Is *cats* one word or two? L2 learners' processing of number marking in English from the viewpoints of form–meaning mapping

Abstract

This study examined number marking comprehension among Japanese learners of L2 English, whose L1 does not have an obligatory number marking system. The study conducted an online sentence comprehension experiment with 96 L1-Japanese learners and 32 native speakers of English, wherein participants engaged in a self-paced reading with Stroop-like number judgment tasks. Participants were required to determine the number of single words in stimuli (e.g., *cat*/*cats*, one word; *the cats*/*the cat*, two-word sets), and their judgment time was measured for singular and plural words. The results indicated that both groups took more time to judge single plural nouns, suggesting that Japanese L2 learners of English automatically activate plurality in online sentence comprehension as native speakers do. In contrast, neither group showed an interference effect of singularity in judging singular two-word noun sets (*the cat*), unless the singularity is explicitly marked by indefinite article (*a cat*). The lack of interference may be because of unmarkedness of singularity.

Keywords: plural morpheme, number, markedness, English

Introduction

The acquisition of English number marking has attracted researchers' attention for decades, particularly as it is closely related to the acquisition of inflectional morphemes in English, such as plural *-s* and third person singular *-s* (3PS), which are difficult for second language learners to acquire (Ionin and Wexler, 2002; Jiang, 2004, 2007; Jiang et al., 2011, 2017; Lardiere, 1998; Shibuya and Wakabayashi, 2008; Tretiakova, 2020). An investigation into the representation of plurality should provide insight into the acquisition of plural morphemes. Further, examining L2 learners' representation of singularity should contribute to the body of research on the acquisition of English 3PS, given that identifying the grammatical number of subjects is critical to 3PS acquisition (Shibuya and Wakabayashi, 2008).

Jiang (2004) presents seminal work on the acquisition of plural morphemes in a sentence processing task. Using online sentence processing, Jiang investigated whether L2 learners were sensitive to English plural inflectional morphology. The study found that L2 learners were not as sensitive as native speakers of English to either addition or omission of the plural morpheme. Although Jiang (2004, 2007) at first did not identify the cause of L2 learners' insensitivity, Jiang et al. (2011) later concluded that insensitivity to the plural morpheme results from L1 influence. According to the morphological congruency hypothesis (MCH) proposed by Jiang et al. (2011), L2 learners whose L1 lacks a similar morphological marking system as used in the L2, such as Japanese-speaking English learners, have difficulty acquiring the morpheme in the target language. However, other researchers have challenged these findings from the viewpoint of cognitive overload (Wen et al., 2010) and structural differences (Choi and Ionin, 2021; Song, 2015).

Although many studies have focused on the acquisition of the English plural morpheme, they tend to target number agreement. In processing number agreement, learners require knowledge not only of the plural morpheme but also of how number agreement should be successfully computed. As such, it is difficult to identify the cause of failure of processing number agreement. Learners may fully acquire the plural morpheme while continuing to have difficulty processing number agreement. To overcome this pitfall, researchers should investigate the acquisition of the plural morpheme independently from the acquisition of number agreement. Therefore, this study considers the acquisition of the English plural morpheme *-s* to involve the creation of a form–meaning mapping, and views this as emphasizing the crucial role of successfully associating form with meaning in language acquisition.[1]

The need to differentiate number agreement and identify the number information of nouns can also be seen in the inconsistent results of previous research on the acquisition of the English plural morpheme (Jiang, 2004, 2007; Jiang et al., 2011; Song, 2015; Wen et al., 2010). Following the first seminal works on L2 learners' acquisition of the English plural morpheme (Jiang, 2004, 2007), Song (2015) provided evidence against the insensitivity view and argued that inconsistency was due to structural differences in number agreement, such as determiner-head agreement and quantifier-head agreement. In addition, Choi and Ionin (2021) demonstrated that even Chinese and Korean L2 learners of English, whose L1s do not have an obligatory plural marking system, could detect the omission of plural marking in self-paced reading tasks.

Further, existing literature on plural morpheme acquisition tends to use the anomaly detection paradigm, investigating responses to the ungrammaticality of target structures (Jiang et al., 2011; Song, 2015; Wen et al., 2010). As native speakers generally notice ungrammaticality, it is argued that evidence of this tendency among L2 learners indicates that

---

[1] In the semantics literature, there is an argument that plurality does not necessarily mean "more than one" in certain contexts (Sauerland et al., 2005). I briefly touch upon this topic because it has relevance for the discussion of markedness of number. However, because it is a complex issue, examining this topic in more detail is beyond the scope of this paper.

L2 learners possess similar knowledge to native speakers. However, failure to notice anomalies in sentences does not necessarily indicate a failure to acquire a certain feature (Trenkic et al., 2014; Vainio et al., 2016).

Research should also examine the successful use of grammatical information, which the anomaly detection paradigm cannot reveal. Successful use of grammatical knowledge in online sentence processing among L2 learners indicates that they have acquired it. To the best of the author's knowledge, however, existing L2 research has not attempted such an investigation. This study's novel experiment strove to overcome inherent issues in the anomaly detection paradigm when using it to identify the acquisition of inflectional morphology. First, identifying the grammatical number of nouns and computing successful number agreement are different, and investigating the number agreement process of L2 learners does not reveal the source of difficulty.

To overcome this problem, this study utilized a Stroop-like number judgment task—originally developed for L1 psycholinguistic research, that investigated the processing of inflectional number morphology (Berent et al., 2005). In this task, participants were required to determine whether one or two words were presented. The number marking was manipulated to include morphologically plural nouns presented as one word and morphologically singular nouns presented as two words. The mismatch between the number of words and their grammatical number was expected to induce a reaction time (RT) delay. Berent et al. (2005) found that participants responded more slowly when judging a one-word plural noun than a one-word singular noun, concluding that the participants automatically process plurality, which consequently interferes with their number judgment.

In the present study, the Stroop-like number judgment was embedded in the self-paced reading task, similar to Patson and Warren (2010). The participants read a sentence one word at a time. When prompted, they were required to judge the number of words (one or two). The advantage of implementing number judgment in sentence comprehension is that it enables the participants to focus on meaning comprehension. It is possible that in presenting the words without contexts, the participants could strategically pay attention to visual (rather than linguistic) stimuli; namely, if there was a space between two words, there were two words, and if not, there was one word. By contrast, in the present study, the participants were required to engage in sentence comprehension, and they did not know in which part of the sentence they would be asked to judge the number of words. Therefore, they needed to process the words first and then judge the number of words. If a noun's number information was automatically processed, the mismatch between the number information (singular or plural) and the number of words (one or two) would induce RT delay.

As this task did not involve any number agreement computation, it targeted the processing of inflectional morphology and number marking based on surface morphology. Additionally, this task was not based on the anomaly detection paradigm, and thereby investigated the successful use of number information, rather than sensitivity to ungrammaticality. Therefore, using this Stroop-like number judgment task enabled us to remove the influence of various processes involved in number agreement and investigate the representation of number itself.

This paper reviews the theoretical underpinnings of the acquisition of inflectional morphology and grammatical number and introduces the distinction between grammatical and conceptual numbers and number markedness.

Background

Clahsen and Felser (2006, 2018) proposed the shallow-structure hypothesis (SSH), which has been widely applied in L2 research. SSH emphasizes that grammatical information is accessible to L2 learners but they underuse it and tend to rely more on semantic or pragmatic information (Clahsen and Felser 2018). SSH has been applied to not only syntactic but also morphological processing (Clahsen et al., 2010; Silva and Clahsen, 2008). Cunnings (2017) argued that L2 learners rely on whole-word processing with less robust encoding systems to process morphologically complex words, resulting in difficulty in acquiring features requiring morphosyntactic processing, such as gender, person, and number marking.

Morphological Processing

As number is represented by inflectional morphemes, the acquisition of number marking is relevant to the long-lasting debate over whether and how L2 learners decompose morphologically complex words into their stem and inflections (see Gor, 2010). One of the key variables in investigating morphological processing is frequency. Previous studies on L1 morphological processing have found that highly frequent morphologically complex words, such as plural-dominant plurals (e.g., *eggs*), did not show a processing disadvantage, as they were likely to be processed as a whole (Baayen et al., 1997; Beyersmann et al., 2015; Biedermann et al., 2013; New et al., 2004). Similarly, in the L2 literature, Portin et al. (2008) found that L1-Hungarian learners of Swedish used morphological decomposition for low- and medium-frequency words but not for high-frequency words. In contrast, L1 Chinese learners showed full-form processing for all frequencies. Therefore, Portin et al. (2008) concluded that L2 learners can acquire morphological decomposition processing, although L1 influence was present. In addition, Portin et al. (2007) found evidence of morphological decomposition in Finnish-speaking learners of Swedish, including a less proficient group, supporting the availability of morphological decomposition in L2. Tamura et al. (2016) investigated L2 morphological processing and compared the frequency dominance of English nouns using a lexical decision task, including singular-dominant nouns (singular was more frequent than plural), plural-dominant nouns, and control nouns (frequency was largely the same). Unlike L1 studies, Tamura et al. (2016) did not find any evidence of whole-word processing regardless of frequency dominance, supporting the availability of morphological decomposition in English plural nouns. These contradictory results suggest that the morphological processing available to L2 learners may be distinct from that of L1 speakers.

However, some studies argue that L2 learners have difficulty decomposing morphologically complex words, processing them as a whole (Clahsen et al., 2010; Silva and Clahsen, 2008). Silva and Clahsen (2008) investigated the priming effect of English regular past tense *-ed* using a visual lexical decision task and found a facilitative priming effect in L1 but not in L2, suggesting that L2 learners do not decompose English regular past tense inflection. This is in line with SSH, as described above.

More recently, Gor et al. (2017) found that while L2 learners can decompose morphologically complex words under cognitive pressure, L2 learners may not always be engaged in the recombination and checking processes that follow morphological decomposition and are necessary to access the morphosyntactic information of inflections. Ignoring the recombination and checking of inflected words may lead to insensitivity to the inflectional morphemes reported in various sentence processing studies (Clahsen and Felser, 2006, Jiang, 2004, 2007; Jiang et al., 2011, 2017). However, Gor et al. (2017) caution that the source of L2 learners' difficulty found in sentence processing research could also lie in syntactic processing at the sentence level, such as parsing agreement dependencies. If this is correct and L2 learners do not access morphosyntactic information carried by inflection (e.g., plurality in *-s*), the learners would not be likely to exhibit the interference effect in a number judgment task.

Grammatical and Conceptual Number

Previous theoretical and psycholinguistic research demonstrated that number marking should be classified as morphological, grammatical, or conceptual (Corbett, 2000; Nickels et al., 2015; Patson et al., 2014).

The morphological number corresponds to the morphological forms of nouns at the surface level and is usually equivalent to the grammatical number, making morphologically singular forms grammatically singular and morphologically plural forms grammatically plural. Nonetheless, the distinction between these two types of number should be clear, as demonstrated by the so-called zero plurals (e.g., *sheep*). The noun *sheep* can create both singular and plural agreement, such as *this sheep* or *these sheep*, without adding any plural morpheme (Corbett, 2000: 66). This observation suggests the existence of a distinction between morphological and grammatical number.

The distinction between grammatical and conceptual numbers should also be noted, as *plurale tantum*, such as *glasses* or *scissors*, refer to single objects despite being morphologically plural and creating plural agreement, making the grammatical number plural. In addition, collective nouns, such as *family* and *army*, may denote conceptual plurality when addressing the individual members of a group, such as father, mother, or brother (in the case of *family*). Existing research recognizes the critical role played by the three levels of number information, as discussed below.

According to previous research, both grammatical and conceptual numbers play a crucial role in sentence processing (Bock et al., 2004; Eberhard, 1999; Eberhard et al., 2005). L1 psycholinguistic scholars argue that the conceptual plurality of local collective singular nouns does not impact agreement attraction (Bock and Eberhard, 1993). However, when the subject head noun is a collective singular noun, a plural verb is more likely to follow compared with when it is a non-collective singular noun (Bock et al., 1999; Haskell and MacDonald, 2003; Humphreys and Bock, 2005; Staub, 2009), suggesting the possible inference of conceptual plurality from singular collective nouns and its influence on the number agreement process. There is little research on L2s examining conceptual number derived from collective nouns. Kusanagi et al., (2015) demonstrated that in a self-paced reading task, L2 learners of English experienced delays when they encountered an agreement mismatch between subject and copula (e.g., *Everyone in the hall was/*were…*), indicating sensitivity to agreement errors. More importantly, this effect decreased when the noun phrase included collective singular nouns (e.g., *Everyone in the team was/*were…*). These results indicated that L2 learners of English could access the conceptual plurality of collective nouns in online sentence processing. Since L1 speakers do not demonstrate this kind of number attraction effect from the conceptual plurality of local collective singular nouns, the result of Kusanagi et al. (2015) may imply that L2 learners are more sensitive to distributive interpretation of collective nouns. Therefore, it might be inferred that there is a notable difference in how L1 and L2 speakers perceive collective nouns. The present study also investigated L2 learners' comprehension of plural and collective singular nouns to examine the processing of grammatical and conceptual plurality.

Markedness of Number

In L2 acquisition research, markedness is often used to ascertain typological differences between languages. Features that are peculiar to a few languages are more marked. Such typological differences are useful for predicting in which areas or to what degree learners have acquisitional difficulty (Eckman, 1977). More recently, L2 research has focused on markedness in relation to the acquisition of morphemes, as a morphologically complex marked form requires processing of inflectional morphemes (Bañón et al., 2020; McCarthy, 2008).

This paper defines markedness of number in relation to the comparison of singular and plural forms. Grammatically speaking, plural forms in English are derived from singular forms by adding the plural morpheme, except in the case of irregular plural forms (e.g., *mouse–mice*) and zero plurals (e.g., *sheep*). As such, the singular form is the default and, therefore, unmarked.[2] In L1 psycholinguistic research on agreement attraction, the idea of singular forms being unmarked has been supported by findings suggesting that unmarked forms are less likely to cause agreement attraction (e.g., *the key to the cabinet/s was/were*); in contrast, marked number (plurality) induces more agreement errors (Bock and Miller, 1991; Eberhard, 1997).

In understanding markedness of number, it is also important to differentiate grammatical number and conceptual number. According to research on agreement attraction, plurality is marked, but other research has argued that singularity is marked in terms of conceptual number. L1 psycholinguistic research on conceptual number has revealed that the representation of conceptual plurality is not specified unless the noun is preceded by a numerical quantifier (e.g., *two*) (Patson et al., 2014). Thus, unspecified plural phrases (e.g., *the cats*) would rarely indicate the quantity of the denoted nouns even for native speakers (e.g., "two," "three"). In contrast, the conceptual number for singular nouns is specific, because singular nouns always denote "one" (Patson, 2014; Patson et al., 2014).

In a similar vein, Sauerland et al. (2005) and Sauerland and Elbourne (2002) contended, from the viewpoint of semantics, that grammatical plurality does not always indicate "more than one" but may denote "one or more," depending on the context. The example below, taken from Sauerland et al. (2005: 413), clearly demonstrates how plural forms allow both single and multiple interpretations:

(1a) You are welcome to bring your children.

(1b) You are welcome to bring your child.

If you are the host of a party and invite your neighbors with their children, saying (1b) means that only one child is allowed to come to the party, even if your neighbors have two or more children. In contrast, (1a) indicates that the neighbors may bring any number of

[2] For various senses of markedness, see Haspelmath (2006).

children. The contrast between the interpretations of (1a) and (1b) led Sauerland et al. (2005) to conclude that while singular forms specify the number as exactly one, plural forms may leave it unspecified, so they are semantically unmarked.

Bale et al. (2011) supported the view that markedness of number is reversed for grammatical and conceptual numbers. In essence, singular forms are conceptually marked but syntactically unmarked, whereas plural forms are conceptually unmarked but syntactically marked. As this study was interested in the acquisition of grammatical number, morphologically marked plural forms were likely to interfere with number judgment, while morphologically unmarked singular forms were not, as demonstrated in previous psycholinguistic research (Berent et al., 2005; Gulgowski and Błaszczak, 2018). However, if singular forms are morphologically marked, an error may be induced in number inference (Gulgowski and Błaszczak, 2018). Therefore, this study included marked singular conditions (e.g., *a cat*, *one cat*) in addition to an unmarked one (e.g., *the cat*). It should be noted that English does not have morphologically inflected singular nouns, and therefore the markedness was realized by adding *a* and *one*.

The Present Study

Currently, little is known about L2 learners' ability to establish form–meaning connections for plural morphemes, a process necessary to successfully compute number agreement. Therefore, this study examined the nature of form–meaning mapping of L1 Japanese learners of English, since Japanese does not necessarily mark number.[3]

In the present study, form–meaning mapping was operationalized as automatic activation of number information in a meaning-focused comprehension task. This study

[3] Japanese has a marker *-tachi* to express plurality. However, *-tachi* can only be attached to animate nouns, and even without it, nouns can stand for more than one entity (see Nakanishi and Tomioka, 2004 for more on the use of *-tachi*). Corbett (2000) considered languages, such as Japanese and Chinese, to have *general number*. In these languages, the absence of a plural marker does not necessarily indicate singularity: bare nouns mean one or more than one entity. In English, however, singularity or plurality must be expressed.

employed a Stroop-like number judgment task to investigate the accessibility of number morphology as mentioned in the introduction.

Various types of number information should be considered, including grammatical and conceptual plurality. The activation of grammatical plurality was investigated using common plural nouns (e.g., *cats*, *boys*), while conceptual plurality was investigated using collective singular nouns (e.g., *family*, *team*). First, it was expected that due to the interference of plurality, RT would be slower for judging common plural nouns as one-word than for singular forms. Second, although collective singular nouns are morphologically singular, participants would be likely to access conceptual plural information that is stored as part of the lexical information (Kusanagi et al., 2015). In this case, RT for judging collective singular nouns would be slower than for baseline common singular nouns. If L2 learners demonstrate this tendency, it would indicate that conceptual plurality of collective singular nouns was accessible to L2 learners. In addition to these two predictions, the possible frequency effect of plural nouns was also investigated. Given that plural-dominant nouns are more likely to appear in plural forms, their association with plurality would be stronger. Moreover, L2 learners use morphological decomposition to process highly frequent plural nouns, just as they do with common plurals (Tamura et al., 2016). Consequently, RTs in number judgments for plural-dominant plurals could also be delayed for L2 learners.

Further, this study investigated the contrast between singular and plural markedness. As discussed above, plural forms are marked, whereas singular forms are unmarked, which should result in a less inhibitory effect of singular forms in the number judgment task (Berent et al., 2005; Gulgowski and Błaszczak, 2018). Therefore, it was expected that judging singular nouns in two-word sets (e.g., *the cat*) would not be slower than judging plural nouns in two-word sets (e.g., *the cats*).

However, an interesting point to investigate was whether marking singularity more explicitly by using an indefinite article or determiner *one* would delay RTs. Gulgowski and Błaszczak (2018) revealed that morphologically overtly marked singular nouns in Polish caused participants to take more time to judge those singular nouns as two words. Although English does not have any overt singular morphemes, presenting singular nouns with an indefinite article and the determiner *one* may delay judgment of singular nouns as two words. This was examined in this study. Figure 1 visually summarizes the predictions of the present study.

[Insert Figure 1]

Participants

Thirty-two native speakers of English and 96 L2 speakers of English whose first language was Japanese were recruited to participate in this study. The native speakers were exchange students (n = 27) and language teachers in Japan (n = 5) from the US (n = 24), the UK (n = 3), Australia (n = 3), Canada (n = 1), and Singapore (n = 1). Their mean age was 24.16 years (SD = 7.44). The L2 Japanese speakers of English were either undergraduate or graduate students at two universities in Japan. Before the experiment, all participants signed a consent form and agreed to participate for a compensation of 2000 JPY.

The mean age of the L2 learner participants was 20.15 years (SD = 2.22), and their majors varied, including education, engineering, physiology, agriculture, science, economics, law, arts and humanities, international development, and information technology. Of the 96 participants, 19 had stayed in one or more English-speaking countries (including Asian or African countries where English is one of the several official languages) for longer than a month. Their English language proficiency was estimated using the Oxford Quick Placement Test (OQPT), which they took at the beginning of the session; mean score was 38.01 out of 60 points (SD = 5.22). The reliability coefficient for the OQPT was alpha = .64 [.54, .74].

Based on the OQPT results, their estimated proficiency levels in the Common European Framework of Reference were A2 (n = 4), B1 (n = 47), B2 (n = 40), and C1 (n = 5). Table 1 shows the participants' demographics.

Table 1. *Participants' demographics*

JEFL		*n*	*M*	*SD*	*Mdn*	*Minimum*	*Maximum*
	Age (years)	95	20.15	2.22	19	18	30
	Study abroad (months)	19	11.85	12.41	9	1	48
	Learning (years)	96	9.2	2.99	8	6	20
	Starting age (years)	95	10.97	2.42	12	1	18
	Self-reported proficiency						
	Reading	96	3.28	0.97	3	1	5
	Writing	96	2.72	0.98	3	1	5
	Listening	96	3.05	1.16	3	1	5
	Speaking	96	2.23	0.97	2	1	4
	Vocabulary	96	2.58	0.93	3	1	5
	Grammar	96	2.79	0.97	3	1	5
NS							
	Age	31	24.16	7.44	21	19	52
	Staying in Japan (months)	28	26.38	61.87	8	0.75	312
	Staying in non-English-speaking countries other than Japan (months)	9	10.67	17.75	4	1	60

Note. NS: Native speakers; JEFL: Japanese EFL learners. The starting age for learning English was calculated by subtracting the years of learning English from the participant's age. As one participant in the JEFL group did not report his age, his starting age could not be calculated. Self-reported proficiency was rated on a five-point Likert scale (1: *poor* to 5: *good*).

Materials

For both one-word and two-word presentations, the target region was set to the end of the sentence. For one-word presentation, all the target nouns were preceded by *the* and an adjective (e.g., *the friendly dogs*) in order to control the variability caused by processing nouns in different environments. Combinations of adjectives and various nouns (common singular, common plural, plural-dominant, collective) were confirmed to be naturally occurring strings based on the online Corpus of Contemporary American English (Davies, 2008).

The vocabulary level was controlled so that 95% of the target nouns were between levels 1 and 3 of JACET 8000 (JACET Committee of Basic Words Revision, 2003). Although ideally other vocabulary profiles, such as frequency and word length (number of letters), should be controlled, this was impossible in this study. Therefore, frequency and word length were included as covariates in RT analysis. Subtitle frequency and British National Corpus (BNC) data were used as indications of word frequency effects (SUBTLEX-US; Brysbaert and New, 2009).[4] The descriptive statistics of frequency information of the test items are available in the Supplementary Materials.

The one-word condition had four levels: common singular nouns, common plural nouns, collective singular nouns, and plural-dominant plural nouns (2).

(2a) The young man had to talk to the older <u>friend</u>. (common singular)

(2b) The young man had to talk to the older <u>friends</u>. (common plural)

(2c) The young man had to talk to the older <u>couple</u>. (collective)

(2d) The young man had to talk to the older <u>residents</u>. (plural-dominant)

---

[4] Subtitle frequency is frequency information from a corpus of subtitles of US films and television series (51 million words in total). Frequency measures based on the SUBTLEX-US corpus are available online at http://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus.

The baseline level was (2a), as the number of words was in accord with grammatical singularity. However, (2b) and (2d) represented number mismatch, as participants had to state that the number of words was one, ignoring the grammatical plurality of *friends* and *residents*. The difference between (2b) and (2d) was in frequency dominance of the plural nouns. *Residents* is twice or more as frequent as *resident*.[5] Likewise, (2c) was a number mismatch condition. However, in contrast to (2b) and (2d), (2c) marked plurality not morphologically but conceptually. The total number of target sentences for the one-word condition was 28, each of which had four levels, as shown in (2a-d).

The two-word condition also had the following four levels with one number-matched baseline level (3a) and three number-mismatched levels (3b-d). Among the three mismatched levels (3b-d), the target noun phrases in (3c) and (3d) manifested singularity more explicitly than did (3b) by the inclusion of the indefinite article and the determiner *one*. As such, (3d) was more likely to cause conflict in judging the number of words as one. Table S2 summarizes the mutual information (MI) scores of the target items, which were calculated based on the BNC. The MI scores were included as a covariate in the analysis of RT data. The number of target items in the two-word condition was 32. Thus, the combined total number of target items was 60, and they were divided into 4 counterbalanced lists to avoid presenting the same sentences under the same conditions. That is, each participant saw each of the (2a-d) types of sentences seven times during the experiment, and each of the (3a-d) types of sentences eight times.

(3a) The girl continued to chase <u>the rabbits</u>. (definite article + common plural nouns)

(3b) The girl continued to chase <u>the rabbit</u>. (definite article + common singular nouns)

(3c) The girl continued to chase <u>a rabbit</u>. (indefinite article + common singular nouns)

---

[5] The decision of whether plural nouns are plural-dominant was based on the British National Corpus. All singular and plural nouns were extracted based on Part-of-Speech tags (NN1 is singular nouns and NN2 is plural nouns). Then, the frequency per million of the singular and plural forms are listed. The frequency of plural-dominant plural forms is two to three times higher than its singular forms (see Supporting Information).

(3d) The girl continued to chase one rabbit. (*one* + common singular nouns)

In addition to 60 test items, 90 filler items were included in the lists. In the one-word condition, two types of filler items were prepared: personal pronouns (*him*, *her*, *them*, *us*, *me*) and material nouns (*gold*, *stone*, etc.). There were 16 personal pronoun fillers, each of which has singular and plural versions, amounting to 32 personal pronoun fillers in total. In addition, there were 14 material noun fillers, which only appeared in their singular form. The personal pronouns were divided into two lists presenting 16 sentences, eight singular personal pronouns and eight plural personal pronouns. Thus, the total number of fillers presented in the one-word condition is 30 (16+14).

The two-word condition also contained two types of fillers: 16 adjectives + nouns (e.g., *beautiful lakes*) and 28 determiners + nouns (e.g., *this ball* / *these balls*). The latter has a singular form (e.g., *this ball*) and a plural form (e.g., *these ball*s); thus, 28 items were divided into two lists, 14 sentences each, for balance. In sum, each participant read 120 items (60 target and 60 filler items),[6] and the number of number-matched and number-mismatched items was equal.

These filler items were meticulously constructed to prevent participants from discerning the true purpose of the experiments. For instance, the critical region for the filler items was deliberately placed in the middle of the sentences (fifth word on average), contrasting with the target conditions where it was placed at the end. This strategy was designed to ensure that participants remained uncertain about when they would be required to make a number judgement during sentence comprehension. Introducing this element of unpredictability was intended to ensure sustained engagement with the semantic content of

[6] Among 90 fillers, personal pronoun fillers in the one-word condition and determiner + noun fillers in the two-word conditions had singular and plural forms, and therefore, half of each set, 16 out of 32 personal pronoun fillers and 14 out of 28 determiners + noun fillers, was presented to each participant. Thus, the total number of filler items that each participant read was 60 (16+14+16+14).

the sentence from beginning to end. In the experiment, all 120 items were presented randomly.

Experimental Task

The experimental task was developed using the programming language Hot Soup Processor, version 3.4 (http://hsp.tv/). Data were collected either individually or at most from two people simultaneously, in separate sections of the same room. When two participants were in the same room, earplugs were provided to help them focus on the experimental task. The experiment was conducted using a laptop computer, and participants were not allowed to use any Internet resources, dictionaries, or their phones during the experiment. The experiment reported here is part of a larger study, and the entire session took approximately 120 minutes for the learner group and 60 minutes for the native group; the time difference was inevitable because the learner group had to take the placement test while native speakers did not.

The task was similar to a word-by-word self-paced reading task. Participants were required to read a sentence word-by-word by pressing a space bar. However, unlike a self-paced reading task, 200 ms after the participant pressed the space bar to call up the target word(s), the screen color changed to blue, signaling that the participant had to judge how many words were presented on the screen (Figure 2).[7] The color screen changed and the response were required in only the target stimulus in each trial, which could consist of one or two words. Except this target region, the stimulus display was word-by-word. As mentioned earlier, the target regions where the number judgment was required differs for the target items and the filler items. In the target items, the number judgment was always required in the last

[7] There are two reasons for the 200ms gap between the presentation of word/s and the change of the screen color. First, if the screen color changes at the same time as the word appears, the participants would be less likely to process the target word/s. Second, the 200 ms interval eliminates potential RT differences that could arise from the word recognition stage due to variations in the lexical properties or semantic feasibility of the target words. A demonstration video of how the experiment works is available at the OSF (https://doi.org/10.17605/OSF.IO/ZRC5N).

region, while in the filler items, the judgment was required in the middle of the sentence. That is, when the participants read the filler sentences, they need to read the rest of the words in the sentence after the number judgment.

Measurement of participants' RTs started the moment the screen turned blue and stopped the moment they pressed the arrow keys. Participants were instructed to press the left arrow key if they thought the number of words was one and the right arrow key if they thought the answer was two. Although there was no time limit for the number judgment, participants were told to judge as quickly as possible. Following this task, they were asked to answer a simple true-or-false comprehension question, except after the filler items.[8]

Participants were allowed to complete the experiment at their own pace, and they could take a break between each trial if they wanted to. The experimental task took 20 to 25 minutes. Thereafter, all participants answered a background questionnaire survey.[9]

[Insert Figure 2]

Before the main experiment, participants were engaged in a practice session with 10 examples and were instructed to use their left hand to press the space bar and their right hand to press the arrow keys. The use of both hands allowed quicker responses. In the practice session, visual feedback on the responses was automatically provided immediately after they

[8] The rationale for having participants answer comprehension questions was two-fold. First, it was necessary to distract their attention from the purpose of the experiment; they would likely consider meaning comprehension as important as number judgment. Second, from a theoretical standpoint, in order to study the acquisition of morphemes, it is necessary to investigate L2 processing behavior during comprehension focused on meaning, as noted by Jiang (2004, 2007). The reason why only the target items were followed by a comprehension question was to reduce the amount of time to complete the task. If all the items, including filler items, were followed by a comprehension question, the participants must have been exhausted after completing only one of the experimental tasks. Therefore, only target items were followed by a comprehension question.

[9] The instructions were in Japanese for Japanese learners of English, while they were in English for native speakers of English. This ensured full understanding of the task procedures, as not all the participants were familiar with the experimental paradigm or psycholinguistic experiments using a computer.

judged the number of words and answered the comprehension question. No feedback was provided in the main session.

Analysis

RT data were analyzed using the generalized linear mixed-effect model (GLMM) with R 3.3.0 software (R Core Team, 2016) and the *lme4* package (Bates et al., 2017). First, cases in which participants answered comprehension questions incorrectly were removed; these were 8.4% and 12.9% of data for L1 and L2 participants, respectively. Second, incorrect number judgment responses were removed, constituting 2.4% and 2.9% of responses with correctly answered comprehension questions for L1 and L2 groups, respectively.

Using the correct responses, the mean and standard deviations were calculated for each participant, and responses three standard deviations above the mean were removed as outliers. This was done separately for one-word and two-word conditions, as response latency and its variance differed across the two conditions. Moreover, as outlier exclusion based on the mean and standard deviation could not eliminate extreme RTs, responses beyond the cut-off point were removed. For the L1 group, responses above 2500 ms for both conditions were removed. For the L2 group, the cut-off point was set to 2500 ms for the one-word condition and 3000 ms for the two-word condition. The different cut-off point for the L2 group was determined on the basis of the visual inspections of their RTs for the two-word condition, which showed greater variance and a slightly heavier tail of RT distribution.[10] In summary, 81.5% and 92.0% of all the L1 group responses were included in the analysis for the one-word and two-word conditions, respectively, while 79.2% responses for the one-word

[10] The SDs in the two conditions were 446 ms (one-word) vs. 460 ms (two-word), and the skewness values were 3.05 (one-word) vs. 3.35 (two-word). When comparing the RT histograms in both conditions, there is a break in the tail of the two-word condition around 3000 ms (not 2500 ms); thus, I used different cut-off points for both conditions. Please refer to the histograms relevant to L2 learners found in the document titled "R_script_JLE.docx" available on the Open Science Framework (OSF) at https://doi.org/10.17605/OSF.IO/ZRC5N.

condition and 85.8% responses for the two-word condition were included for the L2 group. One might consider that removing at most around 20% of the data should be avoided. Nevertheless, as mentioned earlier, the interest of the present study was in the implicit nature of the L2 learners' behavior under meaning-based comprehension; therefore, removing items in which the participants answered incorrectly because they did not pay attention to the task or their focus on meaning comprehension was low, was required. Similarly, longer RTs were removed as outliers, given that slower responses can result from some irrelevant and unexpected influences of the experimental manipulation or carefully considered conscious judgments.

Following this, a series of generalized linear mixed models (GLMMs) using an inverse-Gaussian distribution with identity link function were fitted to the raw RT data. The use of the GLMM was preferred over the linear mixed effects model (LME) because it provided a better fit for the raw reaction time data both theoretically and mathematically (Lo & Andrews, 2015). Additionally, the use of GLMM avoided the need for transforming the reaction time data, which can hinder the interpretation of the relationship between reaction time and explanatory variables. Despite this, one reviewer recommended analyzing the data using LME and comparing the results. While there was one notable difference between the results obtained using LME and GLMM, the majority of the results were the same in both analyses. Note that the raw RT was log-transformed before the LME analysis in order to fit the linear model well. Readers interested in the results obtained using LME can find them in the appendix.

The explanatory variable of the model was the experimental condition (one-word: common-sg, common-pl, pl-dominant, and collective; two-word: *the* + pl, *the* + sg, *a* + sg, *one* + sg). In addition, covariates that could influence RT latency were added, such as trial order, OQPT scores as a measure of L2 proficiency (only for the L2 group), word length, number of syllables, and base and surface frequency based on the BNC and SUBTLEX-US corpuses. For the two-word condition, MI scores, one of the frequency measures of multiple words, were considered, as the strength of the word association could affect RT latency.[11] All continuous variables were z-transformed, and the categorical variable, the experimental condition, was contrast-coded, with the baseline conditions (common-sg for the one-word condition; *the* + pl for the two-word condition) as reference levels. Model selection was performed by following the stepwise forward procedure based on the Akaike information criterion (AIC). For each of the models, we included covariates and random slopes if these improved the model fit. For example, adding the presentation order as a covariate into the model is better than the model without it, based on AIC In determining the random effect structures, sometimes the model with random slopes did not converge. In such cases, the correlation parameters were removed, while the random slopes were retained because there is a risk of a Type I error for intercept-only models (Barr et al., 2013). Thus, the covariates and the random slopes included in the final models differ (see Supporting Information for data and detailed description of the data analysis).

None of the models reported below include the interaction between L1 and L2 groups and the experimental conditions, given that the foci of the analyses lie in the RT differences between the baseline, such as a singular noun in one word condition, and other experimental conditions (common plurals, collective nouns, plural-dominant plurals) for each group. Even so, L1 and L2 similarities can be discussed by comparing whether RT differences are found in each. In addition, including the interaction term would have added unnecessary complexity to the models, and would have made interpretation of the results more difficult. Instead, the

[11] Mutual Information is a type of frequency measure that represents the strength of the connections between the two words. If the word combination appears together more frequently than separately, the MI score is high. Since it is possible that not only frequency of the target nouns in the two-word condition but also the frequency of the two-word set (e.g., *the cat* vs. *the cats* vs. *a cat* vs. *one cat*) could impact the processing, the MI score was included in the data analytic procedure. MI score is often used in investigating the L2 processing of collocations (e.g, Wolter and Yamashita, 2017).

model was built for each group and to examine the differences in RTs between L1 and L2 groups in each of the experimental conditions, which allowed us to compare whether RT differences were found in each group. Furthermore, the proficiency measure was only available to the JEFL group, and therefore it was impossible to build a model that included both the group factor and the proficiency measure—though it turned out that the proficiency measure did not contribute to the model fitness significantly, and it was removed from the final mode reported below.

Results

One-word Condition

Table 2 presents the descriptive statistics of error rates and RTs for the L1 and L2 groups. Both groups showed RT delays for the three number mismatch conditions compared with the baseline conditions, although some RT latencies were small. The best model for the L1 group included the main effect of condition and presentation order with no effect of the other variables, such as word length or frequency indices. The GLMM analysis revealed that the L1 group took longer to judge collective singular nouns and common plural nouns as one word than they did common singular nouns (collective: estimate = 72.95, *SE* = 26.30, *t* = 2.77, *p* = .006; common-pl: estimate = 63.04, *SE* = 24.94, *t* = 2.53, *p* = .011).[12] However, the RT difference between the baseline condition and the pl-dominant condition was not significant (estimate = 0.17, *SE* = 25.07, *t* = 0.01, *p* = .995). The detailed model summary of the L1 group is presented in Table S3.

Table 2. *Mean RTs (ms) and mean error rates (SD in parentheses) in the one-word condition.*

[12] The LME analysis did not find significant difference between the baseline and collective singulars. The possible explanation for this is provided in the discussion section.

Group	Condition	Number mismatch	Example	Mean RT (*SD*)	Mean error rate (*SD*)
NS	common-sg	No (baseline)	*Cat*	777 (375)	.014 (.045)
	collective-sg	Yes	*Couple*	825 (389)	.030 (.063)
	common-pl	Yes	*Cats*	826 (401)	.024 (.079)
	pl-dominant	Yes	*Residents*	789 (361)	.025 (.059)
JEFL	common-sg	No (baseline)	*Cat*	857 (252)	.018 (.055)
	collective-sg	Yes	*Couple*	877 (277)	.022 (.058)
	common-pl	Yes	*Cats*	887 (254)	.024 (.083)
	pl-dominant	Yes	*Residents*	939 (297)	.036 (.086)

Note. Common-sg = singular form of the common noun; collective-sg = singular form of the collective noun; common-pl = plural form of the common noun; pl-dominant = plural form of the plural-dominant noun; RTs = reaction times; MS = milliseconds; SD = standard deviation.

The final model for the L2 group included the main effect of the experimental condition, the order of presentation, and the Zipf surface frequency from the SUBTLEX-US corpus (Table S4). The L2 group demonstrated significant RT delays in all three number mismatch conditions (collective: estimate = 34.36, *SE* = 17.09, $t = 1.99$, $p = .047$; common-pl: estimate = 41.74, *SE* = 15.23, $t = 2.74$, $p = .006$; pl-dominant: estimate = 131.96, *SE* = 15.87, $t = 8.31$, $p < .001$). The Zipf surface frequency from the SUBTLEX-US corpus, although entered into the model, did not reach a significant level (estimate = -6.42, SE = 6.72, $t = 0.96$, $p = .340$).

Two-word Condition

The mean error rates and RTs for the two-word condition are summarized in Table 3.

Table 3. *Mean RTs (ms) and mean error rates (SD in parentheses) in the two-word condition*

Group	Condition	Number mismatch	Example	Mean RT (*SD*)	Mean error rate (*SD*)
NS	*the* + pl	No (baseline)	*The rabbits*	686 (309)	.032 (.079)
	a + sg	Yes	*A rabbit*	732 (300)	.013 (.041)
	one + sg	Yes	*One rabbit*	679 (280)	.030 (.068)
	the + sg	Yes	*The rabbit*	714 (336)	.028 (.055)
JEFL	*the* + pl	No (baseline)	*The rabbits*	870 (301)	.012 (.043)
	a + sg	Yes	*A rabbit*	875 (289)	.073 (.111)
	one + sg	Yes	*One rabbit*	903 (321)	.027 (.071)
	the + sg	Yes	*The rabbit*	847 (283)	.026 (.061)

Note. the + pl = definite article with common plural; *the* + sg = definite article with common singular; *a* + sg = indefinite article with common singular; *one* + sg = numeral quantifier with singular. RTs = reaction times; MS = milliseconds; SD = standard deviation.

The tendency of RT latencies for the number mismatch conditions differed between the L1 and L2 groups. The final model of the GLMM analysis for the L1 group included the main effects of condition, presentation order, and MI scores. A significant RT difference was found only between the baseline *the* + pl condition and *a* + sg condition (estimate = 60.10, $SE = 27.03$, $t = 2.22$, $p = .026$). However, none of the other contrasts were significant (*the* + sg: estimate = 33.47, $SE = 27.12$, $t = 1.23$, $p = .217$; *one* + sg: estimate = 1.08, $SE = 25.80$, $t = 0.4$, $p = .967$). Table S5 provides the results from the GLMM analysis.

The results for the L2 group are shown in Table 3. RT delay appeared only for the *one* + sg condition, which was confirmed by the GLMM analysis. There was a significant RT difference only between *the* + pl condition and *one* + sg condition (estimate = 55.06, $SE = 12.48$, $t = 4.41$, $p < .001$), which was not observed in the L1 group. However, none of the other number mismatch conditions showed any significant RT delays compared with the baseline condition (*a* + sg: estimate = -5.30, $SE = 13.23$, $t = 0.40$, $p = .688$; estimate = -

22.99, *SE* = 14.04, *t* = 1.64, *p* = .102). A more detailed model summary is presented in Table S6.

Discussion

This study aimed to assess the activation of number information using an online sentence comprehension task for L1 English speakers and L2 learners. In the one-word task, both L1 and L2 participants took more time to judge common plurals and collective singular nouns than they did baseline common singular nouns. However, while the L2 group exhibited significant delays in judging plural-dominant plurals, the L1 group did not show this tendency.

In the two-word task, the L2 group's judgment of *one* + N was significantly slower than the baseline condition, while the L1 group demonstrated no significant differences. In contrast, the L1 group's number judgment was significantly slower for *a* + N, whereas the L2 group's judgment showed no significant differences. Moreover, neither group showed RT delays for *the* + N condition regardless of noun phrase singularity. Table 4 summarizes the results of both tasks for both groups of participants.

Table 4. *Results summary*

Condition	Estimated RT difference against the baseline	
	NS	JEFL
collective-sg	**72.95 (*p* = .006)**[13]	**34.36 (*p* = .047)**
common-pl	**63.04 (*p* = .011)**	**41.74 (*p* = .006)**
pl-dominant	0.17 (*p* = .995)	**131.96 (*p* < .001)**
a + sg	**60.10 (*p* = .026)**	-5.30 (*p* = .688)
one + sg	1.08 (*p* = .967)	**55.06 (*p* = .001)**
the + sg	33.47 (*p* = .217)	-22.99 (*p* = .102)

Note. The negative value indicates the RT is longer than the baseline.

[13] Note that the LME analysis did not find statistically significant difference here (*p* = .100).

Processing of Plurals

L2 learners of English whose L1 does not have an obligatory number marking system showed a clear number mismatch effect in judging common plural nouns as one word in the same way as L1 English speakers. This indicates activation of grammatical plurality of nouns and suggests that the participants could acquire the English plural morpheme -*s*, at least in terms of form–meaning mapping. In addition, the number information attached to plural morphemes could be accessed quickly in meaning-focused sentence processing. This is in line with previous findings, that L2 learners can acquire the English plural morpheme (Choi and Ionin, 2021; Song, 2015; Wen et al., 2010) while contradicting the MCH (Jiang, 2004, 2007; Jiang et al., 2011, 2017), given that the MCH presupposes L1 influence on the acquisition of inflectional morphemes. L2 learners' activation of plural information as found in the current study corroborates the ideas of Gor et al. (2017) who suggest that morphological decomposition of inflected words may be available to L2 learners. It can therefore be assumed that L2 learners' processing difficulty may be due to problems of syntactic processing at the sentence level, such as parsing agreement dependency, and not due to morphological processing. Further studies are needed to explore this possibility.

In addition, both L1 and L2 groups showed an interference effect in judging collective nouns as one word regardless of the grammatically singular form. This may be due to conceptual plural information being activated and confusing the number judgment. L2 learners' availability of conceptual number derived from collective nouns is in line with Kusanagi et al. (2015).

Nonetheless, the result must be interpreted with caution because the LME analysis, the L1 group did not demonstrate significant delay in judging the collective singulars ($p = .100$). The log transformation of the raw reaction time data, which was initially right-skewed, produces a normally distributed dataset. This means that the influence of data points on the right tail of the distribution may be greater in the GLMM compared to the LME, where their impact may be restricted. This difference in influence may explain why the effects of the collective singular condition, which could probably have a right-skewed distribution of the reaction times, are different in the two analyses.

Regarding the judgment of plural-dominant plural nouns, L2 group demonstrated an interference from plurality, indicating the possibility that L2 learners might not use whole-word processing for plural-dominant plurals.[14] Therefore, the L2 group may have processed plural-dominant plurals similarly to common plural nouns. Consequently, the plurality was more likely to interfere with number judgment. As discussed above, previous research on L1 lexical processing suggests that highly frequent morphologically complex words, such as plural-dominant plurals, are processed as fast as their corresponding singular forms, as they follow the whole-word route (Baayen et al., 1997; Beyersmann et al., 2015; Biedermann et al., 2013; New et al., 2004). Nevertheless, this might not be the case for L2 learners as suggested by AUTHORS (***). However, a note of caution is due here, as the lack of the interference effect for the L1 group does not necessarily indicate inaccessibility of grammatical plural information. Rather, it is speculated that it was insufficient to interfere with the number judgment process in the present experimental task.

Together with the result of the weak interference effect of collective singulars in LME analysis and the lack of interference found both in LME and GLMM for L1 group, the difference between L1 and L2 group could be attributed to the two kinds of interpretations of group nouns: collective reading and distributive reading as Bock et al.(2004) suggested. In the former interpretation, collective nouns refer to groups that denote a set such as *audience*

---

[14] “Highly frequent morphologically complex words” does not mean that the words themselves are more frequent than other words. Rather the key here is the frequency difference between the base forms (e.g., singular forms) and their inflected forms (e.g., plural forms). Inflected forms are usually decomposed into a stem and inflected morphemes. However, if the inflected forms are more frequent than their base forms, they are processed as a whole (Baayen et al., 1997; Beyersmann et al., 2015; Biedermann et al., 2013; New et al., 2004).

as a whole, whereas in the latter case, they refer to constituents of the group such as *audience* consisted of multiple separate individuals. Conceptual plurality of collective singular nouns arises from this distributive reading. Thus, if the collective interpretation is superior to distributive reading during the task for the L1 group, interference from collective singular nouns would be less likely. In contrast, L2 group might have preferred distributive reading, focusing on the individuals comprising the set. In fact, as reviewed in the background section, Japanese EFL learners would be likely to go with distributive reading in processing collective singular nouns (Kusanagi et al., 2015), thereby showing notional attraction in number agreement. However, the precise reasons for this contrast between L1 and L2 speakers remain unclear, pointing to a need for additional studies to explore this phenomenon.

While the following is merely speculation, the L2 group's preference for distributive reading, which focuses on the individual members of a group, may have contributed to the difference in the processing of plural-dominant plurals between the L1 and L2 groups. While some of the plural-dominant plurals used in the study, such as *skills* and *chemicals*, may be difficult to interpret as sets of individuals, others, such as *parents*, *individuals*, *soldiers*, *kids*, *residents*, and so on, may have been interpreted distributively. Thus, given the tendency of L2 learners and the characteristics of the stimuli used in the study, it is possible that the L2 group was influenced by both the conceptual plurality derived from accessing the individuals in the plural sets and the plural meaning conveyed by the plural morphemes. L1 group, on the other hand, might not have used distributive reading in processing plural-dominant plurals. Furthermore, as discussed above, morphological processing of plural-dominant plurals is different from that of common plurals for L1 speakers. Therefore, L1 speakers might not have shown the interference from plural-dominant plurals.

Processing of Singulars

For both groups, grammatical singularity of *the* + sg (e.g., *the cat*) did not interfere with number judgment. These results corroborate the findings of Berent et al. (2005), who demonstrated that number interference was not observed in the two-word condition, as number was not specified in singular nouns. This observation may also support the markedness hypothesis discussed above, which suggests that plural forms of nouns are marked, as marked forms are derived from default unmarked forms, singular nouns. Previous psycholinguistic research has suggested that marked number information is more likely to influence number agreement attraction (Bock and Miller, 1991; Eberhard, 1997). It is possible that number marking of singular nouns did not have a significant impact on number judgment for the L1 or the L2 group. These findings provide support for the conceptual premise that the activation of number information is asymmetrical: unmarked singular forms are less likely to interfere with processing, while marked plural forms are more likely to be activated during processing.

On the other hand, both groups showed some discrepancies when singularity was overtly marked by *a* and *one*. While the NS group was slower in responding to *a* + sg items (e.g., *a cat*), the JEFL group did not display any delay. Rather, the JEFL group's responses were slower in *one* + sg condition (e.g., *one cat*), even though the NS group did not show that tendency. The fact that both groups showed some influences from the singularity markers, though in different ways, is in accordance with the observation reported in Eberhard (1997). Eberhard revealed that number information that is lexically specified enhances the activation of singularity. It is also in line with Gulgowski and Błaszczak (2018), who demonstrated that in Polish, if the singularity is overtly marked by morphemes, unlike Hebrew (the target language of Berent et al., 2005) or English, it can interfere with the judgment of multiple singular words.

Nevertheless, the question remains as to why the two groups were differently affected by the singularity markers. One possible explanation for the lack of interference effect in *a* + sg conditions for the JEFL group could be that they tend to rely more on lexical information (e.g., *one*) than grammatical information (e.g., *a*), an argument similar to that made by SSH (Clahsen and Felser, 2018). Thus, the JEFL group did not successfully access the singularity of the indefinite article. Furthermore, non-interference from *one* observed for the NS group could be due to the association between the determiner and the noun. As can be seen from Table S2, the MI score was lowest for *one* + sg items. Therefore, it might have been easy for the NS participants to disassociate the word combination and judge it as two words. However, the MI score itself does not fully explain the result given that although the MI score was added to the final model, the main effect did not attain significance (Table S5). This is an important issue for future research.

Limitations and Future Research Directions

This study had some limitations. First, the number of items tested in the experiment was limited, resulting in the broad 95% CI. As the number of collective nouns and plural-dominant nouns was limited, it was difficult to develop more test items for the one-word condition. Similarly, the nouns used in this study were limited to concrete countable nouns. To generalize the results to other nouns, abstract nouns that the plural morpheme can be attached to should also be investigated in future research.

Second, the number of filler items could have been larger so that the participants could not ascertain the purpose of the study. Keating and Jegerski (2015) recommend 75% or more filler items in the design of single variables with two levels, although they also argue “it is…highly desirable to have greater than 50% of non-critical items” (p.17), which is the proportion of the filler items in this study. Since the experiment reported in this paper is part of a larger study and sessions lasted two hours, adding more fillers might have resulted in a fatigue effect with participants losing focus. Nevertheless, future studies should add more fillers while maintaining the balance between the number of number-match and mismatch items, the number of one-word and two-word items, and the sentences' target regions where the participants are required to judge the number of words.

Third, the design of the study could be improved. Ideally, the comparisons should be made between minimal pairs in one-word condition. For example, plural-dominant plurals should have been compared with counterpart plural-dominant singulars, to reduce the possible influence of comparing different words. Comparing different words also raises the issue of how the target words fit with the same sentence context. Though the adjective and noun combinations were checked using Corpus of Contemporary American English (Davies, 2008-), it would have been better to control contextual plausibility in each condition. Nevertheless, given that the baseline should always be singular nouns, it is impossible to prepare the baseline condition for singular collectives (e.g., *family*). As an anonymous reviewer suggested, one way to overcome this issue is to collect lexical decision task data from the same participants and use it to control the possible response differences between different words. Future studies that attempt to use the Stroop-like number judgment task should consider this issue. Therefore, caution should be exercised when interpreting the effects found for collective singulars and plural dominant plurals, even though frequency and length were taken into account in the analysis. In contrast, the key interference effect observed for common plurals can be considered more reliable because the only difference between the baseline and common plurals is the presence of the plural morpheme.

Finally, it would be ideal to have a single regression model that integrates the analysis of both the one-word and two-word conditions, rather than analyzing the two datasets separately. This would allow the study to demonstrate any significant interactions between the number marking of nouns (singular or plural) and the experimental conditions (one-word or two-word). If there is a significant interaction between these two factors and a significant difference in reaction time is only observed in the one-word condition, it would suggest that number marking influences the processing of plurality but not singularity. Future studies should use the same sets of nouns and context sentences in both the one-word and two-word conditions to enable the inclusion of an interaction term in the analysis.

In spite of these limitations, the study does add to our understanding of the nature of L2 acquisition of plural morphemes. Gor et al. (2017) argued that the potential source of L2 learners' difficulties in online processing may be their inefficient morphological processing, such that they ignore the recombination and checking of inflected words after morphological decomposition. However, this study contradicted this claim, as there was a clear interference effect in number judgment. RT delays in judging the plural nouns as one word among L2 learners indicated that they engaged in recombination and accessed morphosyntactic information. Therefore, agreement insensitivity found in online sentence processing research may be due to syntactic problems.

Number judgment tasks can be useful for separating the syntactic agreement process and the processing of number information at the word level. Therefore, future research should integrate experiments that tap into morphological processing (e.g., priming experiments), form–meaning mapping (e.g., number judgment task), and number agreement (e.g., sentence processing experiments). Combining the results from such experiments would provide new insights into the sources of difficulty for L2 learners.

Conclusions

This study investigated the comprehension and activation of number marking among Japanese L2 English learners whose L1 did not have an obligatory number marking system. This study found that L2 Japanese learners of English activated plurality information in online sentence comprehension in processing common plural nouns in a way similar to native speakers, whereas their processing of plural-dominant plurals and singular nouns contrasts with native speakers. Therefore, L2 learners may create form–meaning mapping between inflectional plural morphemes and plural meaning, even though their L1 does not have obligatory plural morphemes, which contradicts MCH (Jiang et al., 2011, 2017). This finding has important implications for the development of a model of L2 processing, such as the SSH proposed by Clahsen and Felser (2018), because revealing the form–meaning mapping has the potential to bridge the gap between research on morphological processing of inflected words and that on sentence processing. However, some questions remain regarding differences in the processing of highly frequent plural nouns and singular nouns among L1 speakers and L2 learners. Future studies should focus on these aspects.

Competing Interests

The author declares no competing interests.

Acknowledgements

This article is partly based on the author's doctoral dissertation, submitted to Nagoya University in 2018. I am deeply grateful to Junko Yamashita, Masatoshi Sugiura, and Koji Miwa for their invaluable support during this process. An earlier version of this paper was presented at the Second Language Research Forum in 2019. I would also like to express my heartfelt thanks to Professor Roumyana Slabakova, Editor of Second Language Research, and the anonymous reviewers for their insightful comments and patience, which significantly improved the quality of the article. A special note of gratitude goes to my best friend, Junya Fukuta, for his thoughtful commentary on the manuscript. Any remaining errors are solely my responsibility.

References

Baayen RH, Dijkstra T and Schreuder R (1997) Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* 37(1): 94–117. DOI: 10.1006/jmla.1997.2509

Bale A, Gagnon M and Khanjian H (2011) On the relationship between morphological and semantic markedness. *Morphology* 21(2): 197–221. DOI: 10.1007/s11525-010-9158-1

Bañón JA, Miller D and Rothman J (2020) Examining the contribution of markedness to the L2 processing of Spanish person agreement. *Studies in Second Language Acquisition*: 1–30. Published online by Cambridge University Press on 10 November 2020. DOI: 10.1017/S0272263120000479.

Barr, D. J., Levy, R., Scheepers, C., and Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3): 255–278. DOI: 10.1016/j.jml.2012.11.001

Bates D, Maechler M, Bolker B and Walker S (2017) *lme4: Linear mixed-effects models using Eigen and S4* [R package version 1.1-13]. Available at: https://cran.r-project.org/package=lme4

Berent I, Pinker S, Tzelgov J, Bibi U and Goldfarb L (2005) Computation of semantic number from morphological information. *Journal of Memory and Language* 53(3): 342–358. DOI: 10.1016/j.jml.2005.05.002

Beyersmann E, Dutton EM, Amer S, Schiller NO and Biedermann B (2015) The production of singular- and plural-dominant nouns in Dutch. *Language, Cognition and Neuroscience* 30(7): 867–876. DOI: 10.1080/23273798.2015.1027236

Biedermann B, Beyersmann E, Mason C and Nickels L (2013) Does plural dominance play a role in spoken picture naming? A comparison of unimpaired and impaired speakers. *Journal of Neurolinguistics* 26(6): 712–736. DOI: 10.1016/j.jneuroling.2013.05.001

Bock K and Eberhard KM (1993) Meaning, sound and syntax in English number agreement. *Language and Cognitive Processes* 8(1): 57–99. DOI: 10.1080/01690969308406949

Bock K, Eberhard KM and Cutting JC (2004) Producing number agreement: How pronouns equal verbs. *Journal of Memory and Language* 51(2): 251–278. DOI: 10.1016/j.jml.2004.04.005

Bock K and Miller CA (1991) Broken agreement. *Cognitive Psychology* 23(1): 45–93. DOI: 10.1016/0010-0285(91)90003-7

Bock K, Nicol J and Cutting JC (1999) The ties that bind: Creating number agreement in speech. *Journal of Memory and Language* 40(3): 330–346. DOI: 10.1006/jmla.1998.2616

Brysbaert M and New B (2009) Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4): 977–990. DOI: 10.3758/BRM.41.4.977

Choi SH and Ionin T (2021) Plural marking in the second language: Atomicity, definiteness, and transfer. *Applied Psycholinguistics*, 42(3): 549-578.DOI: 10.1017/S0142716420000569

Clahsen H and Felser C (2006) Grammatical processing in language learners. *Applied Psycholinguistics* 27(1): 3–42. DOI: 10.1017/S0142716406060024

Clahsen, H and Felser C (2018) Some notes on the shallow structure hypothesis. *Studies in Second Language Acquisition* 40(3): 693–706. DOI: 10.1017/S0272263117000250

Clahsen H, Felser C, Neubauer K, Sato M and Silva R (2010) Morphological structure in native and nonnative language processing. *Language Learning* 60(1): 21–43. DOI: 10.1111/j.1467-9922.2009.00550.x

Corbett GG (2000) *Number*. Cambridge: Cambridge University Press.

Cunnings I (2017) Interference in native and non-native sentence processing. *Bilingualism: Language and Cognition* 20(4): 712–721. DOI: 10.1017/S1366728916001243

Davies M (2008) The corpus of contemporary American English (COCA): 520 million words, 1990-present. Available at: https://corpus.byu.edu/coca/

Eberhard KM (1997) The marked effect of number on subject–verb agreement. *Journal of Memory and Language* 36(2): 147–164. DOI: 10.1006/jmla.1996.2484

Eberhard KM (1999) The accessibility of conceptual number to the processes of subject–verb agreement in English. *Journal of Memory and Language* 41(4): 560–578. DOI: 10.1006/jmla.1999.2662

Eberhard KM, Cutting JC and Bock K (2005) Making syntax of sense: Number agreement in sentence production. *Psychological Review* 112(3): 531–559. DOI: 10.1037/0033-295X.112.3.531

Eckman FR (1977) Markedness and the contrastive analysis hypothesis. *Language Learning* 27(2): 315–330. DOI: 10.1111/j.1467-1770.1977.tb00124.x

Gor K (2010) Introduction. Beyond the obvious: Do second language learners process inflectional morphology? *Language Learning* 60(1): 1–20. DOI: 10.1111/j.1467-9922.2009.00549.x

Gor K, Chrabaszcz A and Cook S (2017) Processing of native and nonnative inflected words: Beyond affix stripping. *Journal of Memory and Language* 93: 315–332. DOI: 10.1016/j.jml.2016.06.014

Gulgowski P and Błaszczak J (2018) Stroop-like interference of grammatical and visual number. *Acta Linguistica Academica* 65(2–3): 259–291. DOI: 10.1556/2062.2018.65.2-3.3

Haskell TR and MacDonald MC (2003) Conflicting cues and competition in subject–verb agreement. *Journal of Memory and Language* 48(4): 760–778. DOI: 10.1016/S0749-596X(03)00010-X

Humphreys KR and Bock K (2005) Notional number agreement in English. *Psychonomic Bulletin & Review* 12(4): 689–695. DOI: 10.3758/BF03196759

Ionin T and Wexler K (2002) Why is 'is' easier than '-s'?: Acquisition of tense/agreement morphology by child second language learners of English. *Second Language Research* 18(2): 95–136. DOI: 10.1191/0267658302sr195oa

JACET Basic Words Revision Committee (eds) (2003) *JACET List of 8000 Basic Words (JACET8000).* Tokyo: JACET.

Jiang N (2004) Morphological insensitivity in second language processing. *Applied Psycholinguistics* 25(4): 603–634. DOI: 10.1017/S0142716404001298

Jiang N (2007) Selective integration of linguistic knowledge in adult second language learning. *Language Learning* 57(1): 1–33. DOI: 10.1111/j.1467-9922.2007.00397.x

Jiang N, Novokshanova E, Masuda K and Wang X (2011) Morphological congruency and the acquisition of L2 morphemes. *Language Learning* 61(3): 940–967. DOI: 10.1111/j.1467-9922.2010.00627.x

Jiang N, Hu G, Chrabaszcz A and Ye L (2017) The activation of grammaticalized meaning in L2 processing: Toward an explanation of the morphological congruency effect. *International Journal of Bilingualism* 21(1): 81–98. DOI: 10.1177/1367006915603823

Keating GD and Jegerski J (2015). Experimental designs in sentence processing research: A methodological review and user's guide. *Studies in Second Language Acquisition*, 37(1): 1–32. DOI: 10.1017/S0272263114000187

Kusanagi K, Tamura Y and Fukuta J (2015) The notional number attraction in English as a foreign language: A self-paced reading study. *Journal of the Japan Society for Speech Sciences* 16: 77–96.

Lardiere D (1998) Dissociating syntax from morphology in a divergent L2 end-state grammar. *Second Language Research* 14(4): 359–375. DOI: 10.1191/026765898672500216

Lo S and Andrews S (2015) To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, 6. DOI: 10.3389/fpsyg.2015.01171

McCarthy C (2008) Morphological variability in the comprehension of agreement: An argument for representation over computation. *Second Language Research* 24(4): 459–486. DOI: 10.1177/0267658308095737

Nakanishi K and Tomioka S (2004) Japanese plurals are exceptional. *Journal of East Asian Linguistics* 13(2): 113–140. DOI: 10.1023/B:JEAL.0000019058.46668.c1

New B, Brysbaert M, Segui J, Ferrand L and Rastle K (2004) The processing of singular and plural nouns in French and English. *Journal of Memory and Language* 51(4): 568–585. DOI: 10.1016/j.jml.2004.06.010

Nickels L, Biedermann B, Fieder N and Schiller NO (2015) The lexical-syntactic representation of number. *Language, Cognition and Neuroscience* 30(3): 287–304. DOI: 10.1080/23273798.2013.879191

Patson ND (2014) The processing of plural expressions. *Language and Linguistics Compass* 8(8): 319–329. DOI: 10.1111/lnc3.12085

Patson ND, George G and Warren T (2014). The conceptual representation of number. *Quarterly Journal of Experimental Psychology* 67(7): 1349–1365. DOI: 10.1080/17470218.2013.863372

Patson ND and Warren T (2010) Evidence for distributivity effects in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36(3): 782–789. DOI: 10.1037/a0018783

Portin M, Lehtonen M, Harrer G, Wande E, Niemi J and Laine M (2008) L1 effects on the processing of inflected nouns in L2. *Acta Psychologica* 128(3): 452–465. DOI: 10.1016/j.actpsy.2007.07.003

Portin M, Lehtonen M and Laine M (2007) Processing of inflected nouns in late bilinguals. *Applied Psycholinguistics* 28(1): 135–156. DOI: 10.1017/S014271640607007X

R Core Team (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Available at: http://www.R-Project.org

Sauerland U, Anderssen J and Yatsushiro K (2005) The plural is semantically unmarked. In: Kepser S and Reis M (eds) *Linguistic Evidence.* Berlin: Mouton de Gruyter, pp. 413–434. DOI: 10.1515/9783110197549.413

Sauerland U and Elbourne P (2002) Total reconstruction, PF movement, and derivational order. *Linguistic Inquiry* 33(2): 283–319. DOI: 10.1162/002438902317406722

Shibuya M and Wakabayashi S (2008) Why are L2 learners not always sensitive to subject-verb agreement? *EUROSLA Yearbook* 8: 235–258. DOI: 10.1075/eurosla.8.13shi

Silva R and Clahsen H (2008) Morphologically complex words in L1 and L2 processing: Evidence from masked priming experiments in English. *Bilingualism: Language and Cognition* 11(2): 245–260. DOI: 10.1017/S1366728908003404

Song Y (2015) L2 processing of plural inflection in English: L2 processing of plural inflection in English. *Language Learning* 65(2): 233–267. DOI: 10.1111/lang.12100

Staub A (2009) On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language* 60(2): 308–327. DOI: 10.1016/j.jml.2008.11.002

Suzuki Y (2017) Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics* 38(5): 1229–1261. DOI: 10.1017/S014271641700011X

Suzuki Y and DeKeyser R (2017) The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning* 67(4): 747–790. DOI: 10.1111/lang.12241

Tamura Y, Morita, M and Nishimura Y (2016, September 12) Word frequency dominance and L2 word recognition. [Paper presentation]. Vocab@Tokyo. Tokyo, Japan.

Trenkic D, Mirkovic J and Altmann GTM (2014) Real-time grammar processing by native and non-native speakers: Constructions unique to the second language. *Bilingualism: Language and Cognition* 17(2): 237–257. DOI: 10.1017/S1366728913000321

Tretiakova ES (2020) Japanese EFL Learners' interpretation of plural morphology. *Journal of Psycholinguistics* 4: 122–133. DOI: 10.30982/2077-5911-2020-46-4-122-133

Vafaee P, Suzuki Y and Kachisnke I (2017) Validating grammaticality judgment tests. *Studies in Second Language Acquisition* 39(1): 59–95. DOI: 10.1017/S0272263115000455

Vainio S, Pajunen A and Hyönä J (2016) Processing modifier-head agreement in L1 and L2 Finnish: An eye-tracking study. *Second Language Research* 32(1): 3–24. DOI: 10.1177/0267658315592201

Wen Z, Miyao M, Takeda A, Chu W and Schwartz BD (2010) Proficiency effects and distance effects in nonnative processing of English number agreement. In: Franich K, Iserman KM and Keil LL (eds.), *Proceedings of the 34th Annual Boston University*

Conference on Language Development. Somerville, MA: Cascadilla Press, pp. 445–456.

Wolter, B, and Yamashita, J (2018) Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance? *Studies in Second Language Acquisition*, *40*(2): 395–416. https://doi.org/10.1017/S0272263117000237

Figure captions

Figure 1. Predictions of the present study: The left- and right-hand side figures show the one-word and two-word presentation conditions, respectively.

Figure 2. Schematic of the Stroop-like number judgment task in the target conditions. In the filler item conditions, the screen color changed in the middle of the sentence.