Author/s:
Zhao, H;Shirai, Y

Title:
Frequency Effects in Chinese Learners' Acquisition of the English Article Construction

Date:
2022

Citation:
Zhao, H. & Shirai, Y. (2022). Frequency Effects in Chinese Learners' Acquisition of the English Article Construction. Chan, M (Ed.). Benati, AG (Ed.). Challenges Encountered by Chinese ESL Learners, (1), pp.237-263. Springer Nature Singapore.

Persistent Link:
https://hdl.handle.net/11343/313672

Frequency Effects in Chinese Learners' Acquisition of the English Article Construction

Helen Zhao and Yasuhiro Shirai

Abstract The current study, built on the usage-based approach to language, investigated frequency effects in Chinese learners' acquisition of English articles. We carried out type and token frequency analysis of article usages in academic written essays sourced from a written English corpus of Chinese learners. We adopted an existing usage-based article cue coding scheme, which allowed us to implement a refined frequency analysis of all form-function mappings in learner texts. Our findings suggested that learners' article usage follows the Zipfian distribution in terms of token frequency. Learners show a heavier reliance on a very limited number of high-frequency cues than native speakers. Non-definites (indefinite article and zero article) outnumber definite articles in terms of token and type frequency in learner texts. Yet definite articles show a significantly higher type/token ratio than non-definites, suggesting that learners develop a more complex and heterogeneous profile of definite article usage. We argue for more research and pedagogical attention to frequency and complexity effects in the acquisition of articles.

Keywords Usage-based · Construction learning · Frequency · English articles

1 Introduction

Despite the high input frequency, English articles (*a/an*, *the* and the zero article Ø) remain a feature generally acknowledged to be difficult for learners of English who come from an article-less background such as Chinese, Japanese, Korean and Slavic languages (Celce-Murcia & Larsen-Freeman, 1999). Teachers of English as a second language (L2) often find it difficult to understand how or why their students choose to use English articles the way they do. As a result, acquiring the article system remains

H. Zhao (✉)
The University of Melbourne, Melbourne, Australia
e-mail: helen.zhao@unimelb.edu.au

Y. Shirai
Case Western Reserve University, Cleveland, OH, USA

M. Chan and A. G. Benati (eds.), *Challenges Encountered by Chinese ESL Learners*,
https://doi.org/10.1007/978-981-16-5332-2_10

an elusive goal. In the current study, we focused on Chinese learners who have been extensively studied for the stages of their English article acquisition (Crosthwaite, 2016; Diez-Bedmar & Papp, 2008; Robertson, 2000; Thomas, 1989). We adopted the usage-based approach to language acquisition and investigated how frequency affected Chinese learners' acquisition of articles.

2 Adult Second Language Article Acquisition

The framework most widely adopted for analysing English articles in previous L2 article acquisition studies is Bickerton's (1981) semantic wheel framework. This model has two defining features: specific reference (SR) and hearer knowledge (HK). SR means whether the speaker refers to a specific referent, whereas HK means whether the hearer of an utterance has, or is assumed by the speaker to have, shared knowledge about the referent. Based on the two features, Bickerton specified four semantic types. Type 1 [-SR, +HK] is the generics category. The function of genericity can be expressed by the definite article (e.g. ***The*** *elephant is the largest land animal*), the indefinite article (e.g. ***An*** *elephant never forgets*) or the zero article (e.g. **Ø** *Elephants are the largest land animals*). The speaker does not refer to a specific object, while the hearer is assumed to have knowledge of the generic category. Type 2 [+SR, +HK] includes all the referential definites such as previous mentions (e.g. *I know* ***the*** *girl*), anaphoric referents (e.g. ***the*** *Harvard faculty*), uniqueness in all contexts (e.g. ***the*** *Sun*) and uniqueness in a given speech context (e.g. *Pass me* ***the*** *salt*). The referent is uniquely identifiable by both the speaker and the hearer. All the definite article usages, except for generics and some conventional uses, belong to this type. Type 3 [+SR, −HK] is the referential indefinite. These are the first mentions and can take both the indefinite article (e.g. *I repaired* ***a*** *car*) and the zero article (e.g. *I have been repairing* **Ø** *cars all day*). Only the speaker knows which specific object(s) is or are being referred to, while the hearer does not share the knowledge. Type 4 [−SR, −HK] includes all the non-referential nouns and is used with the indefinite article (e.g. *I need* ***a*** *car*) and the zero article (e.g. *I need* **Ø** *cars*). Neither the speaker nor the hearer presumably cares which specific object is being referred to.

The Bickerton model has motivated a strong body of research that has investigated the sequence of acquiring articles. Adult L2 learners have been observed to first associate specificity [+SR] with the definite article (Huebner, 1983; Thomas, 1989). First marking for specificity leads to learners using *the* for all Type 2 and Type 3 noun phrases. This can result in an overuse of *the* in Type 3 contexts where the hearer does not share the knowledge of the specific referent and, as such, *a* should be used, as in "Tom is visiting *a* boy from his class" (Butler, 2002; Ionin et al., 2008; Parrish, 1987). In contrast, learners are slower to account for the hearer's knowledge [±HK] (Butler, 2002; Ionin et al., 2008; Thomas, 1989).

The Bickerton model has also been applied to research that has a contrastive linguistic focus. Diez-Bedmar and Papp (2008) and Crosthwaite (2016) were both learner corpus studies that compared the sequences of article acquisition by learners of different L1s, but the two studies obtained very different results. Diez-Bedmar and Papp (2008) found that college-level Chinese learners of English (proficiency unspecified) showed significantly lower accuracy of article use than Spanish learners in written essays. Also, they concluded the hierarchy of accuracy for Chinese learners was Ø > *a* > *the*. In contrast, Spanish learners showed an overall significantly higher accuracy than Chinese learners and followed the acquisition sequence of *a* > *the* and Ø.

Crosthwaite (2016) reported much higher accuracy rates for Chinese learners' written performance in another learner corpus. He concluded that Chinese learners of low-intermediate to upper-intermediate proficiency in his sample had little trouble with article use and had equally good performance on definite, indefinite and zero articles. Korean and Thai learners who also come from article-less L1 backgrounds produced more errors and resembled the performance of the Chinese learners in Diez-Bedmar and Papp (2008).

Bickerton's semantic wheel was an effective model in guiding linguistic research of article analysis and acquisition. But the model also created issues for researchers to address. First, the four semantic types in the model are balanced in terms of the number of features [±SR, ±HK], but are not balanced in terms of the number of functions. Type 1 has only one function (generics), whereas Type 2 has many more functions. The elicited number of tokens in the four types also differs to a great extent (see Crosthwaite, 2016; Diez-Bedmar & Papp, 2008; Thomas, 1989). The imbalance creates problems for meeting statistical assumptions. Second, the model explains the use of many high-frequency usages but gives up a large number of idiosyncratic functions that reflect idiomatic and conventional article use in English (e.g. **Ø** *hand in* **Ø** *hand*, *in* ***the*** *1990s*, ***the*** *Mississippi River*, **Ø** *Michigan Lake*). Though these functions may seem peripheral, they constitute an indispensable part of the English article system and pose a serious challenge for L2 learners.

3 Usage-Based Approach to Language Acquisition

Almost none of the previous research works investigated frequency effects in article acquisition. Previous researchers have been primarily concerned with learners' accuracy of article use and have largely ignored learners' frequency of exposure to article use and their frequency of using article forms and functions. But frequency is found to be one of the most important predictors of language acquisition in the literature that follows the usage-based approach to language and language acquisition. Frequency effects in language use are typically shown in a Zipfian distribution (Zipf, 1935): Frequency is inversely proportional to its rank in the frequency table. The most frequent usages account for the majority of occurrences in a given category.

The Zipf's law has been consistently shown to be applied to linguistic constructions such as verb argument constructions (Ellis et al., 2016).

Frequency effects in language learning refer to the acquisition of linguistic knowledge based on cumulative experiences with language (Crossley et al., 2019; Ellis, 2002, 2006). Frequency of exposure from the input forms memory and interacts with the form-function associative learning mechanism in learner cognition. Form-function mappings are strengthened and consolidated with increased frequency in learners' input and interactional usage. High-frequency form-function mappings have the advantage of being acquired earlier than low-frequency constructions (Ellis & Ferreira-Junior, 2009). As L2 competence increases, learners gradually expand their use of L2 constructions to incorporate functions that are less prototypical and less frequent (Eskildsen, 2015; Eskildsen & Kasper, 2019).

Among other determinants of language acquisition under the usage-based framework, L1 transfer is another factor that is closely related to L2 article acquisition. As Ellis (2006) noted, "The L2 learners' neocortex has already been tuned to the L1, incremental learning has slowly committed it to a particular configuration, and it has reached a point at which the network can no longer revert to its original plasticity" (p. 184). The reduced plasticity of the brain interferes with the functioning of the associative learning mechanism, making it harder for adult learners to develop sensitivity to distributional probabilities of form-function mappings to the native level (Tachihara & Goldberg, 2020).

In the context of the current study, Mandarin Chinese does not have a system of articles that exists in English (Li & Thompson, 1981). Yet there is a widespread use of determiners which function in part to signify definiteness and indefiniteness. Robertson (2000) found that two distinct features in Mandarin were related to transfer phenomena that he observed in Chinese learners' use of English articles. First, the distal demonstrative *nà-* "that" and the unstressed numeral *yi* "one" begin to take on some of the functions of the definite and indefinite articles *the* and *a* in English, respectively (Huang, 1999; Li & Thompson, 1981):

1) 你 認識 不 認識 那個 人
nǐ rènshi bu rènshi **nèi**-ge rén?
You know no know that-classifier person
'Do you know the/that person?'

2) 他/她 買了 一個 帽子
tā mǎi-le **yi**-ge màozi
3sg buy-perfective one-classifier hat
'S/he bought a/one hat.'

(Robertson, 2000, p. 144)

Another productive structure in the grammar of Mandarin noun phrases (NP) is the "NP-*de*-NP" structure. Two NPs are chained by the particle -*de* which indicates associative relationship (especially possession) (Li & Thompson, 1981). The first NP is often a personal pronoun as in (3). It can also be an animate (3) or inanimate entity (4).

3) 我的 襯衫
wǒ-**de** chènshān
I-particle shirt
'my shirt'

4) 那個 飯店的 菜
nèi-ge fàndiàn-**de** cài
that-classifier restaurant-particle food
'the food of that restaurant'
(Robertson, 2000, p. 144)

Robertson (2000) proposed that Chinese learners had the tendency to transfer the two constructions to English productive use. He reported that some Chinese learners used the English demonstrative *this* and the numeral *one* as the definite and indefinite markers, respectively. He did not make specific observations about whether and how transfer might have been related to the "NP-*de*-NP" structure.

4 Frequency Analysis of English Articles

Master (2013) was one of the very few studies that examined frequency in English article usage. Master focused on the genre of research articles in science and technology. He analysed the token frequencies of the definite article *the*, the indefinite article (*a, an*) and the zero article (Ø) in a self-composed database of research articles published by native English-speaking authors. He found that the articles (*the*, *a/an*, Ø) accounted for the majority of determiners by a very large margin (90.3%). Of the articles, the most frequently occurring form in his database is the zero article Ø (51.2%), followed by the definite article *the* (37.8%) and the indefinite article *a(n)* (11.0%). He concluded that although *the* is the most frequent word, the zero article is in fact the most frequent free morpheme in the English language.

Master (2013) also analysed the frequencies of other types of determiners. Determiners are obligatory prehead structures in English noun phrases that generally serve the function that indicates a specification of definiteness and indefiniteness (Huddleston & Pullman, 2002). Determiners include articles, demonstratives (e.g. *this, those*), possessives (e.g. *our, my*), quantifiers (e.g. *all*, *many*), cardinal numerals

(e.g. *one, two*), assertives/nonassertives (e.g. *some*, *either*) and negatives (e.g. *no*). Non-article determiners are strong competitors to articles because these forms can also serve similar functions as articles. Many of them add a more refined semantic specification to definiteness and indefiniteness and thus are semantically "heavier" than articles. So far very few studies have compared usage-based properties (e.g. frequency of usage) of articles with other determiner types. Master's (2013) study made a valuable contribution to this research gap by investigating the frequency of distribution of various types of determiners in texts produced by native-speaking English expert writers. We used his findings of the distributional frequency as the L1 norm to compare with the L2 learner data in our study.

Zhao and MacWhinney (2018) was the only study that has applied the usage-based approach to a comprehensive investigation of English articles. In their analytical framework (MacWhinney, 2012), forms (e.g. *the*, *a*) compete for mapping onto functions (e.g. second mention). The winner forms in the competition (e.g. *the*) serve as *cues* for the activation of functions (e.g. second mention). For convenient naming of the form-function mappings, Zhao and MacWhinney (2018) called them article cues (e.g. cue "second mention|*the*", symbolising the conditional probability that a second mention interpretation will occur given the formal cue *the*).

Unlike many previous article acquisition studies (Butler, 2002; Huebner, 1983; Thomas, 1989) that adopted Bickerton's (1981) semantic wheel framework (±SR, ±HK) which treated the four semantic categories on an equal footing, Zhao and MacWhinney (2018) emphasised the importance of carrying out a more refined analysis of the form-function mappings in the English article system. From a usage-based perspective, they pointed out the inherent differences in the four semantic categories. For example, the generics category (−SR, +HK) has low token frequency and type frequency in English language usage, but has complex form-function mappings (i.e. the function of genericity mapped onto three forms *the, a,* Ø). The referential indefinites (+SR, −HK) and the non-referentials (−SR, −HK), in contrast, have very high token frequency, as these two meanings represent the most frequent article usages in English, but have low type frequency since these are mostly the few first mention cues that take either the indefinite article or the zero article depending on the countability of the head noun. The referential definites (+SR, +HK) have high token frequency and high type frequency because this category covers all definite article usages in English (except the generic *the*). There was an attempt, though not widely adopted, to extend the four-category framework by adding the fifth category which would include idioms and all idiosyncratic usages (Diez-Bedmar & Papp, 2008; Ekiert, 2004). This fifth category has relatively low token frequency but high type frequency as it contains a large number of context-specific form-function mappings. The semantic wheel framework has been very useful and powerful in accounting for some commonly observed learner errors while considering crosslinguistic influence on L2 acquisition. However, the four semantic categories represent four distinct usage-based profiles which make them not entirely comparable. The usage-based approach to article analysis advocated by Zhao and MacWhinney (2018) broke the categorical restriction of the semantic wheel and treated each form-function mapping

in the article system as an analytical unit. This approach greatly facilitates the identification and analysis of native speakers' and L2 learners' article usages, making it easier to predict and explain learners' specific difficulties in article acquisition.

Zhao and MacWhinney (2018) identified a full range of 86 article cues in the English language and did an L1 corpus analysis of written text (26,468 words) sampled from ten common genres (academic, encyclopaedia, magazine, newspaper, novel, drama, children's story, recipe, etc.) on a wide range of topic areas (politics, economy and finance, education, history, geography, technology, entertainment, sports, travel, food, etc.). Meanwhile, they computed the frequencies and reliabilities of the identified article cues following the corpus count method specified by the theoretical framework of the Competition Model (McDonald & MacWhinney, 1989). They found that native English speakers' use of article cues obeyed the Zipf's law (Zipf, 1935). The top ten cues with the highest token frequencies accounted for 76.3% of all the article tokens in their corpus sample.

5 The Present Study

In the current study, we adopted Zhao and MacWhinney's (2018) analytical framework of English articles and applied it to the investigation of frequency effects in second language learning. In line with the previous literature, we expect to observe frequency effects in L2 acquisition of the English article construction. We assume that the article cues that are used in learner production have emerged in their interlanguage system and have become available to them. Furthermore, we predict that learners should demonstrate increased knowledge of infrequent article cues as their L2 competence increases.

We adopted the corpus-based approach and analysed the L2 learner data from the written section in the Spoken and Written English Corpus of Chinese Learners (Version 2.0) (SWECCL) (Wen et al., 2008). SWECCL is one of the largest corpora for Chinese-L1 learners learning English as a foreign language (EFL). Learners' written texts in this corpus are collected from Chinese-speaking college students in 34 universities in mainland China. The sampling of the universities has a good coverage of geographic areas and of different rankings.

Same as most learner corpora, SWECCL does not offer quantitative data that can be used to indicate learners' L2 proficiency. However, it includes written data obtained from both English majors and non-English majors, both of which are included in the analysis of the current study. Although it is not necessarily the case that an English-major learner will have higher L2 competence than a non-English major, it is fair to assume that in the EFL context English majors have more exposure to the target language and more opportunities of producing the L2 than non-English majors. English majors in Chinese universities take skill-based and content-based courses with input materials and the medium of instruction all in English. They primarily produce written English essays for coursework. Non-English majors, on the other hand, have relatively limited exposure and use of English in college education. With

the above consideration, the current study assumes an overall higher L2 competence among English-major learners than non-English-major learners. Specifically, we aim at investigating the following research questions:

1. What is the frequency distribution of learners' usage of English determiners?
2. What is the frequency distribution of learners' usage of English articles? Specifically,
   (a) What article cues are used by learners?
   (b) Does learners' use of the article cues follow the Zipfian distribution?
3. Do English-major learners show more nativelike usage of articles than non-English-major learners? Specifically,
   (a) Do English-major learners use more infrequent article cues than non-English-major learners?
   (b) What are the type and token frequencies of English majors' and non-English majors' usage of the definite (*the*) and the indefinites (*a/an/Ø*)?

6 Methods

The majority of the written texts in the SWECCL corpus were argumentative essays based on prompts. The essay prompts are available in the Appendix. The corpus includes two types of text, timed and untimed, depending on whether students were given time restriction for the written task. Texts were initially collected from learners' handwritten documents and then were manually typed into digital form.

We included both English majors and non-English majors in our analysis. In the corpus, English majors' texts were available from students in Year 1 to Year 4 at college, whereas non-English majors' texts were only available from students in Year 1 and Year 2. For a fair comparison, we only analysed Year 1 and Year 2 essays from both majors. We randomly sampled approximately 20 texts from the timed essays in the four subgroups (English-major Year-1, English-major Year-2, Non-English-major Year-1 and Non-English-major Year-2). Timed measurements tend to elicit learners' implicit knowledge (Ellis et al., 2009) and are more likely to better reflect the status quo of learners' interlanguage development than untimed measurements. We only selected essays with more than 150 words. Many essays shorter than 150 words are found to be incomplete and lack a clear essay structure.

Four samples with a sum of 16,989 words were generated based on the above criteria (Table 1): English-major Year-1 (4707 words), English-major Year-2 (5858 words), Non-English-major Year-1 (3683 words) and Non-English-major Year-2 (2741 words). We adopted Parrish's (1987) and Tarone and Parrish's (1988) methods of coding all types of noun phrases (NPs) including articles, quantifiers, possessives and demonstratives. A total of 3004 noun phrases were identified as obligatory contexts for the use of all types of determiner, including articles and other non-article

Table 1 SWECCL data sample

	English majors		Non-English majors	
	Year 1	Year 2	Year 1	Year 2
Texts	18	21	19	17
Words	4707	5858	3683	2741
Words per text	261.50	278.95	193.84	161.24
NPs with all determiners	847	1085	626	446
NPs with all determiners per text	47.06	51.67	32.95	26.24
Quantifiers per text	4.50	5.14	2.58	2.12
Possessives per text	4.94	6.57	5.42	4.18
Demonstratives per text	2.06	1.76	1.05	0.82
Obligatory NPs for article use (token)	640	801	454	315
Obligatory NPs for article use per text (token)	35.56	38.14	23.89	18.53
Obligatory NPs for article use (type)	200	234	168	149
Obligatory NPs for article use per text (type)	11.11	11.14	8.84	8.76

determiners (quantifiers, possessives and demonstratives). 2210 tokens of article use (*the*, *a*, *an*, Ø) were identified.

The first author of the current article and a native English speaking research assistant manually coded all article tokens for (a) cue type and for (b) accuracy of usage in obligatory contexts (SOC). The two coders reached an interrater reliability of 0.86 after discussion and resolution of disagreements. We will only report frequency results in this article. The accuracy data is reported in a parallel study (Zhao & Fan, 2021).

Cue types were coded with a coding scheme consisting of 86 article cues developed by Zhao and MacWhinney (2018), which were extracted from descriptive grammar books (Celce-Murcia & Larsen-Freeman, 1999; Huddleston & Pullum, 2002; Quirk et al., 1985; and an ESL textbook focusing on articles, Cole, 2000). To illustrate the current coding, for the sentence "*So **the** children must learn how to compete to protect themselves*", the use of the definite article *the* is an error since the author intends it to be a general category of children rather than refers to a specific group of children. Here, *the* was coded as an error token of the cue "*plural*\Ø" in the coding scheme (Use Ø with plural nouns unless they are uniquely identifiable). More examples of article cues will be discussed below; see especially Table 3.

Certain forms were excluded from analysis. When there are two parallel NPs, both of them are coded when there is no involvement of non-article premodifiers such as possessives or quantifiers. For example, in the phrase "a lot of troubles to **college** and **society**", both "college" and "society" were coded. Both of them are considered as obligatory contexts for article use. But in the phrase "for **your commanders** or **commercial partners**", only the first NP "commanders" was coded for possessive use. The second NP "commercial partners" was excluded from our coding, since we

cannot judge whether the zero article was used due to the use of the possessive (*your*) or due to the cue *plural|Ø*.

We also excluded the erroneous forms that invite ambiguous interpretations. For example, the NP "*foreigner"* in the sentence "I think communicating with **foreigner** is the thing you really want to do" was excluded since it is most appropriate to interpret this error as a morphological error due to the omission of the plural marker *-s*. However, the error could also be interpreted as an omission error of the indefinite article *a*. Such NPs were excluded to avoid ambiguous interpretations in coding. The errors related to misuses of parts of speech were also excluded from coding. For instance, the NP "*independence*" in the sentence "We can learn to be **independence** in universities" was a grammatical error since an adjective (*independent*) rather than a noun was required in the slot. Similarly, we also excluded coding on the adjective "*healthy*" in the sentence "The good **healthy** for them are very important." Gerunds were also excluded from coding.

We distinguished between *tokens* of article cues, counting all the tokens of an article cue, and *types* of article cue, counting only one instance of each article cue which occurred in the text. That is, when multiple forms of the same article cue (e.g. *plural|Ø*) occurred in a text, such as *children*, *schools* and *companies*, we counted the token frequency as 3 and the type frequency as 1.

We grouped the analysis of the indefinite article (*a*, *an*) and the zero article (Ø) as one category of non-definite articles. Only three indefinite article cues were identified in the learner texts: "*singular countable|a/an*" (Use *a/an* when the singular countable noun is not made concrete or instantiated by any modifier); "*positive 'few' or 'little'|a/an*" (Use *a* with words "few" or "little" expressing a positive meaning); and "*a XX of|a/an*" (Use *a/an* with structures like "a number of", "a handful of" and "a pair of"). Only the cue "*singular countable|a/an*" had a reasonable size of coded tokens. The other two *a/an* cues had a very small token size. It is not fair to statistically contrast the indefinite article (*a/an*) cues with the cues of the definite article and of the zero article in terms of token and type frequency. Therefore, we grouped the indefinite article cues with the zero article cues as the non-definite article cues (*a/an/Ø*).

When analysing results, we compared learners' frequency patterns to native speaker patterns. For the comparison of frequency distribution of determiner use, we compared the learner results to the native speaker data reported in Master (2013) since this is the only study that has provided the most comprehensive examination of determiner use frequency in English academic texts. Regarding article cue distribution, we compared our learner results to the native speaker pattern reported in Zhao and MacWhinney (2018).

7 Results

7.1 Frequency Distribution of the Determiners

Table 1 presents the overall determiner usage (including articles) in the learner essays. The average percentage of determiners per number of words was 17.7. This percentage is comparable to 18.3 for the published research articles written by native English speakers reported in Master (2013). The articles *the*, *a(n)* and Ø in the learner production accounted for the majority of determiner use (73.6%), though this proportion is much smaller compared to 90.3% reported in Master (2013). The reduced slice for article tokens in our sample was taken by other determiners including possessives (13.3%), quantifiers (9.1%) and demonstratives (3.6%).

Possessives had a surprisingly high frequency in the learner corpus. Possessives accounted for only 2.4% of determiner use in Master's (2013) analysis in which *their* was reported as the most frequently used possessive determiner in published academic texts (47% of total possessive use). The most frequently used possessives in our data were *our*, *their*, *my* and *your*. We analysed the percentage of each possessive out of the total number of possessive use in each learner group and obtained the following patterns of results (Table 2). *Our* accounted for the majority of possessives by a large margin among both Year 1 and Year 2 groups of non-English majors. Non-English majors heavily relied on first person possessives (*our*, *my*) in their argumentative essays. English majors used the plural form of third person possessive (*their*) much more frequently than non-English majors. English majors also used a higher percentage of *your* in the writing.

The second largest category of non-article determiners in our data is quantifiers which account for 9.1% of determiner use. Similar to the case of possessives, the proportion of quantifiers in our data was drastically higher than its proportion in the published research articles (2.06%) reported by Master (2013). In our data, *some* was found to be the most frequently used quantifier. Master (2013) differentiated two forms of *some* as a determiner: an unstressed form (e.g. *Medical care is worse in some poor villages*) that indicates an indeterminate amount, and a stressed form (e.g. *Some people think …; but in my opinion…*) meaning "certain unidentified" referent (Greenbaum & Quirk, 1990, p. 74). Both forms of *some* appeared in our data, with the unstressed form being the more frequently used form of *some*. *Many* was the second most frequently used quantifier in our data, though it was not reported in Master's (2013) analysis of published research articles.

Demonstratives occupied 3.6% of determiner use in our data, which is slightly lower than the proportion (4.5%) reported in Master (2013). *This* was reported as the most frequently used demonstrative in Master's analysis. In our data, *this* was also the most frequently used demonstrative. But there was not a clear pattern of results regarding the frequency distribution of demonstrative use in the learner groups. This was due to the limited number of demonstrative tokens (see Table 1) that appeared in the learner's essays.

Table 2 Percentages of possessives, demonstratives and quantifiers by learner group

Learner group	Percentages of possessives
English-majors year-1	*Their* (30.7%) *our* (29.5%) *my* (12.5%) *your* (11.4%)
English-majors year-2	*Their* (37.7%) *our* (25.4%) *my* (10.1%) *your* (8.7%)
Non-English-majors year-1	*Our* (58.3%) *my* (19.4%) *their* (4.9%) *your* (1%)
Non-English-majors year-2	*Our* (46.5%) *my* (15.5%) *their* (12.7%) *your* (5.6%)
Master (2013) native data	*Their* (47%) *its* (35.1%) *our* (15.1%) *his* (1.6%)
	Percentages of quantifiers
English-majors year-1	*Some* (28.4%) *many* (21.0%) *other* (11.1%) *every* (4.9%)
English-majors year-2	*Some* (33.0%) *many* (16.5%) *every* (8.3%) *no* (6.4%)
Non-English-majors year-1	*Many* (32.0%) *some* (20.0%) *all* (12.0%) *one* (10%)
Non-English-majors year-2	*Some* (38.9%) *one* (13.9%) *many* (11.1%) *another* (5.6%)
	Percentages of demonstratives
English-majors year-1	*This* (62.5%) *those* (16.7%) *these* (12.5%) *that* (8.3%)
English-majors year-2	*Those* (45.2%) *these* (25.8%) *that* (16.1%) *this* (12.9%)
Non-English-majors year-1	*This* (69.2%) *these* (23.1%) *those* (7.7%)
Non-English-majors year-2	*This* (44.4%) *that* (22.2%) *those* (33.3%)

7.2 Frequency Distribution of Article Cues

Out of the total number of 86 article cues in the coding scheme (Zhao & MacWhinney, 2018), 42 article cues were observed in the learner essays. Despite the fact that half of the L1 cues did not appear in the L2 texts, the frequency distribution of L2 article cues is Zipfian, similar to the Zipfian distribution of L1 article cues. The most frequent article cues overall accounted for the majority of all the tokens. Figure 1 plots the token frequency distribution of all the 42 article cues identified in the learner texts as cue frequency (Y axis) against frequency rank (X axis). Since the frequency ranks in different learner groups varied, we used the rank in the Year-1 English-majors group as the benchmark to plot Fig. 1. The Year-2 English majors have the sharpest Zipfian pattern of distribution compared to the other three groups.

The overall trend of all the learner groups is that the cues with high L1 frequency also have high L2 frequency. Table 3 lists the token frequency of all the article cues

Table 3 Number of tokens for article cues in learner groups

L1 rank	Cue	Example	English-major year 1	English-major year 2	Non-English-major year 1	Non-English-major year 2
1	*Plural|Ø*	**Ø** students	152	210	67	42
2	*Non–countable|Ø*	**Ø** water	136	158	139	92
3	*Singular countable with post-modifiers|the*	**the** man she is dating	19	29	13	13
4	*Singular countable|a/an*	**a** Shakespearean drama	80	97	78	67
5	*Preposition-modifier non-uniqueness|a/an/Ø*	I need **a** translator with more experience.	17	21	19	19
6	*Plural with post-modifiers|Ø*	**the** letters I received today	9	9	4	5
7	*Part of|the*	I'm returning this coat for a refund. **The** zipper broke.	23	15	16	3
8	*Second mention with variation|the*	I saw a peacock at the zoo. **The** bird was beautiful.	9	13	2	4
9	*Clause-modifier non-uniqueness|a/an/Ø*	Help me find **a** word that fits in this sentence.	17	14	3	3
10	*Second mention|the*	I saw a peacock. **The** peacock was beautiful.	17	36	9	5
11	*Names of countries, cities or states|Ø*	**Ø** Australia	11	7	0	1
12	*Non-countable with post-modifiers|the*	**the** wealth of her parents	20	37	12	10

(continued)

Table 3 (continued)

L1 rank	Cue	Example	English-major year 1	English-major year 2	Non-English-major year 1	Non-English-major year 2
13	*Ranking words\the*	**the** first, **the** beginning	6	12	9	4
14	*Superlative\the*	**the** best	11	7	8	5
15	*Uniqueness\the*	**the** Sun, **the** Moon	9	5	16	7
17	*Anaphoric reference in phrase\the*	**the** Black Sea Coast, **the** Harvard faculty	4	8	4	1
20	*Specific collectives of people\the*	**the** Republican Party	1	13	0	0
21	*Habitual location\Ø*	go to **Ø** school	35	7	14	24
22	*Time of the day/week/season\the*	in **the** morning, in **the** summer	1	1	0	0
24	*Historic periods\the*	**the** 1990s	0	1	1	1
25	*Obvious physical place\the*	**the** airport, **the** beach	1	2	0	1
26	*Political and military institution used alone\the*	**the** air force, **the** Department of Commerce	1	7	0	0
29	*Disease name\Ø*	**Ø** cancer	0	2	0	0
30	*Ranking words for time\Ø*	**Ø** next month	1	0	1	0
31	*Comparative \ Ø*	**Ø** more years	18	14	6	5
32	*Positive "few" or "little"\a/an*	**a** few friends, **a** little time	0	1	1	4
33	*A XX of\a/an*	**a** handful of	5	9	7	1

(continued)

Table 3 (continued)

L1 rank	Cue	Example	English-major year 1	English-major year 2	Non-English-major year 1	Non-English-major year 2
34	*Generic animals|the*	**the** elephant	0	0	0	2
40	*Company names used alone|Ø*	**Ø** Microsoft	1	0	0	0
41	*Language, religion|Ø*	**Ø** English, **Ø** Buddhism	0	12	2	2
42	*Xx university|ø*	**Ø** Yale University	4	0	0	0
43	*Situational uniqueness|the*	Jimmy, get your feet off **the** table!	3	2	4	0
48	*Generic inventions|the*	**the** computer	2	15	5	2
52	*Abstract adjectives for people|the*	**the** poor, **the** disadvantaged	1	6	1	1
57	*Same|the*	**the** same song	5	3	2	1
59	*Past, future|the*	**the** past, **the** future	6	3	7	1
61	*Negative "few" or "little"|Ø*	**Ø** few choices, **Ø** little doubt	0	0	0	1
62	*Body parts|the*	He got hit in **the** eye	0	0	0	2
63	*Modifying words in phrase|the*	in **the** year 2018	1	1	1	1
67	*Double comparative|the*	**the** more, **the** better	0	0	0	2
82	*Ranking words for prizes|Ø*	**Ø** first prize	0	1	0	0
85	*Singled out words|Ø*	How do you spell **Ø** "cat"?	0	1	0	0

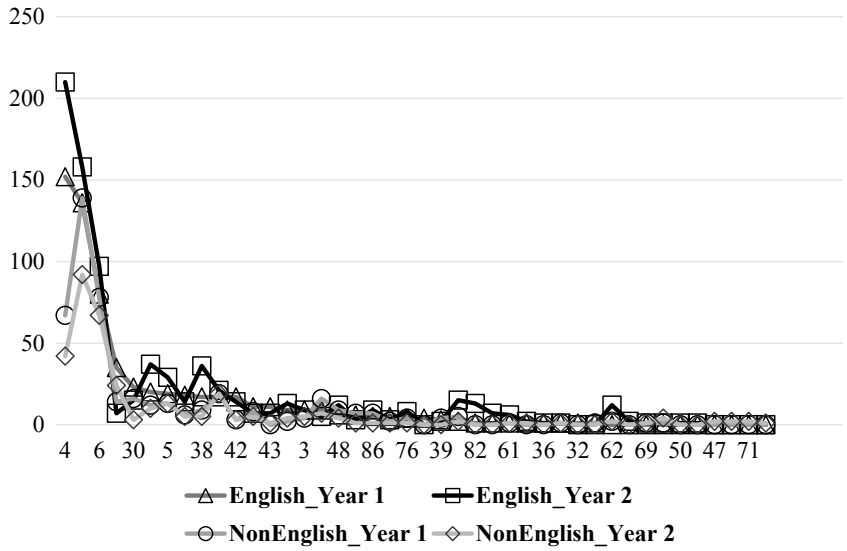

Fig. 1 Frequency distribution of article cues in learner groups

identified in the learner essays. To facilitate comparison across groups, the cues are ranked according to their L1 frequency ranks (Zhao & MacWhinney, 2018).

The three cues with the highest L2 token frequency in Table 3 are *plural*|Ø, *non-countable*|Ø and *singular countable*|*a/an*, ranking 1, 2 and 4 in L1 frequency. The English majors followed the L1 frequency rank (1 > 2 > 4), whereas the non-English majors deviated from the L1 rank (2 > 4 > 1) (see Table 3). The non-English majors in both years produced more non-countable nouns (or mass nouns) in the essays than plural nouns. For example, non-English majors used a large number of abstract nouns in their essays (e.g. *knowledge*, *education*, *environment* and *success*), which require the zero article cue *non-countable*|Ø.

Among the top 10 article cues in the L1 frequency rank (Table 3), there are two cues that do not rank as high in the L2 frequency rank: *part of*|*the* and *second mention*|*the*. The partonomy cue (*part of*|*the*) describes bridging relations between new entities and a previously mentioned entity in the discourse. The two entities can be linked via lexical relation through synonymy, hyponymy, meronymy, or thematic roles and through the interlocutors' shared pragmatic and world knowledge. For example, in the sentence "*I bought a car, but* ***the handle*** *is broken*", the new entity "*handle*" (i.e. steering wheel) is a constituent part of the old entity "*a car*" and is therefore registered for its unique identifiability in the discourse. This partonomy cue ranks sixth in the L1 frequency ranking but is used relatively less frequently in our L2 data. Similarly, the second mention cue (*second mention*|*the*) was also pushed out of the top 10 in the L2 frequency ranking.

Idiosyncratic cues, despite its overall low L1 token frequency, were extensively tallied for the various types in L1 data in Zhao and MacWhinney (2018). However, they were rarely found in our L2 data. Even some of the relatively more frequently used idiosyncratic cues were rarely used or did not appear in the L2 texts. These cues include "*collective group names*|Ø" (e.g. *The Republican Party*); "*profession as*

identifier|Ø" (e.g. *Actor Brad Pitt*); "*historic periods|the*" (e.g. *the 1990s*); "*time of the day/week/season|the*" (e.g. *in the morning/summer*); "*directional terms → the*" (e.g. *to the north*). This is partially due to the nature of argumentative writing as the genre of the current L2 learner data. Zhao and MacWhinney's (2018) L1 frequency analysis of article cues was based on samples of ten written genres on a variety of topic areas. It could be that the above idiosyncratic cues might have very low L1 frequency when the genre of academic writing was singled out from their L1 text analysis. If this is true, the weak presence of idiosyncratic cues in L1 English academic texts would likely pose a challenge for L2 learners to acquire this category of article usage, as L1 English academic genre may well constitute a major source of input to learners. Another likely account of restricted idiosyncratic cue use in the learner data is that the limited discourse boundary set by the essay topics (Appendix) fails to create semantic needs for learners to produce some of the above idiosyncratic cues.

7.3 Frequency Effects in Learner Groups

We expected to find frequency effects in construction learning to be applied to L2 article acquisition. We hypothesised that the English-major learners would demonstrate stronger knowledge of infrequent article cues than the non-English-major learners. The pattern of our results was in compliance with this prediction. The English majors produced a larger variety of article cues than the non-English majors (see Table 3). The numbers of cue types produced by the learner groups were: Year-1 English majors (32 types), Year-2 English majors (35 types), Year-1 non-English majors (28 types) and Year-2 non-English majors (32 types).

The Year-2 English majors, in particular, used the largest amount of infrequent article cues among the four learner groups. They produced quite a number of tokens for three cues ("*language, religion|Ø*", "*generic inventions|the*" and "*abstract adjectives for people|the*") that rank low in the L1 frequency ranking (ranked 41st, 48th and 52nd, respectively, see Table 3). They were the only group that used the cues "*ranking words for prizes|Ø*" and "*singled out words|Ø*" which are rarely observed in L1 texts (ranked 82nd and 85th, respectively). They produced an observable number of tokens for the cues "*specific collectives of people|the*" and "*political and military institution used alone the*", which ranked 20th and 26th in the L1 frequency ranking but were barely used by the other three learner groups. In contrast, the Year-1 non-English majors produced the smallest amount of cue types and used very few infrequent cues.

In short, there is evidence suggesting that the English majors demonstrated increased knowledge of infrequent article cues than the non-English majors. Based on this finding, we infer that the overall stronger L2 competence allowed the English majors to expand their article usage from the more frequent cues to the less frequent ones. Meanwhile, regardless of majors, the Year-2 students produced more cue types

than the Year-1 students. But this contrast was not as clear as the contrast between the two majors.

7.4 *Definite Article and Non-definite Article Cues Used by Learner Groups*

Among the 2210 article tokens in our sample, we identified 558 *the* tokens (25.2% of the total number of tokens) and 1652 *a/an/Ø* tokens (74.8% of the total number of tokens). These two proportions were somewhat comparable to the proportions reported in Master's (2013) analysis of published research articles: *the* (37.8%) and *a/an/Ø* (62.2%). Note that the higher percentage of non-definite article tokens in our data was not due to omission or erroneous use of the definite article in obligatory contexts. Unsupplied or inaccurate *the* tokens were coded into *the* cues in our coding method.

In terms of token versus type frequency, we observe very different patterns in the distributions of *the* and *a/an/Ø* cues (see Figs. 2 and 3). *A/an/Ø* cues had significantly higher token frequency than *the* cues by a large margin. The contrast was much smaller when considering type frequency. We treated each learner text as one participant and ran three separate 2 Form (*the* vs. *a/an/Ø*) × 2 Major (English vs. Non-English) × 2 Year (one vs. two) repeated measures ANOVA with token frequency, type frequency and type/token ratio as dependent variables.

The ANOVA on token frequency yielded the main effects of article form [$F(1, 71) = 191.84$, $p < 0.0001$, $\eta^2 = 0.73$] and major [$F(1, 71) = 62.24$, $p < 0.0001$, $\eta^2 = 0.47$]. The form × major interaction was also significant, $F(1, 71) = 9.05$, $p = 0.004$, $\eta^2 = 0.11$. The mean token frequency of *a/an/Ø* cues (mean = 21.91) was significantly higher than that of *the* cues (mean = 7.39); the English majors produced significantly more article tokens per text (mean = 18.47) than the non-English majors (mean = 10.82). The form × major interaction was due to the larger contrast between

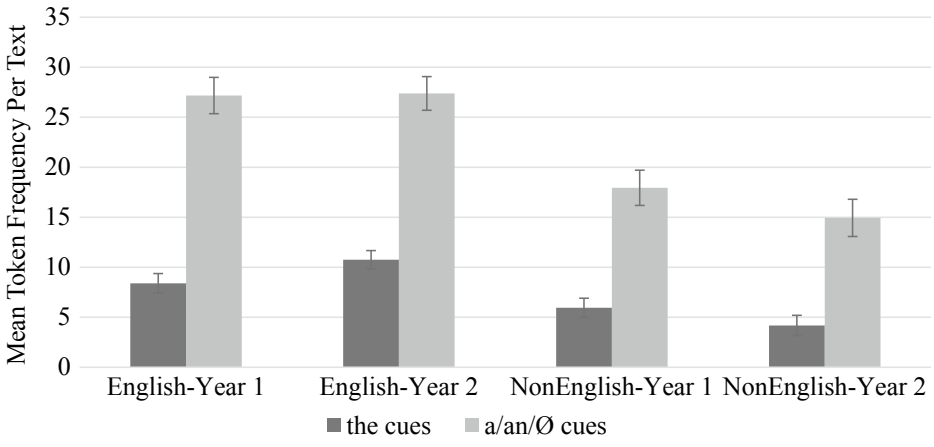

Fig. 2 Mean token frequency of the and a/an/ø cues per learner text

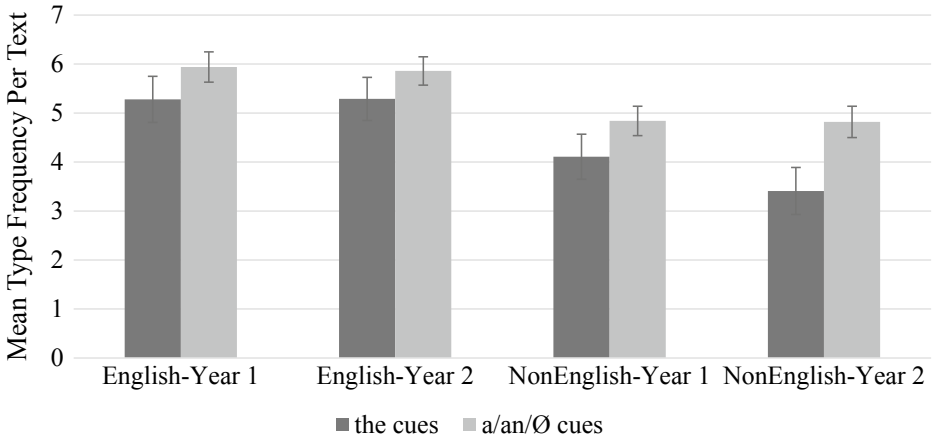

Fig. 3 Mean type frequency of the and a/an/ø cues per learner text

English majors and non-English majors in the production of *a/an/Ø* tokens than in *the* tokens.

The analysis on type frequency only revealed a main effect of article form, $F(1, 71) = 10.12$, $p = 0.002$, $\eta^2 = 0.13$, and no other main effects or interactions. The mean type frequency of *a/an/Ø* cues (mean = 5.37) was still significantly higher than that of *the* cues (mean = 4.53), though with a much smaller effect size than that of token frequency.

The type/token ratio analysis showed a main effect of article form, $F(1, 71) = 189.54$, $p < 0.0001$, $\eta^2 = 0.78$, and no other significant results. The mean type/token ratio for *the* cues (mean = 0.69, standard error = 0.03) was significantly larger than that of *a/an/Ø* cues (mean = 0.28, standard error = 0.01). None of the above analyses indicated a significant effect of the year of study in college.

8 Discussions

8.1 Evidence for Avoidance and Transfer?

Compared to the percentage of article tokens in the overall determiner use (90.3%) in Master (2013), the percentages of article tokens in the learner essays in the current study (73.6%) are much lower. The learners used a significantly higher proportion of other types of determiner (possessives 13.3%; quantifiers 9.1%; demonstratives 3.6%) as premodifiers for noun phrases. This contrast has to be carefully interpreted since the L1 texts that Master analysed are science and technology related research articles, whereas the L2 texts in our analysis are argumentative essays about general societal issues and topics. Nevertheless, what we observed is that the Chinese EFL learners used less amounts of articles for referentiality in written essays than the amount used by L1 expert academic writers. Does this finding constitute evidence

for learners' avoidance of using articles? Our interpretation is that the Chinese EFL participants made use of a variety of referential resources to compensate for their insufficient knowledge about articles. Some of their heavier reliances on non-article determiners can be evidence of L1 transfer.

Robertson (2000) argued for L1 transfer based on his findings that Chinese learners used demonstratives (particularly *this*) to replace the definite article and used the numeral *one* as the replacement for the indefinite article. But Robertson's transfer argument was made based on his observation of samples in his qualitative data. He did not report the actual frequency of occurrences of the argued substitutions (*this* for *the*; *one* for *a/an*). In our results, all the demonstratives accounted for 3.6% of all determiner use, which is comparable to the proportion of non-article determiners in L1 research articles in Master (2013). Similarly, we did not observe an obvious overuse of the quantifier *one* in our sample, contrary to Robertson (2000). As a result, we did not find robust quantitative evidence in support of lexical transfer that Robertson argued for in his study.

Instead, what may be regarded as stronger evidence of transfer in our findings is the high proportion of possessives in the overall distribution of determiners. This transfer can be traced to the very productive use of the particle *-de* in Mandarin noun phrases. This frame of "NP-*de*-NP" in Mandarin could be easily transformed into the "possessive NP" structure in English, thus enabling learners to transfer a productive L1 constructional frame. Luk and Shirai (2009) argued that because of the similarity between L1 and L2, possessive *–s* is acquired earlier than predicted by Krashen's Natural Order by Chinese, Korean and Japanese learners of English. This is consistent with the present findings.

It is noteworthy that the non-English majors in our sample primarily relied on first person possessive pronouns (particularly *our*) for determiner use, whereas the English majors used the third person possessive pronoun (*their*) more frequently. At some point, the English majors were more proficient at projecting a more objective voice by distancing themselves from the topics under discussion. The non-English majors (particularly in Year 1) seemed to be heavily influenced by a Chinese discursive pattern that emphasises the collective "we" that engenders solidarity (Diani & Bison, 2004). Such collective discourse centres on the value of family and self while emphasising the differentiation between *us* and *them* as the two opposite ends of a dichotomy (Zhou & Yang, 2018). The following excerpt comes from an essay written with high-frequency use of the possessive *our* typical of the Year 1 non-English majors. The collective voice in discourse *demands* high use of the possessive and consequently fewer uses of articles.

> *Firstly, universities offer us more stages to show* ***our*** *talents, which will add* ***our*** *self-confidence. Thirdly, we can broad* ***our*** *eyes on evreything, such as making friends, adapting to changeble environment and so on. Forthly, we can improve* ***our*** *abilities through studing in universities, because we will meet with a lot of problems in* ***our*** *studies and lives, we must solve them by ourslves.* (Note: The excerpt is selected from the original learner data and contains grammatical and spelling errors.)

8.2 Definite Article and Non-definite Articles

The non-definite article cues outnumbered the definite article cues in terms of token frequency and type frequency. The contrast was significantly larger for token frequency. This was within expectation. The top two most frequent cues in the English language are zero article cues (Zhao & MacWhinney, 2018): *plural*|Ø and *non-countable*|Ø, and jointly account for 27% of the entire article tokens in the L1 texts analysed therein. The Zipf's law seems to be magnified in the learners' article use. The two cues mentioned accounted for 43.3% of all the article tokens in our data. The learners heavily relied on the most frequent cues for their own usage. This pattern is roughly the same in each of the four learner groups. In fact, the Year-2 English majors showed the sharpest Zipfian pattern of distribution compared to the other three groups and used the two most frequent cues that accounted for 45.2% of all their article tokens.

The definite article cues outnumbered the non-definite article cues in terms of type/token ratio. Given the same amounts of tokens, the participants used significantly more types of *the* cues than types of *a/an/Ø* cues. What contributed to this result was that a very small number of *a/an/Ø* cues (the top three) accounted for almost all the tokens of the non-definite cue category and the rest of the *a/an/Ø* cues had very few tokens. But no definite article cues accounted for the majority of tokens in the definite cue category. The token numbers were relatively more widely spread out among *the* cues. In other words, the learners constructed a more distributed, heterogeneous and "adventurous" profile in their definite article usage. Among the infrequent cues that the learners produced (see Table 3), the majority of them were *the* cues. The learners took more risks of trying out infrequent *the* cues in their written output.

8.3 Language Competence and Frequency Effects

The L2 learner sample in the current study, English majors and non-English majors alike, showed a Zipfian frequency distribution in their L2 article usage, similar to the distributional pattern in L1 article usage (Zhao & MacWhinney, 2018). The finding confirms that L2 production frequency is intimately tuned to input frequency (Ellis & Ferreira-Junior, 2009). Zhao and Fan (2021), which is a parallel study to the current one, adopted the analytical method of structural equation modelling and further corroborated this finding.

The English majors in our sample showed an overall stronger level of written performance than that of the non-English majors. In a timed test situation, the English majors produced significantly longer written texts with a much denser use of noun phrases that take articles or other types of determiners. The English majors also produced a larger variety of article cues than the non-English majors. The English majors demonstrated a stronger capacity to expand article usage from more frequent types to less frequent ones, which aligns with previous empirical findings of the

usage-based approach to L2 acquisition (Eskildsen, 2015; Eskildsen & Kasper, 2019). The usage-based approach assumes a bottom-up exemplar-based learning process in which high-frequency exemplars that have a strong association with the target construction play a decisive role in helping learners identify structural regularities in construction use and formulate functional understanding of the construction (Ellis, 2002). High-frequency exemplars allow learners to learn faster and build up an abstract prototype of the construction so that they can generalise the schemata of the construction use and extend it to non-prototypical exemplars or novel uses at the later stage of construction learning (Goldberg, 1995, 2006; Sung & Kim, 2020).

Zhao (2020) reported that higher proficiency Chinese EFL learners demonstrated a stronger competence to differentiate form-function mappings in the English article construction than the lower proficiency learners at college level. The Michigan Test of English Language Proficiency (MTELP) was used to differentiate proficiency levels. Given the very complex form-function mappings in the article system (4 forms and at least 86 functions according to Zhao & MacWhinney, 2018), the task of learning the English article construction includes schemata development based on high-frequency prototypes and more importantly a refined analysis of the distributional characteristics of language input. In this process, learners analysed cue distinctions and drew analogies among form-function mappings based on semantic and structural similarities (and distinctions) observed from the cues in the input. Meanwhile, learners developed contextualised constructional knowledge (i.e. they learned that functional meanings such as definiteness or countability are in fact context dependent). They needed to look for other distributional features in discourse in order to generate a more accurate analysis of form-function mapping. Therefore, the process of moving from prototypical cue use to less frequent cue use is all part of this process of association, generalisation, analysis, differentiation and category formation.

Finally, in contrast with the observed distinction between the two majors, the year of study did not yield a significant difference in terms of frequency effects. It is not surprising to have this finding for the non-English majors. Because the amount of English exposure is rather limited for non-English majors in Chinese universities. Many of these students rely on their College English classes as the main source of English exposure. Year-2 non-English majors do not necessarily receive more hours of college English training than Year-1 non-English majors. In comparison, we did expect a significant effect of the year of study among the English majors, but the results suggested otherwise. Our speculation was that many programmes for English majors in Chinese universities would be arranged in the first two years of undergraduate studies for English language skills and in the final two years for the more advanced translation and interpretation skills and discipline-specific knowledge (literature, cultural studies, linguistics, translation, etc.). The respective curriculum arrangements for Year-1 and Year-2 English majors, which would indicate the amount of input exposure to students, may not be significantly different.

9 Conclusion

To our knowledge, this is the first study that has investigated frequency effects in L2 acquisition of English articles. Our findings suggest that both frequency and L1 transfer play important roles in influencing learners' article use. Chinese learners "avoid" using articles to a certain extent, which is suggested by a denser use of possessives and quantifiers compared to that of native English academic writers. Their use of possessives and quantifiers suggests traces of lexical and discursive transfer from the L1 discursive convention. Regarding article use, we identified 42 form-function mappings (i.e. cues) in our sample of Chinese learners' college essays. The use of the 42 cues follows the Zipfian distribution, with a heavier proportion of tokens accounted for by few top-ranking cues in the frequency rank. The English majors demonstrated increased knowledge of infrequent article cues than the non-English majors. We conclude that more exposure to English and more opportunities to use the target language allow the English majors to extend their article usage from more prototypical cues to less frequent ones.

The study generates important pedagogical implications for article learning. First, only 42 article cues out of 86 were observed in the learner essays. What about the other half set of cues? A large number of unused cues were low-frequency cues including idiosyncratic and conventional usages of articles (e.g. *geographic features*|*the*; *political/military institution*|*the*; *construction names*|*the*) that may not frequently appear in the genre of academic essays unless they are for specific topics such as politics and geography. Meanwhile, learners may not be exposed to such cues that often, apart from not having enough chances to produce these cues. Nonetheless, these low-frequency cues are an indispensable part of the English article system and will become more important for learners to acquire at an advanced stage of construction learning. It seems clear that the current academic writing practices for English education at college in the Chinese EFL context have not provided a sufficient ground for usage and feedback on these low-frequency cues, thus hindering the ultimate attainment of the English article construction. Instruction can aim to increase students' exposure to these cues in authentic language input of diverse genres and topic areas and create more opportunities for them to extract structural regularities and make generalisations.

Second, there appear to be apparent gaps between non-English majors and English majors, as well as between L2 learners and native speakers with regard to certain high-frequency article usages. For example, non-English majors find the cue of "*plural*|Ø" more difficult to apply compared with English-majors and native speakers, while L2 learners in general find it harder to apply the cue of "partonomy" than native speakers. These are high-frequency cues and should have been well-acquired after years of L2 learning. More explicit types of instruction on these cues are necessary since it may be difficult for learners to acquire them from input exposure. Learners can be encouraged to use the computer-based article tutoring system developed by Zhao and MacWhinney (2018) to compensate for the lack of individualised explicit instruction in language classrooms. A demo version of the tutor is available via the link (http://sla.talkbank.org/English/demo). The tutor has been designed with the usage-based

assumption about construction learning as well as support from instructed second language acquisition theory. Such research is still at the emerging stage. There is great demand for more future research on the usage-based approach to the acquisition and instruction of the English article construction.

Appendix

Essay Topics in the Learner Corpus Sample

No	Topics
1	Does modern technology make life more convenient, or was life better when technology was simpler? Write an essay to state your own opinion.
2	Education is expensive, but the consequences of a failure to educate, especially in an increasingly globalized world, are even more expensive. Write an essay to state your own opinion.
3	Some people think that famous people are treated unfairly by the media, and they should be given more privacy, while some others think that this is the price of their fame.
4	Some people say the government shouldn't put money on building theatres and sports stadiums; they should spend more money on medical care and education. Do you agree or disagree? State the reasons for your view.
5	Some people think that university education is to prepare students for employment. Others think that it has other functions. Discuss and say what other functions you think it should have.
6	Which skill of English is more important for Chinese learners? Some people think that we should give priority to reading in English, while others think speaking is more important. Write an essay to state your own opinion.
7	Some people think that children should learn to compete, but others think that children should be taught to cooperate. Express some reasons of both views and give your own opinion.
8	Will modern technology, such as the internet ever replace the book or the written word as the main source of information? Write an essay to state your opinion.
9	Nowadays, more and more college students rent apartments and live outside campus. Is it appropriate? State your opinion about this.

References

Bickerton, D. (1981). *Roots of language*. Karoma Publishers.

Butler, G. Y. (2002). Second language learners' theories on the use of English article: An analysis of the metalinguistic knowledge used by Japanese students in acquiring the English article system. *Studies in Second Language Acquisition, 24*, 451–480. https://doi.org/10.1017/S0272263102003042

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course*. Heinle & Heinle.
Cole, T. (2000). *The article book: Practice toward mastering a, an, and the* (2nd ed.). The University of Michigan Press.
Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*, *41*(4), 4, 721–744. https://doi.org/10.1017/S0272263118000268
Crosthwaite, P. R. (2016). L2 English article use by L1 speakers of article-less languages. *International Journal of Learner Corpus Research, 2*(1), 69–101. https://doi.org/10.1075/ijlcr.2.1.03cro
Diani, M., & Bison, I. (2004). Organizations, coalitions, and movements. *Theory and Society, 33*, 281–309.
Diez-Bedmar, M. B., & Papp, S. (2008). The use of the English article system by Chinese and Spanish learners. In G. Gilquin, M. B. Diez-Bedmar, & S. Papp (Eds.), *Linking up contrastive and learner corpus research* (pp. 147–175). Rodopi.
Ekiert, M. (2004). Acquisition of the English article system by speakers of Polish in ESL and EFL settings. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *4*(1), 43–78.
Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*, 143–188. https://doi.org/10.1017/S0272263102002024
Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27*(2), 164–194. https://doi.org/10.1093/applin/ami038
Ellis, N. C., & Ferreira-Junior, F. (2009). Construction learning as a function of frequency, frequency distribution, and function. *Modern Language Journal, 93*, 370–385. https://doi.org/10.1111/j.1540-4781.2009.00896.x
Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. Hoboken, Wiley Limited.
Ellis, R., Shawn, L., Elder, C., Reinders, H., Erlam, R., & Philp, J. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*. Multilingual Matters. https://doi.org/10.21832/9781847691767
Eskildsen, S. W. (2015). What counts as a developmental sequence? Exemplar-based L2 learning of English questions. *Language Learning, 65*(1), 33–62. https://doi.org/10.1111/lang.12090
Eskildsen, S. W., & Kasper, G. (2019). Interactional usage-based L2 pragmatics: From form-meaning pairing to construction-action relations. In N. Taguchi (Ed.), *The Routledge handbook of second language acquisition and pragmatics* (pp. 76–191). Routledge.
Goldberg, A. E. (1995). *Construction: A construction grammar approach to argument structure*. Chicago University Press.
Goldberg, A. E. (2006). *Construction at work: The nature of generalization in language*. Oxford University Press.
Greenbaum, S., & Quirk, R. (1990). *A student's grammar of the English language*. Longman.
Huang, S. (1999). The emergence of a grammatical category *definite article* in spoken Chinese. *Journal of Pragmatics, 31*, 77–94. https://doi.org/10.1016/S0378-2166(98)00052-6
Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge University Press.
Huebner, T. (1983). *A longitudinal analysis of the acquisition of English*. Karoma Publishers.
Ionin, T., Zubizarreta, M. L., & Maldonado, S. B. (2008). Sources of linguistic knowledge in the second language acquisition of English articles. *Lingua, 118*, 554–576. https://doi.org/10.1016/j.lingua.2006.11.012
Li, C. N., & Thompson, S. A. (1981). *A functional reference grammar of Mandarin Chinese*. University of California Press.

Luk, Z. P., & Shirai, Y. (2009). In the acquisition order of grammatical morphemes impervious to L1 knowledge? Evidence from the acquisition of plural -*s*, articles, and possessives '*s*. *Language Learning, 59*(4), 721–754. https://doi.org/10.1111/j.1467-9922.2009.00524.x

MacWhinney, B. (2012). The logic of the Unified Model. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 211–227). Routledge/Taylor & Francis.

Master, P. (2013). A contrastive study of determiner usage in EST research articles. *International Journal of Language Studies, 7*(1), 33–58.

McDonald, J. L., & MacWhinney, B. (1989). Maximum likelihood models for sentence processing research. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing* (pp. 397–421). Cambridge University Press.

Parrish, B. (1987). A new look at methodologies in the study of article acquisition for learners of ESL. *Language Learning, 37*, 361–383. https://doi.org/10.1111/j/1467-1770.1987.tb00576.x

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Robertson, D. (2000). Variability in the use of the English article system by Chinese learners of English. *Second Language Research, 16*(2), 135–172. https://doi.org/10.1191/026765800672262975

Sung, M-C., & Kim, H. (2020). Effects of verb-construction association on second language constructional generalizations in production and comprehension. *Second Language Research*, 1–25. https://doi.org/10.1177/0267658320932625

Tachihara, K., & Goldberg, A. E. (2020). Reduced competition effects and noisier representations in a second language. *Language Learning, 70*(1), 219–265. https://doi.org/10.1111/lang.12375

Tarone, E., & Parrish, B. (1988). Task-related variation in interlanguage: The case of articles. *Language Learning, 38*(1), 21–44. https://doi.org/10.1111/j.1467-1770.1988.tb00400.x

Thomas, M. (1989). The acquisition of English articles by first- and second-language learners. *Applied Psycholinguistics, 10*, 335–355. https://doi.org/10.1017/S0142716400008663

Wen, Q. F., Liang, M. C., & Yan, X. Q. (Eds.). (2008). *Spoken and written English corpus of Chinese learners (Version 2.0) [Handbook + CD-ROM]*. Foreign Language Teaching and Research Press.

Zhao, H. (2020). The emergence of second language categorisation of the English article construction. *Languages*, 5*(4)*, 54. https://doi.org/10.3390/languages5040054

Zhao, H., & Fan, J. (2021). Modeling input factors in second language acquisition of the English article construction. *Frontiers in Psychology*, *12*, 653258. https://doi.org/10.3389/fpsyg.2021.653258

Zhao, H., & MacWhinney, B. (2018). The instructed learning of form-function mappings in the English article system. *Modern Language Journal, 102*(1), 99–119. https://doi.org/10.1111/modl.12449

Zhou, Y., & Yang, Y. (2018). Mapping contentious discourse in China: Activists' discursive strategies and their coordination with media. *Asian Journal of Communication, 28*(4), 416–433.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin.

Helen Zhao is a Lecturer in Applied Linguistics in the School of Languages and Linguistics, The University of Melbourne. She completed her Ph.D. in second language acquisition at Carnegie Mellon University. Her primary research focus is second language pedagogy on English grammatical structures. She has worked with several major grammatical systems including articles, prepositions, tense-aspect, and modality. Employing a usage-based approach to language, she has conducted empirical research that explores the intricacy of acquiring these complex grammatical structures. She aims at making innovative use of educational technology to enhance students' second language learning experiences.

Yasuhiro Shirai (Ph.D., Applied Linguistics, UCLA) is a Professor in the Department of Cognitive Science at Case Western Reserve University, USA. His research interests include first and second language acquisition of grammatical constructions, in particular of tense-aspect and relative clauses, and cognitive models of language acquisition and processing. His publications appeared in *Applied Psycholinguistics, Frontiers in Psychology, Journal of Child Language, Journal of Pragmatics, Language, Language Learning, Linguistics, Memory & Cognition, Studies in Second Language Acquisition, Studies in Language,* among others. He has also (co-)authored and (co-)edited more than ten books/special issues of journals, including *The Acquisition of Lexical and Grammatical Aspect* (Mouton de Gruyter), *Handbook of East Asian Linguistics: Japanese* (Cambridge University Press) and *Connectionism and Second Language Acquisition* (Routledge). He is an Associate Editor of *First Language* (Sage). Prior to his current appointment, he was an Assistant/Associate Professor of linguistics at Cornell University and Professor of linguistics at the University of Pittsburgh.