Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

Challenges in the annotation of article errors in Spanish learner texts

María del Pilar VALVERDE IBAÑEZ
[a]*Department of European Studies, Aichi Prefectural University, Japan*
valverde.mp@gmail.com

Abstract: Annotating learner texts with article error information is a difficult task. To identify which are the main difficulties for annotators, we carry out an annotation experiment in Spanish texts written by Japanese learners. Two expert and two non-expert raters annotate 300 noun phrases containing a definite, indefinite or zero article. We calculate inter-annotator agreement and analyse the sources of disagreement. We find article usage governed by pragmatic factors causes disagreement the most, while lexico-semantic factors are the most reliable. Finally, a learner corpus sample of 30,000 words is annotated with a revised version of the annotation scheme.

Keywords: articles, error annotation, learner Spanish, inter-annotator agreement, reliability, corpus annotation

1. Introduction

The annotation of learner texts with error information is necessary for linguistic research as well as for the development of educational applications for language learning.While research has focused on the development of learner corpora and tools for English as a foreign language, the field of Grammatical Error Detection (GEC) is expanding and there is a need to develop resources for other languages. However, learner corpora annotation and evaluation best practices are still an open issue.

First, inter-rater reliability for error annotation can vary widely: while for some errors, like number and gender agreement, rules are clearly defined, and using one rater may be acceptable, other kind of errors like article or preposition presence and choice are harder to annotate (Tetreault et al. 2010), so using only one annotator is not enough reliable. 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. For prepositions, Tetreault and Chodorow (2008) found that even in native texts, "native raters can disagree with each other by 25% in the task of preposition selection". In spite of this, learner corpora are typically annotated only once because double annotation would be too expensive, and few annotation projects provide measures of inter-annotator agreement (Rozovskaya and Roth, 2010; Lee et al., 2012). The need to improve annotation quality has been put forward by the NLP community, that has found difficulties for evaluating error detection systems in the last GEC shared tasks: in the first Helping Our Own task (2011), systems were penalized for valid corrections not annotated in the data while in the last three tasks (H00 2012, CoNLL 2013 and 2014) teams could request the organizers to make changes in the annotation (Tetreault et al. 2014).

Second, as noted by Reidsma and Carletta (2008), there are different types of disagreement: chance disagreement, caused by random slips or lack of knowledge of the annotators, and systematic disagreement, due to different intuitions of the annotators or to a misinterpretation of the annotation guidelines. This distinction is crucial for the development of gold standards, since systematic disagreement can have a worse effect on machine learning than noise-like disagreement.

In this scenario, how can we improve the quality of annotations? Multiple annotations by more than one trained annotator is unrealistic for large projects and while crowdsourcing has shown good results for English preposition error annotation in a pilot study (Tetreault et al. 2010), more research is needed to

deal with other languages and error types. To date, there has not been found a definitive method to improve the quality and number of annotations for learner texts

The goal of this paper is to investigate the main difficulties faced in the annotation of article errors in Spanish learner texts, so that measures can be taken to improve the quality of future annotation efforts. To do that, we carry out an experiment on article error annotation with a preliminary annotation scheme. In section 2 we describe the annotation principles, in 3 we briefly describe the data collection and annotation procedure, and in section 4 we investigate the sources of disagreement and main difficulties. In section 5 we apply the revised annotation scheme to a learner corpus sample and in 5 we present the conclusions.

2. Annotation principles

2.1. Article errors

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 (base form in bold face).

Table 1: Spanish articles (gender and number)

	Definite		Indefinite	
	Masculine	Feminine	Masculine	Feminine
Singular	***el***	*la*	***un***	*una*
Plural	*los*	*los*	*unos*	*unas*

The definite article *el* 'the' is the most frequent word in Spanish and article usage is also one of the most frequent grammatical errors[1] among learners (specially for speakers of languages that do not have articles like Chinese, Japanese, Korean or Russian.), because article choice it is the result of interacting pragmatic, semantic, syntactic and lexical constraints.

We consider the following type of errors: missing article, extraneous article and article confusion. We are only concerned with article presence and choice, so we did not tag malformation (e.g. spelling or agreement errors) or order errors. A missing article occurs when the learner does not use any article but the noun phrase should contain one. An extraneous article occurs when the article used by the learner is not necessary (zero article is correct). A confusion error occurs when the learner used a definite instead of an indefinite, or vice versa, or when the learner uses a different type of determiner instead of an article.

2.2. Level of confidence in the judgments

It was expected that the annotators would sometimes be unsure about the acceptability of article usage in a given sentence, or unable to determine the most likely correction. With regard to the level of confidence in the annotators' judgments, annotated corpora do not explicitly provide confidence levels for every annotated item. Only in some annotation experiments the annotators are asked to indicate their level of confidence (as "low" or "high") (Tetreault and Chodorow, 2008).

We did not want to force the annotators to make a best guess in "difficult" sentences because that would lower inter-annotator agreement. Instead, we gave the possibility of marking such sentences as "difficult to judge" (as in Han et al. 2006), so later we could look at the sentences marked as problematic, and analyse what they have in common.

[1] Fernández (1997) found 2.2 article errors per 100 words in a 4,433 words sample for Japanese learners.

2.3. Number of tags

With regard to the number of possible analysis a sentence can receive, error-annotated learner corpora typically contain 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 the error analysis is doubtful. In our experiment, we decided to allow only one tag per item to detect the sources of disagreement. After the revision of the annotation scheme, double tagging would be allowed in some specific cases (4.3.2).

3. Experiment

We carried out an experiment on article error annotation with the following objectives. First, calculate inter-annotator agreement for this task, which can be considered as the limit for an automatic article error detection system. Second, 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, so that measures can be taken to refine the annotation scheme and future annotation can be improved.

3.1. Data collection

A teacher of Spanish as a Foreign Language extracted sentences containing at least one article error from students' written assignments,[2] 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. The distribution of the data is as Table 2 shows. In every sentence only one highlighted noun phrase had to be annotated, so the number of sentences and the number of annotated noun phrases is the same.

Table 2: Number of noun phrases and article they contain.

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

3.2. Annotation procedure

The 300 noun phrases were tagged by 4 annotators. The annotators were two experts (teachers of Spanis 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. They were asked to classify article usage for every noun phrase using one of the following tags: missing definite (AD), missing indefinite (AI), extraneous article (E), confusion error (C), difficult to judge (NC), article is correct (OK). They were not given any more guideline or training about the expected level of intervention in the texts: they were only asked to classify the noun phrases in one of the categories.

[2] The texts were written by 4th grade Japanese students of Spanish with an intermediate level of proficiency, at Aichi Prefectural University.

4. Inter-annotator agreement

Tables 3 and 4 show the confusion matrices for expert and non-expert annotations. Observed agreement, defined as the proportion of items on which annotators agree, is 0.79 for expert annotators and 0.76 for non- experts.

Table 3: Confusion matrix for Expert 1 and Expert 2 annotators

E1↓ E2→	AD	AI	C	E	NC	OK	Total
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 Non-expert 1 and Non-expert 2 annotators

NE1↓ NE2→	AD	AI	C	E	NC	OK	Total
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
Total	39	8	32	68	6	147	300

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 chance agreement would equate to 0. For the whole set of noun phrases (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). If we exclude 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 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 noun phrases all of them containing rather "safe" article errors) kappa is 0.79 and 0.61. Although kappa values vary depending on the set of sentences used to calculate it, agreement is over 0.60, which indicates "substantial agreement".

These figures are slightly lower than those for English. In Han et al. (2006) annotators classify noun phrases in the same categories as our experiment with a kappa of 0.86, excluding correct noun phrases and sentences where they are unable to determine correct usage, which for us was 0.79 and 0.61 for experts and non-experts. The difference in the kappa values can in part be explained by the different proportion of article types in the data: while in our experiment article types are balanced (one third of noun phrases for every article type), in real texts like those used in Han et al. (2006) the zero category is the most common (followed by the definite and indefinite article) and this category also has the highest inter-annotator agreement[3], which may raise the total kappa value.

In the following sections we examine different types of disagreement: disagreement due to the annotators' individual biases (4.1), due to the annotation scheme (4.2) and genuine disagreement (4.3).

[3] 3Full agreement (that is, by the four annotators) in sentences with an indefinite article is lower (45%) than in sentences with the zero article (71.0%) $\chi^2(4, N = 299) = 16.7$, $p = 0.02$.

4.1. Disagreement due to the annotators' individual biases

As expected, non-expert annotators are less reliable than experts. First, non-expert annotators make more mistakes (they use tags which are incompatible with certain noun phrases, e.g. a missing article tag in a noun phrase already containing an article). To avoid this kind of mistakes, we should constrain the available tags depending on the input (e.g. if there is already an article in the noun phrase, do not allow the "missing" tag). Second, even though non-experts are supposed to be less confident on their annotation because pointing out errors in a text is a task for which they have no previous training, in fact they are less cautious than experts when they correct texts. 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).

Part of the variability in annotators' attitude 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.

In relation to that, annotators should be instructed about the halo effect, by which the judgment 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. While expert annotators (teachers of a foreign language) are trained in evaluation methods and therefore they are aware of the importance of reliability in students' evaluation, 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, non-experts lack this training and do not know how to perform a fair evaluation -annotation. Therefore, non-expert annotators should receive training in evaluation methods to be able to reliably correct learner texts.

4.2 Disagreement due to the annotation scheme

With regard to the reliability of the 6 tags used for annotation, "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: exclude the sentence if a reasoned annotation is considered impossible (e.g. in incomprehensible fragments of text), or use more than one error tag if both are equally possible (as we will see in 4.3.2).

4.3 Genuine disagreement

Article presence and choice can be determined by different types of factors: it mainly depends on pragmatic factors (in our data, 69.0% of noun phrases), lexico- semantic factors (20.7%) and syntactic factors (10.3%).

As for pragmatic factors, for example the definite article is used to generalize, that is, to refer to a whole class of things or people, as in (1) (we underline the noun phrase and indicate the type of article in

brackets) and to refer to something that is identifiable to the listener, as in (2).[4]The indefinite is used to refer to any object of a particular class, as in (3), and no article is used 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. [DEFINITE]
'Children cause a great deal of trouble.'

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

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

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

As for lexico-semantic factors, for example, place names usually have no article (*México*), while the definite is obligatory for rivers, mountains, seas and oceans (*el Mediterráneo*), and 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'). As for syntactic factors, for example two or more 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).

Leaving aside sentences tagged as "correct" by 4 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. Vacillation between definite article and zero article

Frequently both the definite and zero article are acceptable for the same noun phrase. This happens when the noun phrase can have two pragmatic interpretations: it can refer to a whole class of things or people in general (definite article) as in (1) or to an indefinite amount of something (zero article) as in (4). 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 to express both situations.

In our experiment, when the two pragmatic interpretations are possible for a given sentence, annotators inconsistently choose one of them: some annotators tag the noun phrase for a missing article in (5) (OK|AD|AD|OK)[5] while they tag it for an extraneous article in (6) (E |NC|OK|E), although both noun phrases can have the same pragmatic interpretations.

[4] In (2) Maria's son must be identifiable for the listener because a) Maria has only one son, or b) we have talked about him before.

[5] For very example from the learner data, in parenthesis we indicate the tags chosen by the 4 annotators, in the following order: Expert 1, Expert 2, Non-expert 1, Non-expert 2. We also indicate in brackets the article choice of the learner.

(5) Los políticos hablan en público y manifiestan sus opiniones con el objeto de conseguir votos de ciudadanos [...] [NO ARTICLE]

'Politicians talk in public and show their opinion with a view to get votes from the citizens [...].'

(6) Concretamente los cursos que consiguieron participantes japoneses y que ofrecen los certificados oficiales como IMEC(Instituto de Medicina China) continuarán existiendo [...]. [DEFINITE]

'Specifically the courses which obtained Japanese participants and offer official certificates like IMEC (Chinese Medicine Institute) will continue existing [...].'

In these cases, when both the definite and the zero article are acceptable, according to the principle of minimal change, we opt for leaving the learners' choice unchanged if it is acceptable.

4.3.2 Vacillation between indefinite article and zero article

Sometimes 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 (8) (E|C|C|E), and the annotators propose different corrections: it can be an extraneous article if the noun phrase refers to an indefinite amount of something (zero article), or a confusion error if the noun phrase refers to any object of a particular class (indefinite).

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

When the two are equally acceptable and the annotator considers she cannot make a reasoned choice, we allow two error tags (E/C). In our experiment, this only happens with the pair of tags E and C.

4.3.3 Lack of world knowledge

In some sentences, annotators have insufficient extra-linguistic knowledge to be able to determine the right article usage. For example, in (9) (OK|E|E|E) 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).

(9) 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. [DEFINITE]

'It runs approximately every thirty minutes from Nagoya station and stops in nine very interesting and touristy places, for example Nagoya Castle.'

For future annotation, if the learner's choice is acceptable in some context, as in (9), 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 Lack of knowledge about 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 clear rules about what is considered correct and incorrect by the linguistic norm. These rules are part of language planning efforts by the Spanish language academy, but native speakers -even experts- do not have sufficient knowledge about them and as a result sometimes do not follow them when they annotate learner texts. For example, in (10) (AD|AD|OK|OK) experts marked as erroneous an article usage that is

actually accepted (RAE, 2006): the zero article between the preposition *a* ('to') and the relative pronoun *que* ('which').[6]

> (10) [...] el capítulo 2 dice sobre el proceso del portuñol y los problemas a que el portuñol se enfrenta actualmente. [NO ARTICLE]
> '[...] 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 they should also consult existing rules and recommendations published in reference dictionaries and grammars as RAE (2006) and RAE (2009) to avoid contradictions between their corrections and what the linguistic norm actually says.

4. Corpus annotation with the revised annotation scheme

After analysing the main sources of disagreement, we have revised the annotation scheme as explained in Valverde & Ohtani (2014). Afterwards, we have applied the revised annotation scheme to the annotation of an approximately 30,000 words sample of the CEDEL2 learner corpus (Lozano & Mendikoetxea 2013), as shown in table 5. The texts in the sample were written without preparation, by learners whose first language is English.

Table 5. 30,000 words sample from the CEDEL2 learner corpus.

Level	Words	Texts/Learners
Beginner	10390	40
Intermediate	9960	22
Advanced	10293	20
Total	30643	82

The following categories were used: 1) Missing definite article, 2) Missing indefinite article, 3) Extraneous definite article, 4) Extraneous indefinite article, 5) Confusion error: indefinite instead of definite, 6) Confusion error: definite instead of indefinite, 7) Confusion error: another determiner instead of definite, 8) Confusion error: indefinite instead of another determiner

Annotation has been carried out by one trained annotator with the software UAM Corpus Tool (O'Donnell 2010).We have found 196 errors in 30643 words, that is 0.64/100 words. Results are shown in Table 6. As expected, the most frequent error type involves the presence/absence of article: 92 missing articles (as in 10) and 95 extraneous articles (as in 11) give account of 95.41% of errors, and confusion errors represent only 4.59% (as in 12). This proportion is very close to that found in English learner texts: Han et al. (2006) found 21.5% of extraneous articles, 58.6% of missing articles and only 6.2% of *a-the* confusion in English texts written by Japanese learners. However, the frequency of extraneous articles in our texts –very close to missing articles- is higher than in the English texts.

Table 6. Frequency of error tags by language level

Error tag	Beginner		Intermediate		Advanced	
	N	%	N	%	N	%
Missing article	51	40.80	25	64.10	16	50.0
Extraneous article	68	54.40	12	30.77	15	46.88
Confusion error	6	4.80	2	5.13	1	3.12
Total	125	100	39	100	32	100

[6] The definite article is also acceptable but not *obligatory*. The definite article would be obligatory if the antecedent referred to a person, or if the subordinate clause was negative.

(10) Jude Law tiene pelo rubio y es Ingles. [0 → DEFINITE]
'Jude Law has blond hair and is English'

(11) Fui en el junio y no llovó allí. [DEFINITE →0]
'I went in June and it did not rain there'

(12) Me encanta ir a la Universidad porque es la experiencia Buena.
[DEFINITE→INDEFINITE]
'I love going to the University because it is a good experience'

Among missing articles, the most frequent is the omission of the definite (88/92), as shown in Table 7. Among extraneous articles, the proportion of definite and indefinites is more balanced (37 vs 58).

Table 7. Frequency of error types by language level

Error type	Beginner		Intermediate		Advanced	
	N	%	N	%	N	%
Missing type	**51**	100	**25**	100	**16**	100
Missing definite	50	98.04	25	100.00	13	81.25
Missing indefinite	1	1.96	0	0.00	3	18.75
Extraneous type	**68**	100	**12**	100	**15**	100
Extraneous indefinite	30	44.12	2	16.67	5	33.33
Extraneous definite	38	55.88	10	83.33	10	66.67
Confusion type	**6**	100	**2**	100	**1**	100
Definite instead of indefinite	5	83.33	0	0.00	0	0.00
Indefinite instead of definite	0	0.00	0	0.00	0	0.00
Indef. instead of another det.	0	0.00	1	50.00	1	100.00
Another det. instead of definite	1	16.67	1	50.00	0	0.00
Total	125		39		32	

From this data we can extract some statistically significant differences among learners. For example, beginner learners tend to overuse articles (68 of 125 of article errors, that is 54.40%), while intermediate learners underuse them (25 out of 39 article errors, that is, 64.10%), while among advance learners omission and addition errors are balanced. As for the type of missing article, among beginner learners the indefinite is very rare (1 out of 51), but not so rare for advanced learners (3 out of 16).

5. 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 4 annotators (2 experts and 2 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. Non-experts are less reliable than experts, and the annotation scheme (the tag "difficult to judge") is also responsible for part of the disagreement.

As for genuine disagreement among annotators, some pragmatic distinctions are specially problematic: the distinction between a) a whole class of things or people in general (definite article) and b) an indefinite amount of something (zero article), and the distinction between a) an indefinite amount of something (zero article) and any object of a particular class (indefinite article). In addition to that, some times more world knowledge is needed to determine whether article presence and choice is acceptable or not. As for article usage governed by syntactic and lexico-semantic factors, annotators sometimes

disagree in determining the right article usage because they lack knowledge about the existing prescriptive rules published by the Spanish language academy.

To improve annotation reliability, annotators need to be trained in language evaluation methods and have to consult published prescriptive rules about article usage. After annotating a 30,000 words sample from the CEDEL2 learner corpus with a revised annotation scheme, we have found that the most frequent error types are missing and extraneous articles, which give account of 94.5% of errors. In our annotation sample, beginner learners tend to overuse articles, intermediate learners tend to underuse them and among advanced learners addition and omission errors are equally frequent.

Acknowledgements

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

References

Alonso, R. Castañeda,A. Martínez, P. Miguel, L. Ortega, J. & Ruiz, J. (2013). Students' Basic Grammar of Spanish. Difusión.

Boyd, A. (2010). Eagle: an error-annotated corpus of beginning learner german. *Proceedings of LREC-10*, Malta.

Butt, J. & Benjamin, C. (2014). A New Reference Grammar of Modern Spanish. Routledge.

Chodorow, M. Dickinson, M. Israel, R. & Tetreault, J. (2012). Problems in evaluating grammatical error detection systems. *Proceedings of COLING 2012*, pages 611–628, Mumbai, Desember.

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

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

Fernández, S. (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).

Han, N., Chodorow, M. and Leacock, C. (2006), Detecting errors in English article usage by non-native speakers, Natural Language Engineering 12 (2): 115–129.

Lee, S. Dickinson,M. & Israel, R. (2012). Developing learner corpus annotation for korean particle errors. Association for Computational Linguistics (editor), *Proceedings of the Sixth Linguistic Annotation Workshop, LAW VI*, pages 129–133, Stroudsburg.

Lozano, C., & Mendikoetxea, A. (in press 2013). Learner corpora and second language acquisition: the design and collection of CEDEL2. In N. Ballier, A. Díaz-Negrillo, & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins.

Lüdeling,A., Walter, M., Kroymann,E. & Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of the Corpus Linguistics 2005 Conference*, Birmingham, United Kingdom, July.

O'Donnell, M. (2010) UAM Corpus Tool. Available at http://www.wagsoft.com/CorpusTool/

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.

Reidsma, D & Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Rozovskaya, A. & Roth, D. (2010). Annotating ESL errors: Challenges and rewards. *Proceedings of NAACL'10 Workshop on Innovative Use of NLP for Building Educational Applications*. University of Illinois at Urbana–Champ.

Tetreault, J. & Chodorow, M. (2008). Native judgments of non-native usage: Experiments in preposition error detection. *Proceedings of the Workshop on Human Judgments in Computational Linguistics at the COLING 2008*, pages 24–32.

Tetreault, J, Filatova, E. & Chodorow, M. (2010), Rethinking Grammatical Error Annotation and Evaluation with the Amazon Mechanical Turk, Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pages 45–48, Los Angeles, California, June 2010.

Tetreault, J. & Leacock (2014), Automated Grammatical Error Correction for Language Learners, Turorial at *25th International Conference on Computational Linguistics (COLING 2014)*, Dublin, August 23-29.

Valverde, M.P & A. Ohtani (2014, forthcoming), Annotating article errors in Spanish learner texts: design and evaluation of an annotation scheme. *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC 2014).*