System 76 (2018) 13–24

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system

English article acquisition by Chinese learners of English: An analysis of two corpora

William Leroux*, Tyler Kendall

Department of Linguistics, 161 Straub Hall, 1290 University of Oregon, Eugene, OR, 97403-1290, USA

ARTICLE INFO

Article history:
Received 8 May 2017
Received in revised form 4 April 2018
Accepted 22 April 2018

1. Introduction

Non-native-like use of articles is one of the most common features of learner English (Chuang, 2005, p. 25). While variation between *the*, *a*, and absent articles occurs with high frequency in both spoken and written English, the nature and purpose of this variation can be opaque to learners (as well as native speakers). In addition to alternations between *the* and *a*, L2 speakers also have to learn that absent articles are of, at least, two types, what we will refer to as zero articles and null articles, following Master (2003).

There is a rich literature surrounding the issue of L2 article acquisition. Some researchers have investigated articles in terms of the order of acquisition of different morphemes (e.g., Goldschneider & DeKeyser, 2005), while others have examined various factors that affect accurate article use (Amuzie & Spinner, 2013; Butler, 2002; Geng, 2010). The present study attempts to contribute further to our understanding of the factors that influence article use by English learners. We investigate the effects of several linguistic and demographic factors by examining two corpora of Chinese learners of English. Using these corpora, we ask whether speaker differences within and across the two corpora relate to article use and how factors like speaker proficiency, the pragmatic context of an article, and the abstractness and plurality of a noun influence article choice. One of the corpora, the International Corpus Network of Asian Learners of English (hereafter ICNALE) (Ishikawa, 2013, 2014, 2017), contains the speech of Asians learning English at universities. The second, which we will refer to as the STAAAF data (from an acronym from the original project (Rivers et al., 2009; Kendall, Rivers, & Dodsworth, 2012)) includes Chinese adults who live in New York City. Much of the research on adult L2 article use has been done on college students, likely for practical reasons related to university-based research and teaching (Butler, 2002; Geng, 2010; Tarone & Parrish, 1988; Watcharapunyawong & Siriluck, 2013). While some studies have examined adult L2 article use (e.g., Jin, Afarli, & van Dommelen, 2009), those projects are outnumbered by those done on college students. Through corpus-based analysis of spontaneous speech data, and by considering both adults well past university age and university learners, we seek to build on the observations made by previous researchers.

* Corresponding author. 1290 University of Oregon, Department of Linguistics, Eugene, OR 97403-1290, USA.
E-mail addresses: lwilliamr@gmail.com (W. Leroux), tsk@uoregon.edu (T. Kendall).

https://doi.org/10.1016/j.system.2018.04.011

2. Background

The present study focuses on speakers from Chinese language backgrounds. Mandarin and Cantonese both lack articles. The L1 of a speaker has been shown to affect their L2 article use (Ghisseh, 2009). Research has found that a presence of articles in the L1 of a speaker promotes acquisition of an L2 article system, while an absence has the inverse effect (Ekiert, 2004; Ionin & Montrul, 2010). Article use improves with experience (Ekiert, 2004; Hua & Lee, 2005). However, it is often subject to fossilization in advanced speakers (White, 2003).

There are many different uses for each English article form, and determining the purpose of any given occurrence proves difficult, especially for those who have no L1 article system to support their understanding. Scholars have used various metrics to define the meaning of definite and indefinite articles. Butler (2012) created a semantic map which seeks to explain the production of every English article. His work is exhaustive but is perhaps too complex for an analysis of L2 articles. Another approach, developed by Huebner (1983), is widely used in second language acquisition studies. It makes use of two interacting factors, Hearer Knowledge and Specific Reference, to explain the distribution of articles (pp. 131–133). Hearer Knowledge (HK) is whether or not the listener knows about the speaker's referent. Specific Reference (SR) is whether or not the noun in question is "specific." For example, in the sentence "I like the strawberries grown in marshlands," "strawberries" refers to a specific kind of strawberry, but "marshlands" refers to marshlands in general, not to a specific instance of marshlands.

Research shows that oftentimes learners will associate the +SR feature with the definite article, after initially associating it with +HK (Butler, 2002, p. 474). However, in English, the distinction between a definite article and an indefinite article is not entirely determined by specific reference. The more common distinction comes from the difference between +HK and −HK. Table 1 presents the four possible permutations of the pairing of HK and SR, along with the articles expected in L1 English.

Huebner's (1983) classification system originally did not consider uses of the zero article form which indicate a very definite situation. These have been dubbed "null" articles in other literature (Master, 2003). For example, in the sentence "Dinner was delicious last night," the word "dinner" refers to a specific, known, definite noun. We have included cases of the null article in our analysis, and we differentiate these null forms from other zero cases, which are hereafter referred to as 0.

Beyond Huebner's (1983) work, several studies have found significant differences in how speakers use articles based on HK and SR (Ekiert, 2004; Geng, 2010; Tarone & Parrish, 1988; Young, 1996).

In addition to the pragmatic features of HK and SR, the semantic content of noun phrases is also an important factor underlying article realization. Researchers have considered the distinction between "abstractness" and "concreteness" in previous literature on article acquisition. Amuzie and Spinner (2013) describe "abstract nouns" as nouns that "denote a quality, an attribute, a feeling, or an idea that cannot be seen or touched, while concrete nouns name tangible items that have physical properties" (p. 416). Other studies have found that learners have more difficulty using abstract nouns in native-like ways than concrete nouns (Hua & Lee, 2005).

As indicated in Table 1, plurality also plays a significant role in the production of articles. Singular nouns, plural nouns, and non-count nouns, all interact with articles differently. For example, it is grammatical to say "I have a deadline" in the singular, and "I have 0 deadlines" in the plural, but not "*I have deadline" or "*I have a deadlines." The non-count status of a noun can be the difference between the grammatical "I like lava" and the ungrammatical "*I like a lava." Other literature has found that articles which correlate with plurality, *the* and 0, appear earlier within the development of an L2 grammar (Master, 1997, p. 216). This suggests that plurals will encourage native-like speech, while singular nouns may discourage it. Considering that speakers can often use abstract nouns in both plural and non-count ways (for example, one can say "belief is the cornerstone of faith" and "we respect beliefs from all perspectives"), the issue of plurality becomes even more complex.

In order to cope with these complex and varied factors which affect article use, learners have to develop strategies to pick one article over another. Other researchers (e.g., Butler, 2002) have noted that L2 learners develop strategies to use difficult L2 features accurately. The most desirable strategy would be to internalize the English article system like a native speaker. However, learners may adapt simpler or non-native-like strategies, which they can then update as they gain proficiency. We conceptualize these strategies as methods to increase the overall rate of grammatical (or seemingly grammatical) utterances.

For example, *the* flooding is a phenomenon noted throughout the literature on L2 article acquisition (Master, 2003). It refers to the case of L2 speakers overusing the article *the* across many, often inappropriate, contexts. We argue that this is a strategy for increasing the rate of accurate articles in L2 speech. While the overuse of *the* does lead to its use in inappropriate

Table 1
Hearer knowledge, specific reference, plurality, and articles.

Features	Singular	Non-count	Plural	Example
−HK −SR	a(n)	0	0	"I want an apple."
−HK +SR	a(n)	0	0	"I was carrying a baby."
+HK −SR	a(n), 0, the	0, the	0, the	"I'm looking for the best pizza in town."
+HK +SR	the, null	the, null	the, null	"The man over there is so ambivalent."

Adapted from Ekiert, 2004, p. 11.

contexts, L2 speakers might utilize *the* flooding knowing that *the* can be interpreted as accurate across a wide variety of noun phrase types. Consider the following three examples.[1]

Example 1. *I ate **a** sandwiches at **a** café with **a** police officer.

Example 2. *I ate **0** sandwiches at **0** café with **0** police officer.

Example 3. I ate **the** sandwiches at **the** café with **the** police officer.

In example 1, "*a sandwiches" is incorrect because it is attached to a plural noun, and in example 2, "café" should have an article. Example 3, however, will be interpreted as grammatical. It might not actually present what the speaker is trying to say (maybe the speaker does not mean to talk about a definite café, but an indefinite one) but to listeners, it appears correct. In many situations, the most likely mistake related to *the* would be omitting it. *The* flooding, then, by increasing the appearance of *the*, might act as a strategy to increase overall article grammaticality.

Our study examines how article use compares across our two corpora and asks how the factors just described influence article use in these data. Based on prior work, we expect that speaker proficiency, concrete nouns, plurality, and the articles *the* and 0 would correlate with correct article use. We also expect to find patterns which reveal the simplified strategies our speakers use to produce articles.

3. Data

As mentioned in section 1, the data for this study come from two sources. The first is a dataset developed as a part of a 2009 project which we refer to as the STAAAF dataset (Rivers et al., 2009; Kendall et al., 2012). The data are comprised of three 2-hour long group interviews with a total of twelve L1 Chinese speakers living in New York City, in the United States (U.S.). Four of these speakers are from mainland China, four are from Hong Kong, and four are from Taiwan. Throughout the paper we refer to their areas of origin as speakers' *background*, which reflects somewhat different experiences and L1 histories. Demographic information about the speakers is shown in Table 2. All speakers are male, in professional fields, and at roughly similar ages, so, apart from their places of origin, demographic differences are minimal. (Age of arrival and length of residency in the U.S. are included in the table although these factors did not arise as impacting article use in statistical tests of our data. They are not discussed further in this paper.)

Altogether, these speakers represent a professional, long-term resident population. While all participants had lived in the U.S. for years at the time of their interviews, many described their social circles outside of work as primarily Chinese, with one claiming that "over 80 percent" of his conversations with friends were in Mandarin. However, all were extremely proficient in English, and used it on a daily basis to interact with people outside of the Chinese community.

Speakers were interviewed in three groups according to their background (e.g., all four Hong Kong speakers were interviewed together). They took turns answering questions on a variety of topics, including religion, holidays, work culture, and racism, typically responding individually to the interviewer, though some cases of dialogue between participants occurred.

The second source of data for this project comes from ICNALE, a public corpus of Asian learners of English (Ishikawa, 2013, 2014, 2017). Unlike the STAAAF project, which focuses on spontaneous interview speech, ICNALE was produced in a more structured setting, with prescribed topics and speech duration. Participants for the corpus, Asian college students learning English, were given a set time to either write a short essay or to produce a 1-min speech. Students discussed one of two topics; whether smoking should be allowed in restaurants, or whether students should have part time jobs (Ishikawa, 2013). The

Table 2
STAAAF dataset demographic information.

Group	Speaker ID	Gender	Age	Age of Arrival	Length of Residency
Mainland China	1	Male	33	12	21
Mainland China	2	Male	36	20	16
Mainland China	3	Male	28	7	21
Mainland China	4	Male	27	11	16
Hong Kong	5	Male	32	15	17
Hong Kong	6	Male	34	10	24
Hong Kong	7	Male	26	8	18
Hong Kong	8	Male	26	6	20
Taiwan	9	Male	34	10	24
Taiwan	10	Male	35	12	23
Taiwan	11	Male	35	9	26
Taiwan	12	Male	29	15	14

[1] All examples, unless otherwise indicated, were created by the authors.

present project focuses on only spoken data from ICNALE. ICNALE participants were sorted into four proficiency groups based on English (TOEIC) test scores, as shown in Table 3.

We excluded A2 speakers, judging their speech not to be proficient enough to be utilized for this study. While ICNALE includes hundreds of participants from several Asian countries and language backgrounds, our study concentrates on native speakers of Chinese (Mandarin and Cantonese). All speakers, except five B2 level participants, were mainland Chinese nationals. We also included five B2 Taiwanese participants to increase the number of higher proficiency subjects.

The two datasets are meant to complement each other. The ICNALE corpus contains a much larger set of speakers than STAAAF, but provides very little data per speaker. It also represents a more typical population for analysis of learner behavior. The STAAAF data augment the ICNALE corpus to provide relatively abundant, key information from adults working in an English context. This inclusion of older adults is especially valuable in the field of English L2 article acquisition, which has been understandably biased towards university age learners of English. STAAAF speakers have been living and working in the United States for an average of 18.3 years, and thus we would expect them to represent a stage of L2 development beyond the college level. By focusing on both college-aged learners and adult L2 English speakers, we can examine language in development in different contexts and at different stages.

Our analytic methods follow practices from variationist sociolinguistics (cf. Tagliamonte, 2006), where the articles were conceptualized as alternate possible realizations of a linguistic variable. Articles, including 0s and null articles, in the data were identified and coded manually through a combination of listening to the audio and by close analysis of time-aligned transcriptions.

We include in our analysis all noun phrases, except for proper nouns and those with non-article determiners, like demonstratives, because we saw no evidence of speakers in these corpora affixing articles to these environments; for example, there are no constructions like "*the that apple" in our data. For each noun phrase we coded whether the form used by the speaker matched that "expected" by an L1 speaker of U.S. English. We targeted what we will call "surface level grammaticality." When there was some ambiguity about whether a realized article would have been appropriate, we consider the uttered article grammatical, unless we have evidence to support a different conclusion. Consider the following examples from an intermediate ICNALE speaker (CHN_SMK1_031_B2_1).

Example 4. I think it isn't a good way to preventing **the** smoking because some of 0 people are – are smoking for a long time and some people in 0 restaurant they – they are accustomed to 0 smoking.

Example 5. Those people who often smoking may stay in **the** smoking area, and it will not influence the people who doesn't smoke.

We considered the bolded *the* in example 4 ungrammatical, because we judged that the speaker was talking about smoking in general, and so should have used the 0 article. This is supported by a later use of "smoking" without a *the*. However, we considered the bolded *the* in example 5 to be grammatical, even though the speaker might have been referring to smoking areas in general, and simply forgotten to add the plural –*s* morpheme. This is because *the* can be used in such hypothetical situations, and there was little evidence to suggest that the speaker meant to say something else.

We also coded all noun phrases for a number of factors including Hearer Knowledge (coded as +HK or –HK), Specific Reference (+SR or –SR), plurality status (singular, plural, or non-count), abstractness (abstract or concrete), and idiomatic status (idiom or not an idiom). For our analysis HK, SR, and idiomatic status are combined into a single factor, with five levels, –HK –SR, –HK +SR, +HK –SR, +HK +SR, and Idiom, with the selection of Idiom overriding the HK and SR rating for the noun phrase. Idioms were coded based on a subjective judgment of whether or not a noun would have been learned on a phrasal level. For example, we included the nouns in the phrase "a bird in the hand is worth two in the bush" as idioms, because the articles are fixed in this expression. Our goal was not so much to analyze idiomatic nouns as to isolate nouns that might not have been produced within the HK/SR paradigm.

To illustrate this coding, we consider example 6, the first sentence of a B2 ICNALE speaker (CHN_SMK1_032_B2_0).

Example 6. I think no smoking in **0 restaurant** is a good idea because not all the people find it very – find it acceptable.

The word "restaurant" is being introduced into the monologue without prior mention, so it is –HK. It does not refer to any specific restaurant, so it is –SR. We interpret it as plural, so we mark its plurality status as plural. Restaurants are physical objects in the world, so the word is concrete. The phrase does not appear like something that would be learned wholesale, with prescribed articles, so it is not coded as an idiom.

In the next section, we provide a summary description of the data we obtained and then we review the results of two sets of analyses, first asking what factors condition L2 speakers' use of native-like articles and then asking what factors condition

Table 3
Proficiency levels by TOEIC score.

TOEIC Score	<545	550–670	670–785	785+
Proficiency level	A2	B1_1	B1_2	B2
Number included in present study	0	27	25	20

Adapted from Ishikawa, 2017.

the realization of each main type of article. Following the quantitative analysis, we discuss a series of more qualitative observations that emerge from our project.

4. Results and discussion

4.1. Overview

Altogether, we obtained 1871 article data points, consisting of $N = 1022$ for the STAAAF data, and $N = 849$ for the ICNALE corpus. These counts include 0 and null cases. The relative rates of use for the three realized article types, 0 (which includes null articles), the indefinite article *a(n)* (representing both *a* and *an*), and the definite article *the*, are shown for the two corpora in Fig. 1. Fig. 2 depicts the speakers' rates of accuracy for their article use in comparison to the "expected" article choice, based on the coding by an L1 English speaker (as described above). Here the data are separated into four categories, with 0 and null articles represented individually.

Overall, speakers in both datasets realize relatively similar accuracy rates, although STAAAF speakers are slightly more accurate across the board. For 0 cases and *the*, speakers realize relatively high accuracy rates across both corpora (87.8%–94.7% accuracy) but for *a(n)* rates are somewhat lower. ICNALE speakers are accurate for *a(n)* about two-thirds of the time (67.8%), while STAAAF speakers accurately use *a(n)* 79.2% of the time. Null article accuracies show the greatest difference between the two datasets, with STAAAF speakers accurately omitting null articles 97.2% of the time but ICNALE speakers achieving only 71.4% accuracy.

4.2. What factors account for accurate article use?

Fig. 3 displays speakers' accuracies for the Hearer Knowledge, Specific Reference, and idiomatic status (HKSRI) factor subdivided by plurality. Dashed lines display the overall accuracies for each of the HKSRI factor levels. *N*s are not distributed evenly and for three pairings *N*s are quite low (non-count −HK +SR = 4, plural +HK −SR = 8, and singular +HK −SR = 12) so the overall accuracies (dashed lines in Fig. 3) can differ widely from the individual pairings (bars in Fig. 3) and the rates for the three low *N* pairings need to be interpreted cautiously. As we would expect (see Table 1, above), these two factors are influential in affecting speakers' accurate use of articles. In particular, we can observe that speakers have trouble for singular -SR cases, and, overall, are most accurate for +HK cases.

In order to better understand what factors account for accurate article use, the data were analyzed using mixed-effect logistic regression (using the *lme4* package for R; Bates, Maechler, Bolker, & Walker, 2014). Speakers were included as a random intercept and the various factors coded for the data (described above) were tested as potential fixed effects. As mentioned above, HK, SR, and idiomatic status were tested as a single, combined factor (HKSRI).

Results here report the best statistical model, which was determined after starting with a maximal model (i.e. containing all of the relevant factors as main effects) and then removing factors one at a time that did not significantly improve the model according to likelihood ratio tests. Following this, two-way interactions were tested for the remaining factors and these were added if they significantly improved the model. See Baayen (2008) for information about these statistical modeling procedures. Table 4 presents the statistical results for the significant factors. We consider *p* values below 0.05 as significant and *p* values between 0.05 and 0.1 as marginally significant; *p* values above 0.1 are indicated in the table in brackets. Factors with positive estimates (and significant *p* values) favor the dependent variable (here indicating higher accuracies) and negative estimates disfavor the dependent variable (lower accuracies).

Model results indicate that the combined HK, SR, and idiomatic status (HKSRI) factor is significant, and confirm the view in Fig. 3 that speakers achieve the highest accuracy for articles that are +HK +SR. The high accuracy for +HK +SR environments is not surprising, as these often take *the*, whether they are singular, non-count, or plural. The flexibility of this article likely leads to greater accuracy. It is also worth noting, however, that these environments can take null articles in some situations, so the accuracy of +HK +SR environments implies that many of the speakers have mastered the difference between the null

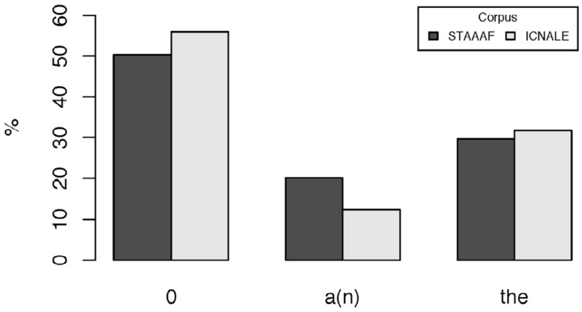

Fig. 1. Use of article types, by corpus.

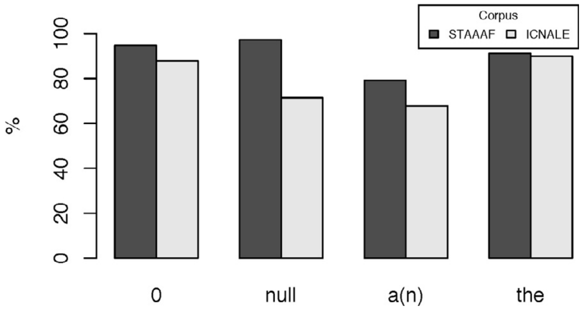

Fig. 2. Accuracy of expected article types, by corpus.

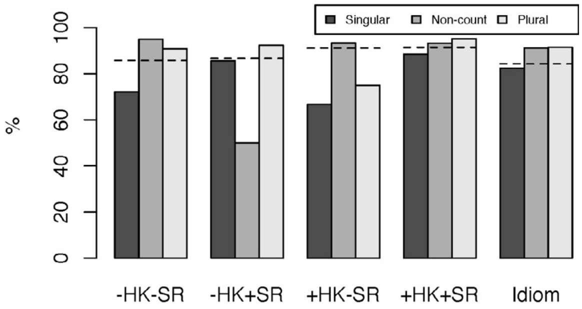

Fig. 3. Accuracy by HKSRI and plurality status.

Table 4
Model results for accurate article use.

Factor	Estimate	Std. Error	*p* value
(Intercept)	1.273	0.207	–
HKSRI (ref = −HK −SR)	–	–	–
−HK +SR	0.605	0.404	[0.1341]
+HK −SR	−0.264	0.652	[0.6856]
+HK +SR	1.050	0.278	0.0002
Idiom	0.620	0.221	0.0051
Plurality (ref = singular)	–	–	–
Non-Count	2.383	0.618	0.0001
Plural	1.449	0.222	<0.0001
Corpus (ref = STAAAF)	–	–	–
ICNALE	−0.700	0.205	0.0006
HKSRI × Plurality Interaction	–	–	–
−HK +SR: Non-Count	−3.993	1.278	0.0018
+HK −SR: Non-Count	−0.195	0.927	[0.8336]
+HK +SR: Non-Count	−1.786	0.746	0.0166
Idiom: Non-Count	−1.518	0.882	0.0853
−HK +SR: Plural	−0.755	0.740	[0.3076]
+HK −SR: Plural	−1.444	1.076	[0.1794]
+HK +SR: Plural	−0.349	0.607	[0.5656]
Idiom: Plural	−0.814	0.600	[0.1753]

*Not selected as improving the model: abstractness, proficiency level (within ICNALE), and group background.

article and *the*. Idioms represent a complex category in the data. Raw accuracies (in Fig. 3) show idioms to be, on average, less accurate than other HKSRI categories, although the statistical results indicate they are slightly, but significantly, more accurate than the reference level for the factor, −HK −SR. This apparent contradiction may stem from speakers having mastered frequent idioms but not mastered infrequent ones, something that our data coding does not allow us to address.

Plurality also arises as an influential factor on article accuracy. Speakers are significantly more accurate for plural and non-count forms than singular forms. Plurality, not surprisingly (see Table 1), interacts with HKSRI status so that speakers are significantly less accurate for non-count +SR (both −HK +SR and +HK +SR) cases.

Finally, we find a significant difference between the corpora. ICNALE learners are significantly less accurate than the STAAAF speakers, adults who are living and working in New York City. While that difference indicates some improvement for

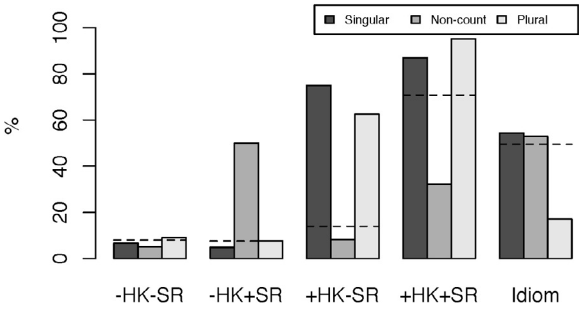

Fig. 4. Use of *the* by HKSRI and plurality status.

second language learners over time, we also note that proficiency level within ICNALE did not emerge as a significant factor within this data. Thus, while there are accuracy differences across the two datasets, finer differences of proficiency among the ICNALE learners do not affect article accuracy.

4.3. *What factors account for the use of specific articles?*

Next we turn to consider what factors significantly influence the specific realization of each of the main article types. We first look at the definite article in these data, asking what factors contribute to speakers choosing *the* over the other article options. Overall, 30.6% of all articles were *the*. Fig. 4 displays the percentage use of *the* across HKSRI and plurality pairings. As in the last figure, dashed lines display the overall percentages for each HKSRI factor level. As we would expect (see Table 1), *the* is associated with +HK environments and rare in −HK environments, although non-count cases disfavor *the* even when +HK.

A closer determination of the factors influencing the use of *the* was made through the use of mixed-effect logistic regression following the same procedures described above. Table 5 presents the factors that emerge as significant predictors of *the* use.

The factors HKSRI, plurality, and corpus all emerge with significant effects. In line with Fig. 4, *the* is significantly associated with +HK cases. Idioms also favor *the* in comparison to both −HK cases. Plurality does not obtain significance as a main effect, but shows some substantial interactions with HKSRI. Non-count +HK forms strongly disfavor *the*, despite the fact that its use is grammatical in such situations. ICNALE speakers are more likely than STAAAF speakers to use *the*. While more detailed measures of proficiency within ICNALE do not emerge as significant, this difference between the less experienced ICNALE speakers and the STAAAF speakers could be interpreted as a greater tendency towards *the* flooding by ICNALE speakers.

Next we turn to consider what factors significantly influence the realization of *a(n)* in these data. As before, we display the relative use of *a(n)* across HRSKI and plurality factors, in Fig. 5, and then we apply mixed-effect logistic regression to

Table 5
Model results for use of *the*.

Factor	Estimate	Std. Error	*p* value
(Intercept)	−2.953	0.320	–
HKSRI (ref = −HK −SR)	–	–	–
−HK +SR	−0.065	0.657	[0.9211]
+HK −SR	3.968	0.749	<0.0001
+HK +SR	4.773	0.355	<0.0001
Idiom	2.897	0.300	<0.0001
Plurality (ref = singular)	–	–	–
Non-Count	−0.643	0.658	[0.3286]
Plural	0.260	0.315	[0.4086]
Corpus (ref = STAAAF)	–	–	–
ICNALE	0.642	0.226	0.0045
HKSRI × Plurality Interaction	–	–	–
−HK +SR: Non-Count	3.250	1.358	0.0167
+HK −SR: Non-Count	−3.319	0.990	0.0008
+HK +SR: Non-Count	−2.143	0.718	0.0028
Idiom: Non-Count	0.543	0.759	[0.4744]
−HK +SR: Plural	0.262	0.907	[0.7731]
+HK −SR: Plural	−0.929	1.078	[0.3886]
+HK +SR: Plural	0.772	0.645	[0.2311]
Idiom: Plural	−1.815	0.524	0.0005

*Not selected as improving the model: abstractness, proficiency level (within ICNALE), and background.

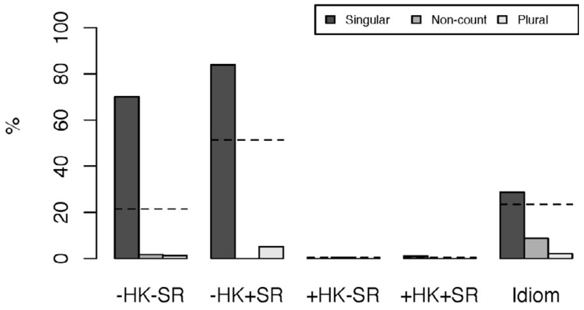

Fig. 5. Use of *a(n)* by HKSRI and plurality status.

Table 6
Model results for use of *a(n)*.

Factor	Estimate	Std. Error	*p* value
(Intercept)	0.848	0.198	–
HKSRI (ref = −HK −SR)	–	–	–
−HK +SR	0.817	0.351	0.0197
Idiom	−1.867	0.216	<0.0001
Plurality (ref = singular)	–	–	–
Non-Count	−3.776	0.540	<0.0001
Plural	−4.932	0.370	<0.0001
Concrete (ref = abstract)	−0.479	0.230	0.0375

*Not selected as improving the model: corpus, proficiency level (within ICNALE), and background.

determine the significant predictors for *a(n)*. As can be seen in Fig. 5, the speakers rarely use *a(n)* for non-singular cases, and almost entirely avoid *a(n)* for +HK cases.

For statistical consideration, we remove the +HK cases, as these are almost exclusively not realized as *a(n)* (only 3 cases are *a(n)* out of 622 total +HK cases). This leaves a total of 1249 tokens in the dataset, with 24.6% realized as *a* or *an*. The results for the best model are shown in Table 6.

HKSRI, plurality, and abstractness emerge as significant predictors for *a(n)* through the statistical model. −HK +SR cases are the largest favoring environment for *a(n)*, significantly favoring *a(n)* in comparison to −HK −SR, the reference level for HKSRI. As should be expected, *a(n)* is quite rare in non-singular cases and therefore is strongly disfavored for non-count and plural forms. It is perhaps surprising then that any non-count or plural cases do occur with *a(n)* for these speakers, but 1.7% (non-count) and 1.4% (plural) of the non-singular noun phrases do occur with *a(n)*. Finally, concrete forms relatively disfavor *a(n)* in comparison to abstract forms.

Lastly, we turn to consider what factors significantly influence the realization of 0 and null forms. Since these forms are indistinguishable in their realization, we assess them together. Again, we display the relative use of these forms in Fig. 6 and then we apply mixed-effect logistic regression to determine the significant predictors. As expected, 0 was rare (though present) in singular cases, because most singular cases do not grammatically allow for 0. Plurals and non-count environments naturally promoted the use of 0.

Considering the 0 and null forms statistically, the results for the best model are shown in Table 7.

As with the model for *the*, the factors HKSRI, plurality, and corpus all significantly contribute to the use of the 0 form. Plurality, as is expected, has strong effects, with non-count and plural cases favoring 0 forms over singular cases. +SR forms

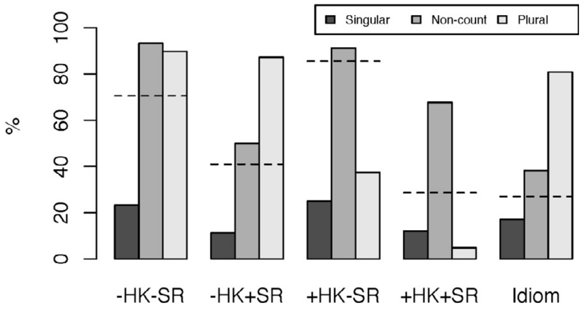

Fig. 6. Use of 0 and null articles by HKSRI and plurality status.

Table 7
Model results for use of 0 and null articles.

Factor	Estimate	Std. Error	*p* value
(Intercept)	−1.070	0.201	–
HKSRI (ref = −HK −SR)	–	–	–
−HK +SR	−1.030	0.438	0.0188
+HK −SR	0.115	0.700	[0.8697]
+HK +SR	−0.844	0.277	0.0023
Idiom	−0.351	0.226	0.1212
Plurality (ref = singular)	–	–	–
Non-Count	4.109	0.555	<0.0001
Plural	3.486	0.228	<0.0001
Corpus (ref = STAAAF)	–	–	–
ICNALE	−0.405	0.193	0.0358
HKSRI × Plurality Interaction	–	–	–
−HK +SR: Non-Count	−1.693	1.222	[0.1660]
+HK −SR: Non-Count	−0.483	0.908	[0.5948]
+HK +SR: Non-Count	−1.264	0.622	0.0420
Idiom: Non-Count	−2.989	0.680	<0.0001
−HK +SR: Plural	0.552	0.668	[0.4087]
+HK −SR: Plural	−2.938	1.037	0.0046
+HK +SR: Plural	−4.416	0.606	<0.0001
Idiom: Plural	−0.576	0.472	[0.2223]

*Not selected as improving the model: abstractness, proficiency level (within ICNALE), and background.

disfavor 0 and null forms and this is especially the case for +HK +SR cases when non-singular. In terms of corpus, ICNALE speakers use 0 and null forms less than the STAAAF speakers.

4.4. Discussion

We turn now to consider some ramifications of these findings. While our corpus-based methods cannot speak definitively to the motivations or grammatical knowledge of our speakers, our findings nonetheless point to ideas and suggestions for further consideration.

We begin by considering factors that proved to be less significant than we would have expected. As mentioned earlier, the ICNALE data include the proficiency level of each speaker. We expected that learners' proficiency levels would correspond to their accurate use of articles, as it did in previous literature (Butler, 2002; White, 2003). However, the data only weakly supported this hypothesis. Our analysis found that speakers in the STAAAF dataset, adult L2 speakers of English who live and work in the English-speaking environment of New York City, were more accurate than the L2 learners of the ICNALE corpus. Further, ICNALE speakers were more likely to use *the* than STAAAF speakers (implying the presence of *the* flooding), and our statistical model for the use of the zero article (0) found that ICNALE speakers were less likely to use 0 than STAAAF speakers. We also note that STAAAF speakers used a higher proportion of *a(n)*, the article with the lowest rate of accuracy, which is presumably difficult to use, and they used it more accurately. These differences between the corpora may be evidence of a broader influence of language experience and proficiency. However, these cross-corpora differences are, on the one hand, not huge, and on the other hand, somewhat weak evidence for proficiency level differences given that differences did not emerge based on proficiency levels within ICNALE. It could also be that the different tasks undertaken by the speakers in the two corpora led to somewhat different rates of article use (something we cannot test here). Overall, proficiency level was of less importance in article use for these data than expected.

For both the STAAAF and the ICNALE data, *a(n)* had the lowest rate of accuracy, while 0 and *the* had higher ones. STAAAF speakers used the 0 article about as accurately as *the*, while ICNALE learners used *the* most accurately, as was seen in Fig. 2, above. Null articles, which have the same surface realization as 0 cases but are used in very specific/definite situations, were used less accurately by ICNALE speakers but rather accurately by STAAAF speakers. Other studies have found that learners from L1 backgrounds without articles tend to master 0 and *the* earlier, although there are some exceptions (Ekiert, 2004; Master, 1997). We would expect articles developed earlier in L2 learning to be present at higher accuracy rates, so the accuracy rate patterns here can be interpreted as in line with the majority of earlier studies.

Speakers were more likely to produce native-like articles in +HK +SR environments than in other environments, and generally −HK cases appeared harder for these speakers to master than +HK cases. This accuracy for +HK +SR cases is not surprising, given that such nouns can take *the* in many contexts. For example, they take *the* across all plurality environments (singular, non-count, and plural). However, speakers still had to contend with the complexity of when to use null articles in +HK +SR environments. The fact that +HK +SR is still the most accurate environment suggests that our participants (particularly the STAAAF speakers) had mastered the null article fairly well.

The rate of accurate article use tends to increase as an article becomes viable in more contexts. The 0 article can be used with non-count nouns, in many idiomatic expressions, and with the majority of plurals, and this fact helps to explain why it is more often accurate than *a(n)*, which occurs only with singular, −HK nouns. 0 is also a “default form” in that speakers who forget to apply articles entirely will, after all, supply it. Of the 102 errors in the STAAAF data, 69 (67.6%) were from speakers using a 0 where they should have used another article, indicating that the majority of errors are errors of omission. As Ekiert noted in her own research, this preponderance of the 0 article is difficult to analyze (2004 p. 15). It is possible that learners are over-applying a semantically present "zero article," but it is also plausible that they, more simply, are forgetting to use, or choosing to avoid, articles entirely.

Moving on to consider other factors, we note that abstractness and concreteness were not found to significantly affect article accuracy. This contradicts previous research, such as Amuzie and Spinner (2013), who found that concrete nouns encouraged accurate article use, and abstract nouns discouraged it. However, speakers were less likely to produce *a(n)* with concrete nouns than with abstract nouns. As of now, we can advance no explanation for this behavior, except that perhaps speakers used a high proportion of learned, abstract phrases with indefinite articles like *a lot*. Amuzie and Spinner (2013) had much more success finding statistically significant results by sorting their data into different levels of abstractness. Future studies should consider taking their approach, to sort out the concepts of abstractness and concreteness into more fine-grained semantic elements.

Plurality also plays a significant role in speakers' accuracy rates, with non-count and plural forms realized with approximately 10% higher accuracy than singular forms in our data. This also explains why articles that can be used in plural environments (*the* and 0) tend to be more accurate. The arbitrariness of certain mass/count noun distinctions (*a love* vs. *love*) may make judgments on the part of L2 speakers difficult or cause misalignment between L2 patterns and L1 patterns (our benchmark for "accuracy").

Also, plural nouns are much simpler to use with articles; they always take either 0, or *the*, and often either would appear grammatical to listeners. Examples 7 and 8 demonstrate this.

Example 7. When I was young, **0** men in my town proposed to me every day.

Example 8. When I was young, **the** men in my town proposed to me every day.

Both appear grammatical, even if they mean slightly different things. Using plurals might be another strategy used by learners of English to increase their surface grammaticality.

However, our data also show that speakers do have issues that go along with plurality. To explore this further, we present the following examples, which come from advanced mainland Chinese learners of English from the ICNALE corpus (CHN_PTJ1_008_B1_1 & CHN_PTJ1_039_B1_2).

Example 9. I agree that it is important for **0** college student to have **0** part-time job.

Example 10. I agree with it and if **0** students still can have **a** part-time job to make his ability and if you have **0** part-time job.

Example 9 refers to college students in general, however, the speaker does not use any plural morphology, such as a plural –*s*. It becomes unclear whether he is referring to one college student as an example, or multiple college students, because the 0 marking before "college student" and "part-time job" indicates plurality, while the lack of plural –*s* morphology indicates singularity. Example 10 talks about "students" having "a part time job." The plurality of the situation is lost halfway through the sentence, and the speaker switches to a singular. The speaker also talks about a hypothetical third person with the pronoun *his*. However, he later uses 0 to describe "part-time job," although the word does not have any plural morphology.

These speakers appear to have three problems with plurality. They misconstrue the plurality of a particular semantic scene, they switch pluralities of concepts throughout their utterance, and they assign 0 to nouns which have no other plural morphology.

The first two problems likely derive from Chinese's lack of a plural. In Chinese, plurality status need not be expressed on every word, as it must be in English. Therefore, Chinese learners of English might have more trouble remembering to mark plurality throughout their discourse, or in keeping that marking consistent. Their L1 does not prepare them for these communicative requirements of English.

The presentation of plural nouns with a 0 and no plural morphology calls for a different explanation. 0s are not just an absence of a form; for count nouns, they partially encode plurality. The word *cats* in the sentence "Genghis-Khan likes cats" is marked for plurality twice: by the plural –*s* attached to it, and by the absence of *a* before it. Therefore, each of the pluralization strategies is redundant. If speakers feel that they can express plurality with just one feature, they may neglect others.

In sum, it appears that plurality and articles interact in learner speech. Learners may use articles to mask their misunderstanding of the plurality of a noun (by using *the* which can be used across all plurality statuses). They may also drop –*s* morphology when plurality can be inferred by context, or by the zero article. Future work could usefully examine the correlation between zero article production and plural –*s* deletion in more detail.

4.5. The case of "some"

Participants in our data often headed noun phrases with the word *some*. In English an article as we have defined it cannot precede the word *some*. "*a some" and "*the some" are both ungrammatical collocations. Noun phrases containing the word *some*, or any other determiner, may then require less effort to process or prepare, as a speaker does not have to categorize the head noun within the complex Hearer Knowledge/Specific Reference paradigm.

While we have not included *some* in our quantitative analysis, we observe that speakers in our data appear to use *some* as an avoidance strategy. Example 11 presents an excerpt from a Chinese B1_1 speaker from ICNALE (CHN_PTJ_031_B1_1).

Example 11. If we choose **some** part-time job now, we only can choose those jobs such as waiter in restaurant and this is – these are not the professional jobs. It cannot give us **some** good – **some** help but this job can give us **some** experience of the society, so if we have spare time, I think we can choose **some** of the part-time job.

The speaker here makes several mistakes. He omits *a(n)* twice, and uses *the* incorrectly. This suggests that this speaker has not fully grasped the English article system yet. The word *some*, then, and likely other determiners, may provide a useful bulwark for L2 learners against the complexities of the English article system.

4.6. Names, pronouns, and determiners

Finally, we turn to a behavioral pattern of learners that generally goes unmentioned in the second language acquisition literature. Researchers can learn much from analyzing L2 mistakes. However, areas in which L2 speakers never make mistakes are also worth considering. English treats pronouns and proper names differently than other nouns. They often fill in for entire noun phrases, and thus do not take articles. Our participants were very good at following this rule. In the entirety of the data, which included thousands of examples of pronouns and names, speakers never once said anything like "*the I" or "*the France." Learners also comprehend the fact that English does not allow the use of two determiners to modify a noun. There are no examples of sentences like "**the* those trees," although logically, one could imagine a case where a speaker would wish to express such a concept.

Our participants avoid these mistakes. It might be because something—linguistic universals, or cross-linguistic characteristics of pronouns and determiners—prevents them from making these mistakes. It might also be that demonstratives, and noun-phrases, already contain markers of definiteness, and thus affixing an article would be redundant. Either way, future researchers should investigate not only conspicuous mistakes in L2 English, but also notable tendencies to avoid mistakes.

5. Conclusion

We began this study with the goal of exploring factors implicit in L2 English article production and to focus on differences across different proficiency levels. Learners' proficiency level and backgrounds were shown to have little effect on the usage of articles in these data. Indeed, even the highly advanced STAAAF speakers, showed only slightly more accurate uses of articles than the ICNALE L2 learners. On the one hand this is surprising but on the other we note that this is, actually, not inconsistent with other studies which have found fossilization in advanced learners (White, 2003). We might surmise that speakers acquire a level of article use that allows them to function, and then they cease to develop their internal grammars further in this area. Perhaps STAAAF and ICNALE speakers have both reached the zenith of their article systems; or, perhaps, the test scores used to determine proficiency differences in the ICNALE do not adequately correlate with mastery of the article system. Future researchers making use of the ICNALE should examine its proficiency ranking system.

Speakers tend to favor the grammatical constructions that can be used in more contexts. The indefinite article, *a(n)*, which was least accurately used, requires a fine-grained distinction between singular and non-count nouns, the knowledge of the hearer, and the idiomatic status of the utterance. *The* and 0 can be used across many more contexts, and so speakers use them more frequently.

Future research on articles should focus on the strategies learners use to simplify the complexities of the L2 grammar system. These include *the* flooding, and simplification tactics like using *some* rather than an article. They should also examine how features of L2 grammar interact with one another; for example, whether or not other speakers will use the 0 article to specify plurality without adding the plural *–s* morphology.

Teachers of English should also bear in mind the tendency toward simplification when teaching the English article system and should find ways of eliciting difficult grammar points from learners to better understand their progress. L2 speakers might master articles more quickly or more accurately if they thoroughly understand a broad range of factors that affect article production, such as plurality. Language patterns such as the *some* strategy and *the* flooding could potentially obfuscate student progress.

Our data has confirmed that speakers seek ways of simplifying the article system, and avoid articles altogether when possible. Future research can further illuminate how learners create these strategies, and what forms these strategies take.

Acknowledgements

First, we would like to thank the anonymous reviewers for their helpful suggestions, which contributed greatly to the quality of this paper. The STAAAF dataset was created thanks to funding by SPAWAR Atlantic/SSC Charleston (SSCC) BAA N65236-08-X-0164 (PI Rivers). We thank William Rivers, and team members and leadership at Integrated Training Solutions, Inc., for creating the STAAAF dataset and for allowing its use for this project. We would also like to thank Professor Shinichiro Ishikawa and his team for creating the ICNALE, and for his encouragement with this project. Professor Helen Southworth, Professor Connie Dickey, Trent Ukasick, Alexi Silva, and Hannah Theobald all provided invaluable feedback for earlier versions of this paper. Finally, thank you to Professor Vera Keller, for her encouragement and advice over the years.

References

Amuzie, G. L., & Spinner, P. (2013). Korean EFL learners' indefinite article use with four different abstract nouns. *Applied Linguistics, 34*(4), 415–434.

Baayen, H. R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge UP.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package (Version 1.1-7) [Software]*. Available from: http://CRAN.R-project.org/package=lme4.

Butler, Y. (2002). Second language learners' theories on the use of English articles: An analysis of the metalinguistic knowledge used by Japanese students in acquiring the English article system. *Studies in Second Language Acquisition, 24*(3), 451–480.

Butler, B. C. (2012). *A semantic map approach to English articles (a, the, and Ø)* (Unpublished doctoral dissertation). Eugene, OR: University of Oregon.

Chuang, F. Y. (2005). Article misuse: A neglected problem in Chinese EAP student writing. In *Proceedings of the 2nd international online conference on second and foreign language teaching and research* (pp. 25–33). The Reading Matrix, Inc.

Ekiert, M. (2004). Acquisition of the English article system by speakers of Polish in ESL and EFL settings. *Working Papers in TESOL & Applied Linguistics, 4*, 1–23.

Geng, J. (2010). The semantic analysis of the definite article misuse by Chinese learners of English. *Asian Social Science, 6*(7), 180–184.

Ghisseh, S. (2009). English article production by Arabic and French speakers. In M. P. García Mayo, & R. Hawkins (Eds.), *Second language acquisition of articles: Empirical findings and theoretical implications* (pp. 37–66). Amsterdam, NL: John Benjamins.

Goldschneider, J. M., & DeKeyser, R. M. (2005). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning, 55*(1), 27–77.

Hua, D., & Lee, T. H. (2005). Chinese ESL learners' understanding of the English count-mass distinction. In L. Dekytstopper, et al. (Eds.), *Proceedings of the 7th generative approaches to second language acquisition conference (GASLA 2004)* (pp. 138–149). Somerville, MA: Cascadilla Proceedings Project.

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

Ionin, T., & Montrul, S. (2010). The role of L1 transfer in the interpretation of articles with definite plurals in L2 English. *Language Learning, 60*(4), 877–925.

Ishikawa, S. (2013). *About the ICNALE corpus*. Retrieved from http://language.sakura.ne.jp/icnale/.

Ishikawa, S. (2014). Design of the ICNALE-spoken: A new database for multi-modal contrastive interlanguage analysis. *Learner Corpus Studies in Asia and the World, 2*, 63–75.

Ishikawa, S. (2017). *ICNALE: The international corpus network of Asian learners of English*. Retrieved from http://language.sakura.ne.jp/icnale/.

Jin, F., Afarli, T. A., & van Dommelen, W. A. (2009). Variability in the L2 acquisition of Norwegian DPs: An evaluation of some current SLA models. In M. P. García Mayo, & R. Hawkins (Eds.), *Second language acquisition of articles: Empirical findings and theoretical implications* (pp. 175–200). Amsterdam, NL: John Benjamins.

Kendall, T., Rivers, W., & Dodsworth, R. (2012). *Grouping speakers and assessing speaker groups: A case study of Chinese Americans in New York City*. Bloomington, IN: Paper presented at New Ways of Analyzing Variation (NWAV) 41.

Master, P. (1997). The English article system: Acquisition, function, and pedagogy. *System, 25*(1), 215–232.

Master, P. (2003). Acquisition of the zero and null articles in English. *Issues in Applied Linguistics, 14*(1), 3–20.

Rivers, W., & colleagues. (2009). *Sociolinguistic tool for the assessment of acculturation and allegiance factors*. Unpublished report. Arlington, VA: Integrated Training Solutions, Inc.

agliamonte, S. (2006). *Analysing sociolinguistic variation*. Cambridge, UK: Cambridge UP.

Tarone, E., & Parrish, B. (1988). Task-related variation in interlanguage: The case of articles. *Language Learning, 38*, 21–44.

Watcharapunyawong, S., & Siriluck, U. (2013). Thai EFL students' writing error in different text types: The interference of the first language. *English Language Teaching, 6*(1), 67–78.

White, L. (2003). Fossilization in steady state L2 grammars: Persistent problems with inflectional morphology. *Bilingualism: Language and Cognition, 6*, 129–144.

Young, R. (1996). Form-function relations in articles in English interlanguage. In R. Bayley, & D. R. Preston (Eds.), *Second language acquisition and linguistic variation* (pp. 135–175). Amsterdam, NL: John Benjamins.