially problematic with plural nouns: in noun phrases with a plural nominal head, agreement by four annotators is less frequent (43.2%) than with singular nouns (66.7%) $\chi^2(2, N = 299) = 18.9, p < 0.001$. Therefore, more care should be paid in the annotation of plural nouns.

If the noun is singular and uncountable, we find the same ambiguous pragmatic distinction as with plural nouns, as in (14) (NC|NC|AD|E||OK), which is tagged as "difficult to judge" by some annotators and "extraneous" by others (the AD tag is a lapsus).

(14) El problema es demanda de la cocaína.
'The problem is demand of cocaine.'

In conclusion, according to the principle of minimal change, when both the definite and the zero article are acceptable, we should leave the learners' choice unchanged.

4.3.2 Indefinite article or zero article

Some times annotators agree in considering a noun phrase as unacceptable but they do not agree in the type of correction. This can happen when the learner wrongly uses a definite article, as in (15) (E|C|C|E||E/CA), and the annotators propose different corrections for it because the noun phrase can refer to *an indefinite amount of something* (zero article) or *any object of a particular class* (indefinite).

(15) En cambio, la cocaína tiene el efecto tóxico.
'On the contrary, cocaine has a toxic effect.'

Only in these cases, we allow adding two error tags (E/CA or E/CD) to the noun phrase.

4.3.3 World knowledge

In some sentences, annotators have insufficient extra-linguistic knowledge to be able to determine the right article usage. For example, in (16) (OK|E|E|E||OK) the annotator needs to know whether in Nagoya there are only nine interesting and touristy places (definite article) or there are more than nine (no article).

(16) Sale cada treinta minutos aproximadamente desde la estación de Nagoya y paran en los nueve sitios muy interesantes y turísticos, por ejemplo El castillo de Nagoya.
'It runs approximately every thirty minutes from Nagoya station and stops in nine very interesting and touristy places, for example Nagoya Castle.'

If the learner's choice is acceptable in some context, as in (16), we do not mark it as wrong. If the learner's choice is not acceptable, we tag the noun phrase as usual.

4.3.4 Syntactic and lexico-semantic rules

Unlike article usage governed by pragmatic factors, which is subject to interpretation by the annotator, for article usage determined by syntactic and lexico-semantic constraints there exist some linguistic norms about what is considered correct and incorrect.

However, native speakers -even experts- do not have a deep knowledge about these rules and some times do not follow them. For example, in (17) (AD|AD|OK|OK||OK) experts marked as error an article usage that is actually accepted, while non-experts tagged it right. It is the use of zero article between the preposition *a* ('to') and the relative pronoun *que* ('which') (RAE, 2006).

(17) [...] el capítulo 2 dice sobre el proceso del portuñol y los problemas a que el portuñol se enfrenta actualmente.
'[...] chapter 2 is about the portuñol process and the problems that the portuñol confronts nowadays.'

Therefore, to determine the acceptability of article usage, annotators should not rely only on their intuition as native speakers but also consult existing rules and recommendations published in reference dictionaries and grammars as RAE (2006) and RAE (2009).

5 Suggestions for reliable annotation

After examining the sources of disagreement in the annotation experiment, we added the following principles to the annotation scheme:

1. It is not recommended to use a tag like NC, "difficult to judge", because it has the lowest reliability. Therefore, we recommend simply not annotating the noun phrase if it impossible to determine the acceptability of the article usage. We did not find any case like that in our data from students with an intermediate level of Spanish.

2. Tags should inform us about the type of error *and* about the correction. This was true for the "add definite", "add indefinite" and "delete" tags, since we indicate which article we should add (definite or indefinite), and we know which article is deleted. The preliminary "confusion" error tag should be broken down into two tags to indicate confusion between definite and indefinite article (CA), and confusion between article and another type of determiner (CD).

3. Follow the principle of minimal change: the sentences should be acceptable rather than perfect. When more than one article choice including the learner's one is acceptable, we leave the learner's choice as correct. The pair definite article-zero article is the most interchangeable (in many sentences both are correct), so annotators should pay attention not to change the learner choice when it is correct.

4. When the learner choice is not acceptable and there are two equally good corrections, we allow double annotation. We found this mainly happens when the learner wrongly uses a definite or indefinite, and the annotators doubt between an extraneous error (zero article) and a confusion error. Only in this cases, we allow double annotation with E and CA or CD tags. There is usually no ambiguity in the appropriate correction for a missing article: annotators usually agree whether a definite or indefinite is necessary (probably for this reason the zero article has a high inter-annotator agreement.)

5. Regarding article usage governed by syntactic and lexico-semantic factors, base annotation not only on annotators' intuitions but first on the rules about article usage published by respected institutions (RAE, 2006; RAE, 2009).

6. When more world knowledge is needed to judge a sentence as correct or incorrect, we do not correct it if the learner's choice is acceptable in some context.

Following these criteria, we have revised the error tags given by the annotators for every sentence and made a decision about the most acceptable tag. The articles in the resulting gold standard set are distributed as Table 7 shows.

Tag	Definite	Indefinite	0 article	Total
AD	-	-	40	40
AI	-	-	6	6
CA	6	16	-	22
CD	0	7	-	7
E	36	18	-	54
E/CA	1	1	-	2
OK	57	58	54	169
Total	100	100	100	300

Table 7: Frequency of error tags in the gold standard per type of article (absolute frequency or %)

Despite the small size of the corpus study, some tendencies are observed in the 300 noun phrases written by Japanese learners:

1. The most frequent error regarding the definite article is extraneous use (83.7%): learners overuse it frequently probably because it is the most frequent article (and word) in Spanish.

2. When zero article is used, the most likely error is omission of the definite article (86.9%), for the same reason.

3. When learners use an indefinite article, the errors they commit are more evenly distributed. Confusion with a definite article or another type of determiner happens in 54.8% of cases and extraneous use in 42.9%.

6 Conclusions

Although article errors have been annotated in a number of small-scale studies, to date there has not been any study about article error annotation and inter-annotator agreement in Spanish learner texts. In this paper we have tested the results of an annotation scheme for article errors in a sample of learner texts written by Japanese learners. We have calculated agreement among four annotators (two experts and two non-experts) and have found kappa values between 0.85 and 0.62 for expert annotators and from 0.73 to 0.58 for non-experts, depending on the collection of sentences considered. The analysis of the disagreement among annotators has served us to find which are the main difficulties for annotators and to refine the annotation scheme according to it. Following more articulated guidelines we have revised the data to create a gold-standard.

The data used for the experiment is available to all interested researchers upon request. We hope the work presented here will facilitate future corpus annotation and development of automatic article error detection systems.

Acknowledgments

This research was partially supported by *kakenhi* (25770207 and 24500189), Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science.

We thank the students from Aichi Prefectural University who gave their permission to use their texts for this research.

References

Rosario Alonso, Alejandro Castañeda, Pablo Martínez, Lourdes Miguel, Jenaro Ortega, and José Ruiz. 2013. *Students' Basic Grammar of Spanish*. Difusion.

Ignacio Bosque and Violeta Demonte, editors. 1999. *Descriptive Grammar of Spanish Language*. Espasa Calpe, (In Spanish: Gramática descriptiva de la lengua española).

Adriane Boyd. 2010. EAGLE: an error-annotated corpus of beginning learner German. In *Proceedings of LREC-10*, Malta.

John Butt and Carmen Benjamin. 2014. *A New Reference Grammar of Modern Spanish*. Routledge.

Martin Chodorow, Markus Dickinson, Ross Israel, and Joel Tetreault. 2012. Problems in evaluating grammatical error detection systems. In *Proceedings of COLING 2012*, pages 611–628, Mumbai, Desember.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Mark Davies. 2005. *A Frequency Dictionary of Spanish: Core Vocabulary for Learners (CD)*. Routledge.

María del Pilar Valverde and Akira Ohtani. 2012. Automatic detection of gender and number agreement errors in Spanish texts written by Japanese learners. In *Proceedings of the 26th PACLIC*, pages 299–307.

Markus Dickinson. 2008. Korean particle error detection via probabilistic parsing. In *Automatic Analysis of Learner Language (AALL'08)*.

Rachele De Felice and Stephen G. Pulman. 2008a. Automatic detection of preposition errors in learner writing. In *Automatic Analysis of Learner Language (AALL'08)*.

Rachele De Felice and Stephen G. Pulman. 2008b. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the COLING 2008*, pages 169–176, Manchester, UK.

Soledad Fernández. 1997. *Interlanguage and Error Analysis in the Learning of Spanish as a Foreign Language*. Edelsa, (In Spanish: Interlengua y análisis de errores en el aprendizaje del español como lengua extranjera).

Michael Gamon, Jianfeng Gao, Chris Brockett, Alex Klementiev, William B. Dolan, Dimitry Belenko, and Lucy Vanderwende. 2008. Using contextual spell checker techniques and language modelling for ESL error correction. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 449–456, Hyderabad, India.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(2):115–129.

Jirka Hana, Alexandr Rosen, Sva, and Barbora Štindlová. 2010. Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop (ACL 2010)*, pages 11–19, Uppsala, Sweden, July.

IC Instituto Cervantes. 2013. *Spanish: a Living Language. 2013 Report*. Instituto Cervantes, (In Spanish: El español: una lengua viva. Informe 2013).

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. SST speech corpus of Japanese learners' English and automatic detection of learners' errors. *International Computer Archive of Modern English Journal*, 28:31–48.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen corpus family. In *7th International Corpus Linguistics Conference*.

John Lee, Joel Tetreault, and Martin Chodorow. 2009. Human evaluation of article and noun number usage: Influences of context and construction variability. In *Proceedings of the Third Linguistic Annotation Workshop (LAW)*, pages 60–63, Suntec, Singapore.

Sun-Hee Lee, Markus Dickinson, and Ross Israel. 2012. Developing learner corpus annotation for korean particle errors. In *Proceedings of the Sixth Linguistic Annotation Workshop, LAW VI*, pages 129–133, Stroudsburg.

Cristóbal Lozano and Amaya Mendikoetxea. 2013. Learner corpora and second language acquisition: the design and collection of CEDEL2. In Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson, editors, *Automatic Treatment and Analysis of Learner Corpus Data*. John Benjamins, Amsterdam.

Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. In *Proceedings of the Corpus Linguistics 2005 Conference*, Birmingham, United Kingdom, July.

Tony McEnery, Richard Xiao, and Yukio Tono. 2006. L2 acquisition of grammatical morphemes. In *Corpus-based language studies. An advanced resource book*. Routledge.

Rogelio Nazar and Irene Renau. 2012. Google books n-gram corpus used as a grammar checker. In *EACL 2012 Proceedings of the Second Workshop on Computational Linguistics and Writing (CLW 2012)*, pages 27–34.

Hiromi Oyama and Yuji Matsumoto. 2010. Automatic error detection method for Japanese case particles in Japanese language learners' writing. In *Corpus, ICT, and Language Education*, pages 235–245.

Real Academia de la Lengua Española RAE. 2006. *Diccionario panhispánico de dudas*. Real Santillana.

Real Academia de la Lengua Española RAE. 2009. *New Grammar of Spanish Language (In Spanish: Nueva gramática de la lengua española)*. Espasa Calpe.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL errors: Challenges and rewards. In *Proceedings of NAACL'10 Workshop on Innovative Use of NLP for Building Educational Applications*. University of Illinois at Urbana–Champ.

Joel Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of the Workshop on Human Judgments in Computational Linguistics at the COLING 2008*, pages 24–32.

Leo Wanner, Serge Verlinde, and Margarita Alonso. 2013. Writing assistants and automatic lexical error correction. In *Proceedings of the eLex 2013 conference*, pages 472–487.

Xing Yi, Jianfeng Gao, and William B. Dolan. 2008. A web-based English proofing system for English as a second language users. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 619–624, Hyderabad, India.