Title: **How classifiers facilitate predictive processing in L1 and L2 Chinese: The role of semantic and grammatical cues**

Authors: Theres Grüter[1]

Elaine Lau[2]

Wenyi Ling[1]

Affiliation: [1)] University of Hawai‘i at Mānoa

[2)] Chinese University of Hong Kong

Address for correspondence:

Theres Grüter

University of Hawai‘i at Mānoa

Department of Second Language Studies

1890 East-West Road, Moore Hall, rm570

Honolulu, HI 96822

U.S.A.

theres@hawaii.edu

Abstract

This study examines long-standing claims that L2 learners rely more on non-grammatical than on grammatical information during sentence processing compared to native speakers. Nominal classifiers in Mandarin Chinese offer an ideal opportunity to test this claim, as they simultaneously encode semantic as well as grammatical form-class cues about co-occurring nouns. This paper reports findings from a visual world eye-tracking experiment with L1 and L2 speakers of Mandarin, which was designed to assess listeners' relative reliance on these two concurrently available cues when creating expectations about an upcoming noun in a sentence. Results show that L2 listeners experienced greater competition than L1 listeners from nouns that were grammatically incompatible with the classifier they heard but shared semantic features associated with it. The greater reliance on semantic cues observed in L2 processing is argued to be an effect of adaptation to the relative reliability of information, serving to maximise L2 processing efficiency.

Keywords: classifier, Mandarin Chinese, eye-tracking, cue weighting, prediction, L2 processing

The incremental integration of cues from various information sources is a hallmark of human language processing. There is ample evidence that listeners use both linguistic and non-linguistic information incrementally and predictively to derive meaning during real-time language comprehension. The bulk of this evidence has come from research on native language (L1) processing, yet there is mounting evidence that the same applies to the processing of a non-native language (L2; see Kaan, 2014, for review). A matter of continued investigation is the relative weighting of different cues, and in particular how this relative weighting may differ depending on factors such as social and discourse context, comprehenders' processing goals, and of key interest here, listeners' linguistic knowledge and experience. The proposal that the relative weighting of cues in L2 comprehension and processing differs from that in L1 processing is not new, as we will review in more detail below. Our goal here is to extend and examine this proposal in the context of prediction in L2 processing. To this end, we report findings from a visual world eye-tracking experiment with L1 and L2 speakers of Mandarin Chinese, which was designed to assess listeners' relative reliance on two concurrently available cues.[i] More specifically, we ask whether L2 listeners allocate greater weight to lexico-semantic than to grammatical cues encoded by prenominal sortal classifiers when creating expectations about an upcoming noun in a sentence, and to what extent this relative weighting of cues differs from what we observe in L1 processing.

Differences in relative cue weighting in L2 vs. L1 processing

In the history of Second Language Acquisition (SLA) research, the idea that L2 learners prioritise certain sources of information over others, and that they may do so in a manner

that is different from what L1 speakers do, has been put forward in a number of different contexts and from a variety of theoretical vantage points. Cue competition is at the heart of the Competition Model (Bates & MacWhinney, 1989), which posits that a key aspect of learning a language lies in determining the relative importance of different morphosyntactic and semantic cues to sentence interpretation. In English, word order tends to be the most reliable cue for identifying the subject and object of a simple transitive clause, and native English speakers prioritise this cue over other potentially informative sources, such as subject-verb agreement and noun animacy. In Spanish, on the other hand, subject-verb agreement tends to be more informative, whereas in Chinese, noun animacy plays a greater role. Research conducted within this framework has shown that L2 learners tend to over-rely on cues that have high reliability and validity in their L1. Accordingly, Spanish learners of English have been shown to rely more heavily on subject-verb agreement (Hernández, Bates, & Avila, 1994), and Chinese learners of English weigh noun animacy more heavily than native English speakers (Liu, Bates, & Li, 1992).

Whereas the allocation of cue weight within the Competition Model has been conceived as determined primarily by cue validity in the L1, others have proposed a more general preference for certain information sources in L2 processing, independent of the nature of learners' L1. Gass (1986), for example, proposed that at early stages of L2 development, learners rely more strongly on semantic than on syntactic heuristics for sentence interpretation, with a gradual shift to greater reliance on syntax with increasing proficiency. In a similar vein, VanPatten (2004) put forward "The Lexical Preference Principle" as a principle of L2 input processing. This principle states: "Learners will tend

to rely on lexical items as opposed to grammatical form to get meaning when both encode the same semantic information" (p. 9). VanPatten draws attention in particular to cases where information from one cue may be redundant in the presence of another. A scenario where such redundancy arises in many languages, including English, is reference to past events, which is often marked redundantly by tense inflection (*-ed*) and adverbs (e.g., *yesterday*). This scenario has been investigated extensively by Ellis and colleagues, who have shown that learners of Latin generally rely more heavily on temporal adverbs than on verbal inflection when judging temporality in simple Latin sentences, although this preference depends at least in part on learners' L1 (Ellis & Sagarra, 2011), and it can be shifted to some degree through specific manipulations of the input during training (Cintrón-Valentín & Ellis, 2016).

The proposals reviewed so far have been investigated primarily through learners' offline sentence interpretation and judgments. Similar proposals have been advanced in the context of real-time sentence processing, notably within Clahsen and Felser's (2006, 2018) Shallow Structure Hypothesis, which holds that "grammatical constraints may be less robust in the L2 than in the L1 and that L2 processing tends to rely more on nongrammatical information than on the grammatical route to interpretation than L1 processing" (Clahsen & Felser, 2018, p. 701). More specifically, Clahsen and Felser propose that "L2 processing may prioritize semantic, pragmatic, or other types of nongrammatical information, with L2 speakers potentially being *more* sensitive to these types of information compared to L1 speakers", adding that "this part of our hypothesis has rarely been investigated" (2018, p. 695). While there is some existing work suggesting that L2 learners may rely more on contextual and/or pragmatic cues than

native speakers do (Foucart et al., 2015; Pan & Felser, 2011; Pan, Schimke, & Felser, 2015), we are not aware of any studies that have directly compared reliance on SEMANTIC versus grammatical cues in real-time L2 sentence processing.

Indirect evidence suggesting that L2 learners are more apt at using semantically informative cues, as opposed to cues whose informativity is purely grammatical, comes from studies of gender-marked determiners in Spanish. Lew-Williams and Fernald (2007, 2009, 2010) showed that native speakers of Spanish take advantage of feminine and masculine gender marking on determiners to facilitate the processing of an upcoming noun; this is the case for nouns referring to humans (*la chica, el chico*; 'the girl', 'the boy'), where gender coincides with sex and is thus semantically informative, but also for inanimate nouns (*la pelota, el zapato*; 'the ball', 'the shoe'), where gender marking is informative at a purely grammatical level, indicating noun class membership. Importantly, English-speaking learners of Spanish in 2nd and 3rd year college-level Spanish classes showed the same processing benefits as native speakers in the case of semantically informative gender marking (Lew-Williams & Fernald, 2009), whereas they did not appear to be able to use the purely grammatical information provided by gender marking on determiners preceding familiar inanimate nouns (Lew-Willams & Fernald, 2010). This asymmetry is intriguing. Yet given that it arises from a comparison across different experiments, stimuli and learners, it cannot speak directly to learners' relative weighting of cues from different information sources when both cues are present at the same time. In the experiment we present here, we take advantage of properties of the Chinese noun class system, where the assignment of nouns to classifier classes involves both semantically transparent and intransparent criteria, as described in more detail

below. It is this dual informativity – semantic as well as grammatical – of classifiers in Chinese that allows us to assess more directly what relative weight L2 learners allocate to these cues, and to what extent this weighting differs from that of native speakers.

Classifiers in L1 processing

Classifiers – also known as measure words – are free morphemes marking the class of the noun they co-occur with. In Mandarin Chinese, their presence is generally obligatory when the noun phrase includes a numeral or demonstrative, as illustrated in (1).

(1) 一 条 绳子

yī * (tiáo) shéngzi

one $CL_{LONG\text{-}THING}$ rope

'a rope'

Sortal classifiers, the type of interest here, categorise nouns into classes broadly based on inherent properties of the object they denote, such as shape, natural kind, and function (Gao & Malt, 2009). While several hundreds of classifiers have been documented in Mandarin Chinese (Chen, Che, Chen, & Zhang, 1988), according to a recent estimate, approximately 50 to 70 sortal classifiers are used frequently in modern Chinese (Li, 2013). The number of nouns associated with a particular classifier varies greatly, and although classifiers are generally associated with certain semantic properties, nouns that belong to the same classifier class do not always form a homogeneous semantic set. The classifier *tiáo*, for instance, is generally associated with nouns that conform to the

description “slender, long-shape thing, often flexible” (Gao & Malt, 2009, p. 1171). Importantly for the purposes of our experiment, however, not all nouns that co-occur with *tiáo* fit this description to the same degree. For example, both *shéngzi* ‘rope’ and *gǒu* ‘dog’ are members of the *tiáo* class, yet only the former straightforwardly fits the semantic criteria associated with the class. In other words, the semantics associated with *tiáo* is predictive of *shéngzi* ‘rope’ to a greater degree than it is of *gǒu* ‘dog’. However, if predictions are based purely on form class membership, then *tiáo* should be equally predictive of both. Moreover, possessing relevant semantic attributes does not necessarily guarantee class membership. For example, a wristwatch could reasonably be described as ‘slender, long-shaped, and flexible’; nevertheless, *shǒubiǎo* (‘wristwatch’) cannot co-occur with the classifier *tiáo*. Thus for both *gǒu* (‘dog’) and *shǒubiǎo* (‘wristwatch’), (non)membership in the *tiáo* class cannot be inferred on purely semantic grounds. Instead, similar to grammatical gender with inanimate nouns in Spanish, class membership is an abstract grammatical property of the noun that is reflected primarily through co-occurrence with a specific classifier. In sum, classifiers serve both grammatical and semantic functions. For most classifier-noun pairs, the grammatical form class and the semantic information encoded by the classifier overlap completely, making it impossible to tease apart whether it is the grammatical form class association, or the semantic features, that facilitate the processing of the subsequent noun. However, the fact that cases like *gǒu* (‘dog’) and *shǒubiǎo* (‘wristwatch’) exist demonstrates that grammatical form-class and semantic relatedness betweeen classifiers and nouns are dissociable. This dissociation will be critical for the design of the present study.

A number of previous studies have investigated the role of classifiers in the real-time sentence processing of native speakers of Chinese. In a visual world study, Huettig, Chen, Bowerman, & Majid (2010) showed that L1 Mandarin speakers were faster to look at a target object when it was named preceded by an informative classifier, i.e., a classifier compatible only with the target noun in the visual scene, than when no classifier was present in the linguistic signal (see also Klein, Carlson, Li, Jaeger, & Tanenhaus, 2012). These findings indicate that native Chinese speakers derive similar processing benefits from prenominal classifiers as native Spanish speakers do from gender-marked determiners. Yet as these studies did not systematically manipulate or control the prototypicality of nouns within their classifier classes, it remains unclear whether these processing benefits are the result of the classifier as a grammatical form-class cue (similar to gender-marked determiners preceding inanimate nouns in Spanish) or its semantic informativity (similar to gender-marked determiners preceding nouns denoting referents with biological sex). The facilitation observed in these studies, therefore, could be the result of preactivation of nouns matching the classifier in FORM-CLASS, or of nouns matching the SEMANTIC FEATURES associated with the classifier, or both.

In order to tease these two sources of informativity apart, Tsang and Chambers (2011) conducted two visual world experiments in which L1 speakers of Cantonese were presented with visual scenes that in addition to the object named in the auditory input also contained an object that had semantic attributes of the target noun's classifier class but was not a member of that class. Tsang and Chambers referred to these as "G-S+ competitors", indicating that they constitute a grammatical mismatch (G-) but a semantic match (S+) for the classifier in the auditory stimulus. Their findings showed that when

the target was a prototypical member of the class (i.e., displaying all its defining semantic features), G-S+ competitors attracted no more looks than unrelated distractors. By contrast, robust competition effects were observed from competitors from the same classifier class (G+), with slightly stronger effects for prototypical (S+) than non-prototypical (S-) members of the class, indicating that semantic information contributed to listeners' expectations, but only within the set of class-consistent nouns. In a second experiment, Tsang and Chambers (2011, Experiment 3) sought to increase the potential for competition by including targets that themselves were not prototypical members of their class. Under these circumstances, a small and short-lived increase in looks to G-S+ competitors was observed. The authors concluded that sortal classifiers influence L1 speakers' predictive processing "primarily through their grammatical constraints" (p. 1065), with classifier semantics acting as a secondary cue that becomes apparent only in certain circumstances, such as when non-prototypical members of a class are involved. In the study we present here, we followed the design of Tsang and Chambers' (2011) Experiment 3 to create an analogous experiment in Mandarin, with the goal of testing whether competition from G-S+ competitors would be stronger, indicating greater reliance on semantic information, among L2 than among L1 listeners.

Classifiers in L2 processing

The role of classifiers in L2 real-time listening was first investigated by Lau and Grüter (2015) in a small-scale visual world study with intermediate-proficiency L2 learners of Mandarin, employing an experimental design closely following that used in Lew-Williams and Fernald's (2007) study of gender-marking in Spanish. Although tempered

by small sample size and limited statistical power, their findings provided at least preliminary evidence that L2 learners were able to take advantage of information encoded on the classifier to facilitate processing of the subsequent noun, with a clear trend towards increased and earlier looks to the target when the classifier was informative.

More recently, Mitsugi (2018) presented findings from a visual world study with 25 L1 and 25 L2 speakers of Japanese, using a similar experimental design. Participants listened to auditory stimuli consisting of simple sentences containing one of two numeral classifiers (*-hon*, used for counting long, string-like objects; *-mai*, for thin, flat-surfaced objects) followed by a noun, and were instructed to click on the matching image in a two-choice picture identification task. Eye-gaze patterns were analyzed in two 1,200-ms temporal windows: a 'predictive region' extending from classifier onset to noun onset, and a 'critical word region' beginning at the onset of the noun (both offset by 200 ms). The results of growth curve models showed a significant effect of Condition in the predictive region, which importantly did not interact with Group, indicating that both L1 and L2 speakers of Japanese took advantage of informative classifiers to direct their gaze to the target before it was named. Analyses of the critical word region revealed somewhat slower and more variable convergence on the target, modulated by proficiency. Proficiency, as assessed by a self-rating questionnaire, however did not modulate any effects in the predictive region. These findings from L2 Japanese confirm and corroborate the trends observed by Lau and Grüter (2015) in L2 Mandarin, and indicate that classifiers present informative cues that facilitate referential processing not just for L1 speakers of Chinese and Japanese, but also for L2 learners at intermediate-to-advanced levels of proficiency.

Importantly, the materials in both of these studies were comprised of nouns that constitute prototypical members of their classifier class. Thus although both Lau and Grüter (2015) and Mitsugi (2018) appeal to the semantic transparency of the cues encoded on classifiers to explain the discrepancy with findings from previous work on grammatical gender, these explanations must remain speculative as long as semantic and grammatical/form-class informativity are conflated in the materials. The design of the experiments by Tsang and Chambers (2011) discussed above presents an ideal solution to tease the two apart. Since Mandarin Chinese is more widely taught as a foreign language than Cantonese, we thus translated and adapted Tsang and Chambers' materials from Cantonese to Mandarin, with the goal of investigating what aspects of classifiers' informativity L1 and L2 listeners draw on during real-time processing. More specifically, we ask: **Do L1 and L2 listeners differ in their use of grammatical form-class information vs. semantics encoded by classifiers to predict an upcoming noun?** Based on previous proposals in the SLA literature, which as discussed above, converge in the general claim that L2 learners tend to rely more heavily on semantics than on grammatical information, we predict that reliance on semantic vs. form-class information will be greater among L2 than among L1 listeners. In terms of Tsang and Chambers' (2011) experimental design, this leads to the prediction of greater competition effects from G-S+ competitors in the L2 than in the L1 group.

Method

Participants

A total of 96 participants took part in this study; 38 identified as native and 58 as non-native speakers of Chinese. The 38 native speakers were recruited from among the international student community at the University of Hawai‘i. Based on information provided in the language background questionnaire, 3 were identified as speakers of Cantonese. Since the classifier systems of Mandarin and Cantonese are not identical, data from these participants was excluded. Data from 11 native speakers had to be excluded from analysis due to poor quality of the eye gaze data, mostly due to difficulties with calibration. The remaining 24 were all born and raised in mainland China or Taiwan, self-identified as native speakers of Mandarin, rated their speaking and listening ability in Mandarin at least 8 out of 10, and indicated using Mandarin on a daily basis. Four speakers in the L1 group also indicated exposure to another Chinese dialect in their childhood homes. All named Mandarin as the language they felt most comfortable with in casual conversation at the present time. Data from these 24 L1 speakers of Mandarin (18 female, mean age: 26 years, range: 19-39) entered the analysis.

L2 speakers were recruited at the University of Hawai‘i (N=28), as well as at Peking University (N=20), the University of Hong Kong (N=7), and the Chinese University of Hong Kong (N=3). Speakers who were taking or had taken 3rd-year Chinese classes (or above) in the U.S., or intermediate/advanced classes in China, were admitted to the study. Data from participants who indicated significant exposure to a Chinese language during childhood (n=12) was excluded from analysis. Data from 3 participants was excluded due to poor quality of the eye gaze data. Of the remaining 43 L2 participants, 34 identified as native speakers of English; other L1s included Spanish (n=3), German (n=2), Dutch, Hebrew, Japanese and Vietnamese (n=1 each). Since

experience with a classifier system in the L1 was considered a potentially confounding factor, data from the L1 speakers of Japanese and Vietnamese was excluded. Thus data from 41 L2 speakers of Mandarin (15 female, mean age: 28 years, range: 20-57) entered the analysis. All of these L2 speakers indicated first exposure to Chinese at or above age 13.

All participants were asked to rate their proficiency in Chinese (speaking, listening, writing, reading, overall) on a scale of 0-10 and complete a brief in-house listening comprehension task.[ii] No significant differences were found between L2 participants tested in the U.S. and China on either self-rating skills (all $t < 1.09$, $p > .28$) or the listening comprehension test (Wilcoxon $W = -0.09$, $p = .93$). Mean self-ratings were significantly higher in the L1 than in the L2 group for all skills (all $t > 7$, $p < .001$; overall proficiency: $M_{L1} = 9.3$, $\text{range}_{L1} = 7\text{-}10$, $M_{L2} = 5.7$, $\text{range}_{L2} = 2\text{-}8$). Scores on the listening comprehension task were generally high in both groups ($M_{L1} = 98.6$, $\text{range}_{L1} = 90\text{-}100$; $M_{L2} = 90.3$, $\text{range}_{L2} = 50\text{-}100$), indicating that the task may have been too easy to capture substantial variability among learners. The difference between groups was nevertheless significant (Wilcoxon $W = 241.5$, $p = .001$).

The study protocol was approved by the Institutional Review Board at the University of Hawai‘i, and participants received a small amount of monetary compensation.

Materials and Procedure

Visual World Experiment

The visual world experiment was designed in analogy to Experiment 3 in Tsang and Chambers (2011, henceforth T&C). Linguistic stimuli consisted of questions as in (2).[iii]

(2) 哪 一 条 是 狗?

Nǎ yī tiáo shì gǒu?

Which one CL is dog?

'Which one is a/the dog?'

The 12 experimental target items contained one of three sortal classifiers (条, *tiáo*, '~long, flexible'; 支, *zhī,* '~stick-like, long'; 张, *zhāng*, '~flat, spread open'). These classifiers were chosen because they are the cognates of the three classifiers used in T&C's Cantonese experiment (*tiu4, zil*, and *jeungl*), and they are introduced in the textbook used in first-year Chinese classes at the University of Hawai'i. The 12 target nouns were chosen to constitute non-prototypical members of their classifier class (see Appendix for a list of all nouns). For example, *gǒu* 'dog' appeared as a target with the classifier *tiáo*, which is generally associated with objects that can be described as long, slender and flexible (e.g., Gao & Malt, 2009). These descriptors do not straightforwardly apply to 'dog'; nevertheless, *gǒu* commonly co-occurs with *tiáo*. Adopting T&C's terminology, targets can thus be characterised as 'G+', reflecting a grammatical match between classifier and noun, and 'S-', indicating a mismatch between the semantic features of the noun and those associated with the classifier (class). In order to substantiate our assumptions regarding what constitutes a semantic (mis)match (S-/+) between the nouns and classifiers in our stimuli, we conducted an independent rating

study (see below, and Supplementary Materials). We followed T&C's rationale for using G+S- items as targets in order to increase the potential for competition from semantically matching (S+; G+ or G-) referents in the visual scene.

Twelve referent sets were created, each consisting of one target, three competitors, and one distractor (see Appendix). Of the three competitors, one consisted of a noun from the target classifier class instantiating the semantic features of the class (G+S+), i.e., a prototypical member of the class (e.g., *shéngzi* 'rope' for a *tiáo* target). The second consisted of a noun that cannot co-occur with the target classifier, but nevertheless possesses semantic properties associated with that class (G-S+). For example, *shǒubiǎo* 'wristwatch' was selected as a G-S+ competitor for a target in the *tiáo* class because (a) *shǒubiǎo* cannot co-occur with the classifier *tiáo*, and (b) it can reasonably be described as 'long, slender and flexible'. The third competitor consisted of a noun that cannot co-occur with the target classifier and does not have any of the semantic attributes associated with the class (G-S-; e.g., *píngguǒ* 'apple' for a *tiáo* target). Finally, distractors were selected by the same criteria as G-S- competitors, that is, they were grammatically and semantically unrelated to the target classifier class.

In selecting target, competitor and distractor nouns, we began by translating T&C's Cantonese stimuli into Mandarin. The item was retained if it fit the following criteria: the translation (i) belonged to the crosslinguistically corresponding class (e.g., *tiu4*/*tiáo*), (ii) was compatible with only one of the classifiers used in the experiment, and (iii) was likely to be familiar to 3rd-year learners of Chinese (based on inspection of textbook vocabulary). Forty-two (out of 60) items were retained.[iv] The remainder were replaced with Mandarin nouns that fit the requirements of the design. All nouns were

judged (by the second and third author) to be familiar to 3rd -year learners of Chinese, and appeared >50 times in the Mandarin SUBTLEX-CH corpus (Cai & Brysbaert, 2010).

Using frequency values (log10W) from the SUBTLEX-CH corpus, we calculated an index of differential frequency for each target-competitor pair (target-log10W minus competitor-log10W) in order to assess to what extent eye gaze may be driven by differences in lexical frequency. Mean values in all conditions were positive (G+S+: M = .41, SD = .68; G-S+: M = .76, SD = .75; G-S-: M = .10, SD = .90), indicating targets were somewhat more frequent overall than competitors. A one-way ANOVA revealed no significant differences between the three conditions, $F(2,33) = 2.2$, $p = .13$, suggesting that differences in lexical frequency are unlikely to contribute substantially to differential looks to competitors across conditions.

In order to assess whether our largely subjective assessment of items' semantic fit to a classifier class (S+/- status) was justified, we conducted an independent rating study including the 60 nouns and images representing target, competitor and distractors in the visual world experiment. The results of the rating study, described in more detail in the Supplementary Materials, showed that items classified as S+ (G+S+ and G-S+ competitors) were rated as significantly more aligned with the semantic features of the classifier they appeared with in the visual world experiment than items classified as S- (G+S- targets, G-S- competitors and unrelated distractors).

Visual scenes included color clipart images of three objects: the target, one of the three competitors, and a distractor (Figure 1). Items in the three conditions (G+S+, G-S+, G-S-) were rotated across three lists, such that each participant saw each target only once in one of the three conditions, for a total of 12 experimental trials, with 4 items in each of

3 conditions. The order of items was pseudo-randomised, and interspersed with 24 filler trials. Fillers consisted of items where the two unmentioned objects were of the same class, items where objects differed by color only, and items where objects shared perceptual or semantic properties but where not from the same classifier class.

<Insert Figure 1 about here>

Auditory stimuli were recorded by a female native speaker of Mandarin in a sound-proof booth at 44.1 kHz, and edited using Praat (Boersma & Weenink, 2017). Experimental stimuli were constructed by concatenating extracted tokens of *nǎ yī* ('which one'), the classifier, *shì* ('is'), and the target noun. The duration of the first three parts was held constant across all items. Silence was added after the classifier and after *shì* such that the duration from classifier onset to noun onset was exactly 1,150 ms in each experimental item. This 1,150-ms time period constitutes the critical region for analysis. Two native Mandarin speakers and two L2 learners checked and confirmed the naturalness of all stimuli.

The experiment was conducted on an SMI RED250 eye-tracker sampling at 250 Hz (for L1 and L2 participants tested in Hawai'i, or a mobile REDn Scientific eye-tracker sampling at 60 Hz (for L2 participants tested in China). Each trial began with a 2,000-ms display of the visual scene, followed by the question. Participants were instructed to click on the correct image in answer to the question. Gaze and mouse-click responses were recorded through SMI Experiment Suite software. Gaze data were classified automatically as fixations, saccades and blinks by the software using default settings.

Fixations were subsequently binned into 20-ms samples for further analysis. Preliminary analyses showed no differences in the structure of the data from the two trackers, thus all data was combined for further analysis.

Vocabulary Test

A vocabulary test was created to assess participants' knowledge of the target classifier-noun pairings. The test consisted of 50 4-alternative forced-choice items, including the 12 target nouns from the Visual World experiment. In these 12 critical items, participants had to complete a numeral or demonstrative phrase (NUM/DEM CL N) with the most suitable classifier, as illustrated in (3). English translation of the phrase was given to ensure the intended meaning of the phrase was clear. Answer options for the critical items did not include the general classifier (个 *ge*), which is known to be overused by L2 learners (Polio, 1994; Zhang & Lu, 2013). The four options were chosen with the intent that one would be clearly more suitable than the other three; this was done by consent between two Chinese speakers (the second and third author).

(3) 一 _____ 鱼

a/one fish

○ 条 ○ 张 ○ 只 ○ 台

Critical items were interspersed with 38 items of the same format, in which missing material consisted of a variety of lexical and functional morphemes. The large number of filler items was necessary because the vocabulary test had to be completed before the

visual world experiment, in order to exclude the possibility that participants' performance on the vocabulary test could reflect learning effects from the main experiment. Interspersing the critical classifier items with fillers was thus intended to avoid that participants would guess that the focus of the study was on classifiers.

The vocabulary test was implemented as a web-based survey, which participants were asked to complete at least four days before the test session in order to minimise any priming effects from the target items in the vocabulary test on performance in the visual world experiment. While not all participants adhered to this schedule, all completed the vocabulary test at least one day before the visual world task, and most completed it substantially earlier (M = 8.7 days, SD = 6.7 days, range: 1–30 days).

Results

We begin by reporting the results from the vocabulary test, as they inform the analysis and interpretation of the eye-gaze data.

Vocabulary Test

One L1 participant did not complete the vocabulary test. Mean accuracy – operationalized as the selection of the option deemed most suitable by the second and third author – across all 50 items was 94.5% (88-100%) for the L1 group, and 72.6% (38-94%) for the L2 group. For the 12 target classifier items alone, however, accuracy was lower, especially in the L2 group (M_{L1} = 89.9%, $range_{L1}$ = 75-100, M_{L2} = 46.3%, $range_{L2}$ = 8-92). Within the L2 group, accuracy on classifiers correlated significantly with

performance on the listening comprehension test (Kendall's tau = .34, p = .006), and marginally with self-rated proficiency (Kendall's tau = .22, p = .07).

Target items and classifiers had been selected carefully from among high-frequency vocabulary introduced in 1st and 2nd year Chinese textbooks, with the expectation that the majority of L2 participants would be able to identify the expected classifier for the majority of nouns. The fact that this expectation was not met is testimony to the general observation that classifiers are exceedingly challenging to master for even advanced L2 speakers (Polio, 1994; Li, 2010). It is worth noting, however, that unlike for grammatical gender in European languages, some variability exists even among native speakers with regard to classifier-noun associations. This is evidenced in the data from our L1 group, where the selection of the classifier originally deemed most suitable varied by item, between 65% and 100%. For the item with 65% accuracy, 张弓 (*zhāng gōng,* CL 'bow'), all 8 native speakers who did not choose the expected 张 instead selected 把 (*ba*), a classifier generally used for objects with handles. Although *zhāng* is the default classifier for 'bow', *ba* is also allowed in certain contexts, especially when the property of having a handle is relevant. Notably, the same item also had the lowest number of expected responses in the L2 group, with *ba* also being the most frequently selected non-target option. This suggests that choice of a non-expected option on this test does not necessarily reflect lack of knowledge of classifier usage. Further investigation of this variability in L1 Mandarin presents an interesting avenue for future research.

Nevertheless, the fact that in more than half of the cases, L2 participants did not select the classifier with which a given noun was paired in the visual world experiment, required us to reconsider our original analysis strategy for the eye gaze data. The original

strategy, consistent with previous related studies on grammatical gender (e.g., Hopp, 2013), was to exclude all trials with target nouns for which the participant did not provide the expected classifier in the vocabulary test. This strategy relied on the assumption that only a small portion of trials would have to be excluded, thus not distorting the overall dataset in any major ways. This assumption was not met by the data from the vocabulary test. We therefore decided to conduct two separate analyses of the eye gaze data. In Analysis 1, no trials were excluded based on participants' performance on the vocabulary test, thus representing listeners' incremental comprehension of classifier-noun sequences more generally, including cases where the association between a given classifier and noun was unknown or dispreferred. This dataset retains the counterbalancing of items and conditions in the original design, and is, arguably, more representative of listening comprehension outside the lab, where one-to-one mappings between classifiers and nouns are not always given—for L2 listeners in particular, but as shown by the variability in the L1 group's performance on the vocabulary test, to a lesser extent for L1 listeners as well. We then present our originally intended analysis as Analysis 2, including only the subset of trials with target nouns for which the participant selected the expected item on the vocabulary test. The exclusion of approximately 10% of the L1 data and 50% of the L2 data from Analysis 2 substantially reduces statistical power. It also raises questions about the generalizability of the patterns observed in the remaining data from the L2 group. At the same time, Analysis 2 presents a more rigorous test of listeners' relative reliance on form-class vs. semantic information encoded by classifiers WHEN BOTH OF THEM ARE KNOWN.

Visual World Experiment

Mouseclick accuracy. Participants' accuracy in selecting the correct image was high overall (M = 93.4%, SD = 9.8), with variability by group (M_{L1} = 98.6, M_{L2} = 90.4) and by condition ($M_{G\text{-}S\text{-}}$ = 95.8, $M_{G\text{-}S+}$ = 94.2, M_{G+S+} = 90.4). A logistic mixed-effect regression analysis (lmerTest, Kuznetsova, Brockhoff, & Christensen, 2017) with Group (contrast-coded, centered) and Condition (treatment-coded; baseline G-S+) as fixed effects (ACCURACY ~ CONDITION + GROUP + (1 | PARTICIPANT) + (1 | ITEM)) showed that L1 speakers were significantly more likely overall to select the correct image than L2 speakers (b = 2.6, p < .001). The overall likelihood of correct choice was marginally lower in the G+S+ compared to the G-S+ (b = -.78, p = .06) condition, while the difference between the G-S- and G-S+ conditions was not significant (b = .53, p = .3). In order to examine the contrast between the G-S- and G+S+ conditions, the model was rerun with the reference level for Condition changed to G-S-. Results showed that the likelihood of correct choice was significantly lower in the G+S+ than in the G-S- condition (b = -1.31, p = .004). Adding the interaction between group and condition did not improve model fit, indicating that the differences by condition were comparable in the two groups.

The decreased accuracy in the G+S+ condition is consistent with the prediction that interference will be greatest when the competitor represents a prototypical member of the classifier class, and suggests that such interference can even impact final answer choices. Yet inaccurate choices may also be reflective of inattention, or lack of knowledge of the target noun. For this reason, all trials with inaccurate mouseclick responses (L1: 4/287, L2: 47/492) were excluded from further analyses.

Eye gaze, Analysis 1: Including all trials with correct mouseclick. A total of 726 trials (L1: 283, L2: 443) originally entered into Analysis 1.[v] Participants' looking behavior, including data from all trials with correct mouseclick responses, is illustrated in Figure 2. For each participant group and condition, Figure 2 shows the difference in the mean proportion of looks to the target minus the mean proportion of looks to the competitor over the course of an experimental trial. A positive value thus indicates more looks to the target than to the competitor, a negative value indicates more looks to the competitor, while zero indicates an equal proportion of looks to the target and the competitor.

<Insert Figure 2 about here>

Visual inspection of looking patterns in the L1 group indicates little difference between the G-S- and G-S+ conditions after the onset of the classifier. This suggests that when competitors were not a grammatical match for the classifier (G-), whether or not their semantic features were aligned with those of the classifier (S+/-) did not influence L1 listeners' expectations. In the G+S+ condition, by contrast, L1 listeners preferentially looked at the competitor – the prototypical member of the classifier class – until after the onset of the (non-prototypical) target noun, indicating that when two candidates presented a grammatical match for the classifier (G+), the one presenting the better semantic match was preferred. In the L2 group, looking patterns in the three conditions begin to diverge approximately 400 ms before the onset of the noun. This presents an indication that L2 listeners, like L1 listeners, were using information encoded by the classifier proactively

to identify the upcoming noun. Notably, unlike in the L1 group, gaze patterns in the G-S- and G-S+ conditions diverge, with temporarily more looks to the competitor in the G-S+ than in the G-S- condition, suggesting that L2 listeners consider referents possessing semantic features associated with the classifier class even when the referent is not a grammatical match for the classifier.

In order to assess participants' relative likelihood to look at the target vs. the competitor statistically, we calculated a 'TargetAdvantage' score for each trial during the critical region of interest. On the standard assumption that it takes approximately 200 ms to execute a ballistic eye movement in response to an acoustic stimulus (Matin, Shao, & Boff, 1993), our region of interest begins 200 ms after the onset of the classifier and extends to 200 ms after the onset of the noun, comprising a temporal window of 1,150 ms. TargetAdvantage scores were calculated by subtracting the number of 20-ms bins in this temporal window that contained looks to the competitor from the number of bins containing looks to the target.[vi] Prior to this calculation, all data points from fixations initiated prior to 200 ms after classifier onset were excluded (following Tsang and Chambers, 2011).[vii]

We used linear mixed-effect regression to model the effects of Condition (treatment-coded; baseline G-S+) and Group (contrast-coded, centered) on TargetAdvantage scores. All models were implemented in R (R Core Team, 2018) using the lmerTest package (Kuznetsova et al., 2017). Fixed effects are presented in Table 1.[viii] Note that the effect of Group in this model reflects a simple effect, i.e., the effect of Group in the baseline G-S+ condition. This effect is negative and significant (b = -9.00, p = .001), indicating that the L2 group looked less at the target – and thus more at the

competitor – than the L1 group when the competitor was a semantically compatible yet form-class incompatible referent. Importantly, no significant effect of Group emerged in the G-S- (b = -1.95, p = .47) condition, whereas in the G+S+ (b = 5.70, p = .04) condition, a significant POSITIVE effect emerged (based on the same model rerun with the respective reference levels for Condition). This indicates that L2 listeners were not GENERALLY less likely to look at the target than L1 listeners, and that L1 listeners appear to be affected more strongly by the prototypicality of referents within a classifier class. The model output also indicates fewer looks to the target in the G+S+ compared to the (baseline) G-S+ condition overall (b = -5.83, p = .003), and this effect is qualified by an interaction with Group (b = 14.70, p < .001). No overall differences are observed between the G-S- and the G-S+ condition, yet a marginal interaction with Group (b = 7.05, p = .07) emerged. To further probe the nature of these interactions, follow-up models were fit to the data from each group separately. In the L1 group, there were no differences between the G-S+ and G-S- conditions (b = -1.70, p = .57), yet significantly fewer looks to the target in the G+S+ condition (vs G-S+: b = -14.59, p < .001; vs G-S-: b = -12.89, p < .001). In the L2 group, by contrast, there were no differences between the G-S+ and G+S+ conditions (b = .09, p = .97), and significantly more looks to the target in the G-S- condition (vs G-S+: b = 5.29, p = .03; vs G+S+: b = 5.20, p = .03).

<Insert Table 1 about here>

In sum, the picture that emerges is that in the L1 group, the two G- conditions (G-S-, G-S+) pattern together and differ from the G+ condition (G+S+), whereas in the L2

group, the two S+ conditions (G-S+, G+S+) align and differ from the S- condition (G-S-). In other words, the form-class (G) cue drives differences in fixations among L1 listeners, while semantic compatibility (S) appears to be the primary driver of eye gaze among L2 listeners. This is further supported by the observation that when the two groups were compared directly, it was only in the G-S+ condition that the L2 listeners were more likely to look at the competitor than the native speakers. These findings are consistent with our hypothesis that the relative weighting of semantic versus grammatical cues is greater among L2 than among L1 listeners in real-time comprehension. In Analysis 2, we probe whether this pattern still holds when only items with correct classifier selection in the vocabulary test are included.

Eye gaze, Analysis 2: Including only trials with correct classifier selection in the vocabulary test. Data from a total of 435 trials (L1: 239, L2: 196) entered into Analysis 2. Figure 3 illustrates participants' fixation patterns in this reduced dataset. The same models as in Analysis 1 were fit to the reduced dataset; fixed effects are presented in Table 2. The output from this model is similar to that from Analysis 1. The simple effect of Group in the (baseline) G-S+ condition remains significant (b = -9.94, p = .007), indicating that L2 listeners are fixating on the target less – and on the competitor more – than L1 listeners when the competitor is semantically compatible yet form-class incompatible. As in Analysis 1, the groups did not differ in the G-S- condition (b = -4.15, p = .24), and the L2 group looked to the target significantly MORE than the L1 group in the G+S+ condition (b = 8.41, p = .02). The output in Table 2 also shows a significant difference between the (baseline) G-S+ and G+S+ conditions, qualified by an interaction

with Group. Follow-up models on the data from each group show the same pattern for the L1 group as in Analysis 1: no differences between the G-S+ and G-S- conditions (b = -1.19, p = .72), and significantly fewer looks to the target in the G+S+ condition (vs G-S+: b = -14.51, p < .001; vs G-S-: b = -13.31, p < .001). In the L2 group, no between-condition comparisons are significant (all $|t|$ < 1.5, p > .18).

<Insert Figure 3 about here>

<Insert Table 2 about here>

In sum, when including only items with correct responses on the vocabulary test, the same pattern as in Analysis 1 emerges for the L1 group: When the competitor is not a form-class match, its semantic compatibility with the target is of no relevance (no difference between G-S- and G-S+). For the L2 group, unlike in Analysis 1, the difference between the G-S- and G-S+ conditions is no longer significant (b = 4.74, p = .19). When comparing the two groups directly, however, L2 listeners continue to be signficantly more likely to look at the competitor than native speakers IN THE G-S+ CONDITION ONLY. Thus while the differences between conditions in the L2 group become weaker and non-significant in Analysis 2, an outcome impacted at least in part by reduced statistical power due to the elimination of approximately 50% of the data, the difference between the L1 and the L2 group in the G-S+ condition remains robust. This indicates that even when taking L2 learners' knowledge of classifier-noun associations into consideration, there continues to be support, albeit somewhat weaker, for the

hypothesis that L2 listeners allocate greater weight to semantic than to form-class cues encoded by classifiers than L1 listeners do.

Discussion and Conclusion

The goal of this study was to examine the relative weight allocated by L1 and L2 listeners to grammatical form-class versus semantic cues encoded by prenominal classifiers in Chinese. Following Tsang and Chambers (2011), we took advantage of the fact that Chinese classifier classes, although generally characterised by semantic and functional criteria, are not always fully transparent with regard to class-membership of individual nouns, at least to a present-day language user. In particular, there are semantically defined classes, such as the *tiáo*-class associated with long, slender and flexible objects, which contain nouns denoting objects whose affordances do not fit straightforwardly with the semantic features of the class (such as *gǒu* 'dog'). At the same time, there are nouns denoting objects whose affordances seem to align with these features, but are not admitted to the class (**yī tiáo shǒubiǎo* 'one wristwatch'). This allowed us to ask whether upon hearing a classifier like *tiáo*, listeners would use it as a grammatical marker of its class and thus preactivate only nouns belonging to the *tiáo*-class (including *gǒu* 'dog', but not *shǒubiǎo* 'wristwatch'), or alternatively, treat it as a semantic cue, similar to an adjective, and preactivate nouns with the relevant semantic features (including 'wristwatch', but not 'dog').

Our results from native speakers of Mandarin largely replicated Tsang and Chambers' (2011) findings from native speakers of Cantonese: L1 listeners used the classifier primarily as a grammatical, form-class cue, indicated by the fact that when the

competitor in the visual display was not from the target classifier class, it made no difference whether or not it had semantic features matching the defining features of the class. The brief temporary competition effect that Tsang and Chambers observed for semantically aligned non-class items (G-S+) was not replicated in our experiment with L1 speakers of Mandarin. Like Tsang and Chambers, however, we found that when L1 listeners had a choice between two class-consistent nouns, they favored the more prototypical one. This indicates that L1 listeners do not ignore the semantic dimension of classifiers, but they use this information only within the set of nouns already confined by the grammatical form-class cue. These findings thus support Tsang and Chambers' conclusion that native speakers use classifiers to create expectations about upcoming nouns "primarily through their grammatical constraints" (p. 1065), with classifier semantics acting as a secondary cue.

This conclusion, which arises from both Tsang and Chambers' and our visual world experiments with native speakers of two different varieties of Chinese, may appear to be at odds with recent evidence from ERP studies, which have found N400, but no P600, effects for mismatching classifier-noun pairs, and concluded that "combinatorial processing of classifier-noun sequences (…) is primarily semantically based" (Qian & Garnsey, 2016, p. 780; see also Chou, Huang, Lee, & Lee, 2014). This is an intriguing contrast that calls for further investigation. We note, however, that as far as we know, none of these ERP studies has included or focused on nouns whose grammatical form-class and semantic properties do not coincide, such as the 'dog' and 'wristwatch' items central to Tsang and Chambers' and our visual world experiments. Future work, ideally including such items, will be needed to characterise more precisely the nature of the

linguistic violation that a classifier-noun mismatch constitutes for native – and for non-native – speakers of Chinese.

What matters for the purposes of the present study is that classifiers constitute a dual source of information: one that can be characterised as purely semantic, similar to an adjective, and another that reflects a formal property of the grammar, namely the classifier system, similar to grammatical gender. It is the simultaneous presence of these two cues, and the potential conflict between them, that allowed us to ask whether L1 and L2 listeners would allocate different weight to semantic versus grammatical cues during real-time listening. In line with several theoretical proposals in the SLA literature, including the Shallow Structure Hypothesis (Clahsen & Felser, 2006, 2018), we hypothesised that L2 listeners would rely more heavily on the semantic than on the grammatical informativity of classifiers in comparison with L1 listeners. We thus predicted that competition from semantically aligned non-class items (G-S+) would be greater among the L2 than the L1 group. Our results broadly support this hypothesis. In both Analyses 1 and 2, i.e., both when including and when excluding items for which participants did not select the expected classifier in the offline vocabulary test, we found that L2 listeners were more likely to look at semantically matching non-class competitors than L1 listeners. Importantly, this was the case only for this particular type of competitor, thus excluding the possibility that the L2 group was generally slower or less proactive in fixating on targets.

Of note, all of these effects emerge in a temporal window prior to the noun, indicating that they must be driven by expectations arising from the nature of the classifier about the identity of the upcoming noun. In other words, they are evidence of

predictive processing, in both the L1 and the L2 group. These findings add to the already substantial body of evidence showing that prediction is a mechanism that contributes to human language processing in various contexts and under various circumstances, including L2 processing (Kaan, 2014; Kuperberg & Jaeger, 2016). At the same time, the differences that emerged between the two groups in this study highlight the observation that prediction is unlikely to work in the same way in all contexts and under all circumstances. As Federmeier (2007, p. 495) had noted in the context of language processing in aging populations, "predictive processing may not be the best – or even a viable – strategy for all individuals at all phases of the lifespan and/or in all processing situations." That is because prediction facilitates comprehension only if its benefits outweigh its costs, i.e., the likelihood of having to retreat from a false prediction and revise. As Kuperberg and Jaeger (2016, p. 32) argue, the level of prediction a language user will engage in is likely to be "a function of its expected utility, which, in turn, may depend on comprehenders' goals and their estimates of the relative reliability of their prior knowledge and the bottom-up input." This logic extends, we believe, beyond the LEVEL to which a comprehender engages predictive mechanisms, to the NATURE OF THE INFORMATION that s/he will use for creating predictions. If a certain type of information, say lexical semantics, presents a more reliable source of knowledge to base a prediction on than another type of information, say classifier-class membership, then the optimally adaptive language user will rely more heavily on the former than the latter. This is precisely what we argue L2 learners of Chinese are doing.

Importantly, this is likely to be the best and most efficient processing strategy for these language users. That is because the semantics of classifiers is likely to be a more

reliable (though by no means perfect) information source for L2 learners than grammatical form-class membership, due on the one hand to formal instruction, which often explicitly highlights the semantic definitions of classifier classes, but also to the nature of language learning after early childhood, which proceeds on the basis of more segmented input and, arguably, leads to weaker or less entrenched associative relations between frequently co-occurring elements, such as classifiers and nouns (for further discussion see Arnon & Ramscar, 2012; Grüter, Lew-Williams & Fernald, 2012, and Siegelman & Arnon, 2015, for grammatical gender; Paul & Grüter, 2016, for classifier-noun associations). By contrast, greater reliance on grammatical form class is likely the more efficient processing strategy for native speakers, due to the entrenched nature of these co-occurrence relations in their mental representations.

The differences between the two groups we observed in this study can thus be seen as a reflection of adaptation to differences in the relative reliability of prior knowledge. Viewed from this perspective, the question of whether L2 learners 'can achieve native-like processing efficiency' or 'come to learn to predict like native speakers' – both common framings of research on L2 processing – becomes somewhat problematic, as we come to recognise that maximal processing efficiency may be achieved through different means in different contexts and under different circumstances. Indeed, if L2 speakers were to rely on the same cues to the same extent as native speakers, their processing efficiency would likely decrease. Thus instead of framing our investigations of L2 processing in juxtaposition to L1 processing, and positing whatever is observed in the latter as a goal to be achieved in the former, we may gain more by considering L1 and L2 processing as just two among many instantiations of adaptive

human language processing. Such a perspective can allow us to side-step the often value-charged debate on 'deficits' in L2, and instead take L2 processing as an opportunity to examine and better understand how humans adapt their language processing given available knowledge, resources and task demands.

Acknowledgments

This research was supported by funding from *Language Learning*'s Small Grant Research Program. We thank Craig Chambers for sharing materials from Tsang and Chambers (2011), and Victoria Lee for assistance with data collection. We are also grateful for valuable feedback from the audiences at CUNY 2017, BUCLD 2017, and ISBPAC 2018, as well as from three anonymous LCN reviewers.

References

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*, 292–305.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*, 457-474.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The Crosslinguistic Study of Sentence Processing* (pp. 3-76). New York: Cambridge University Press.

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.36, retrieved 11 November 2017 from http://www.praat.org/

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. *Plos ONE, 5(6),* e10729.

Chen, B., Che, G., Chen, H., & Zhang, Z. (1988). *Hanyu liangci cidian* [A dictionary of Chinese numeral classifiers]. Fuzhou, China: Fujian Renmin Chubanshe [Fujian People's Publishing House].

Chou, C.-J., Huang, H.-W., Lee, C.-L., & Lee, C.-Y. (2014). Effects of semantic constraint and cloze probability on Chinese classifier-noun agreement. *Journal of Neurolinguistics, 31*, 42-54.

Cintrón-Valentín, M.C., & Ellis, N.C. (2016). Salience in second language acquisition: Physical form, learner attention, and instructional focus. *Frontiers in Psychology, 7,* 1284. doi: 10.3389/fpsyg.2016.01284.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27*, 3–42.

Clahsen, H., & Felser, C. (2018). Some notes on the Shallow Structure Hypothesis. *Studies in Second Language Acquisition, 40*, 693-706.

Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. *Studies in Second Language Acquisition, 33*, 589–624.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44*, 491-505.

Foucart, A., Garcia, X., Ayguasanosa, M., Thierry, G., Martin, C. D., & Costa, A. (2015). Does the speaker matter? Online processing of semantic and pragmatic information in L2 speech comprehension. *Neuropsychologia, 75*, 291–303.

Gao, M. Y., & Malt, B. C. (2009). Mental representation and cognitive consequences of Chinese individual classifiers. *Language and Cognitive Processes, 24*, 1124-1179.

Gass, S. M. (1986). An interactionist approach to L2 sentence interpretation. *Studies in Second Language Acquisition, 8*, 19-37.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research, 28,* 191–215.

Grüter, T., Lau, E., & Ling, W. (2018). L2 listeners rely on the semantics of classifiers to predict. In A. B. Bertolini & M. J. Kaplan (Eds.), *Proceedings of the 42nd Annual Boston University Conference on Language Development* (pp. 303–316). Somerville, MA: Cascadilla Press. (retrievable at http://www.lingref.com/bucld/42/BUCLD42-24.pdf)

Hernández, A. E., Bates, E. A., & Avila, L. X. (1994). Sentence interpretation in Spanish–English bilinguals: What does it mean to be in-between? *Applied Psycholinguistics*, *15*, 417–466.

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research, 29*, 33-56.

Huettig, F, Chen, J, Bowerman, M, & Majid, A. (2010). Do language-specific categories shape conceptual processing? Mandarin classifier distinctions influence eye gaze behavior, but only during linguistic processing. *Journal of Cognition and Culture, 10*, 39-58.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism, 4*, 257–282.

Klein, N. M., Carlson G. N., Li, R., Jaeger T. F., & Tanenhaus, M. K. (2012). Classifying and massifying incrementally in Chinese language comprehension. In D. Massam (Ed.), *Count and mass across languages* (pp. 261-282). Oxford, UK: Oxford University Press.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*, 32–59.

Kuznetsova, A, Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*, 1–26.

Lau, E., & Grüter, T. (2015). Real-time processing of classifier information by L2 speakers of Chinese. In E. Grillo & K. Jepson (Eds.), *Proceedings of the 39th Annual Boston University Conference on Language Development* (pp. 311–323). Somerville, MA: Cascadilla Press.

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science, 18*, 193-198.

Lew-Williams, C., & Fernald, A. (2009). Fluency in using morphosyntactic cues to establish reference: How do native and non-native speakers differ? In J. Chandlee, M. Franchini, S. Lord & G. Rheiner (Eds.), *Proceedings of the 33rd Annual Boston University Conference on Language Development* (pp. 290-301). Somerville, MA: Cascadilla Press.

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language, 63*, 447-464.

Li, S. (2010). Corrective feedback in perspective: The interface between feedback type, proficiency, the choice of target structure, and learners' individual differences in working memory and language analytic ability (Unpublished doctoral dissertation). Michigan State University.

Li, X.P. (2013). *Numeral Classifiers in Chinese: The Syntax-Semantics Interface.* Berlin: De Gruyter.

Liu, H., Bates, E. & Li, P. (1992). Sentence interpretation in bilingual speakers of English and Chinese. *Applied Psycholinguistics, 13*, 451–484.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception, & Psychophysics, 53*, 372–380.

Mitsugi, S. (2018). Generating predictions based on semantic categories in a second language: A case of numeral classifiers in Japanese. *International Review of Applied Linguistics.* (published online 2018-07-03)

Pan, H.-Y., & Felser, C. (2011). Referential context effects in L2 ambiguity resolution: Evidence from selfpaced reading. *Lingua, 121*, 221–236.

Pan, H.-Y., Schimke, S., & Felser, C. (2015). Referential context effects in non-native relative clause ambiguity resolution. *International Journal of Bilingualism, 19*, 298–313.

Paul, J. Z., & Grüter, T. (2016). Blocking effects in the learning of Chinese classifiers. *Language Learning, 66*, 972–999.

Polio, C. (1994). Non-native speakers' use of nominal classifiers in Mandarin Chinese. *Journal of the Chinese Language Teachers Association*, *29*, 51-66.

Qian, Z., & Garnsey, S. (2016). A sheet of coffee: an event-related potential study of the processing of classifier-noun sequences in English and Mandarin. *Language, Cognition, and Neuroscience, 31*, 761-784.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Siegelman, N, & Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language. *Journal of Memory and Language, 85*, 60-75.

Tsang, C., & Chambers, C. (2011). Appearances aren't everything: Shape classifiers and referential processing in Cantonese. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1065-1080.

VanPatten, B. (2004). Input processing in second language acquisition. In Bill VanPatten (Ed.), *Processing instruction: Theory, research, and commentary* (pp. 5-31). Mahwah, NJ: Erlbaum.

Zhang, J., & Lu, X. (2013). Variability in Chinese as a foreign language learners' development of the Chinese numeral classifier system. *The Modern Language Journal, 97*(s1), 46-60.

Table 1. Fixed effects from the linear mixed-effect model including all trials with correct mouseclick responses.

	Estimate	*SE*	*t*	*p*
(Intercept)	0.84	1.34	.63	.53
ConditionG-S-	2.46	1.89	1.30	.19
ConditionG+S+	-5.83	1.93	-3.02	.003 **
Group	-9.00	2.74	-3.29	.001 **
ConditionG-S- × Group	7.05	3.85	1.83	.07
ConditionG+S+ × Group	14.70	3.92	3.75	<.001 ***

Note: Formula: TargAdv ~ Condition * GrpCtr + (1 | Participant) + (1 | Item)

Table 2. Fixed effects from the linear mixed-effect model excluding trials with incorrect classifier selection in the Vocabulary test.

	Estimate	*SE*	*t*	*p*
(Intercept)	1.85	1.84	1.01	.32
ConditionG-S-	1.40	2.44	.57	.57
ConditionG+S+	-6.27	2.48	-2.53	.01 *
Group	-9.94	3.65	-2.72	.007 **
ConditionG-S- × Group	5.79	4.90	1.18	.24
ConditionG+S+ × Group	18.35	5.00	3.68	<.001 ***

Note: Formula: TargAdv ~ Condition * GrpCtr + (1 | Participant) + (1 | Item)

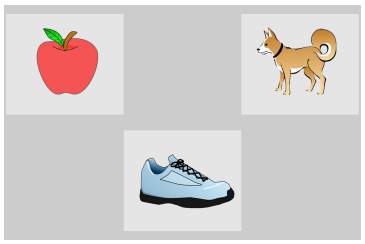
Figure captions

Figure 1. Examples of visual scenes for target [*gǒu*, ‘dog’] for each condition.

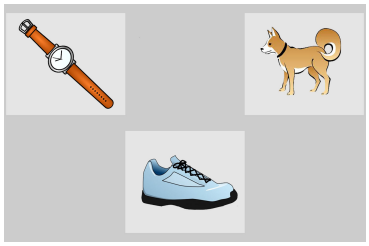
Figure 2. Difference in proportion fixations to target vs. competitor, by Condition and Group; including all trials with correct mouseclick response.

Figure 3. Difference in proportion fixations to target vs. competitor, by Condition and Group; excluding trials with non-target responses in the Vocabulary Test.

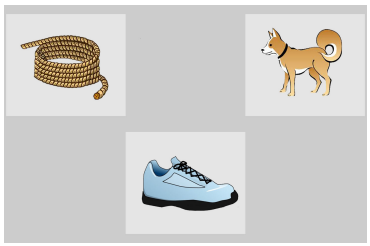
G-S-

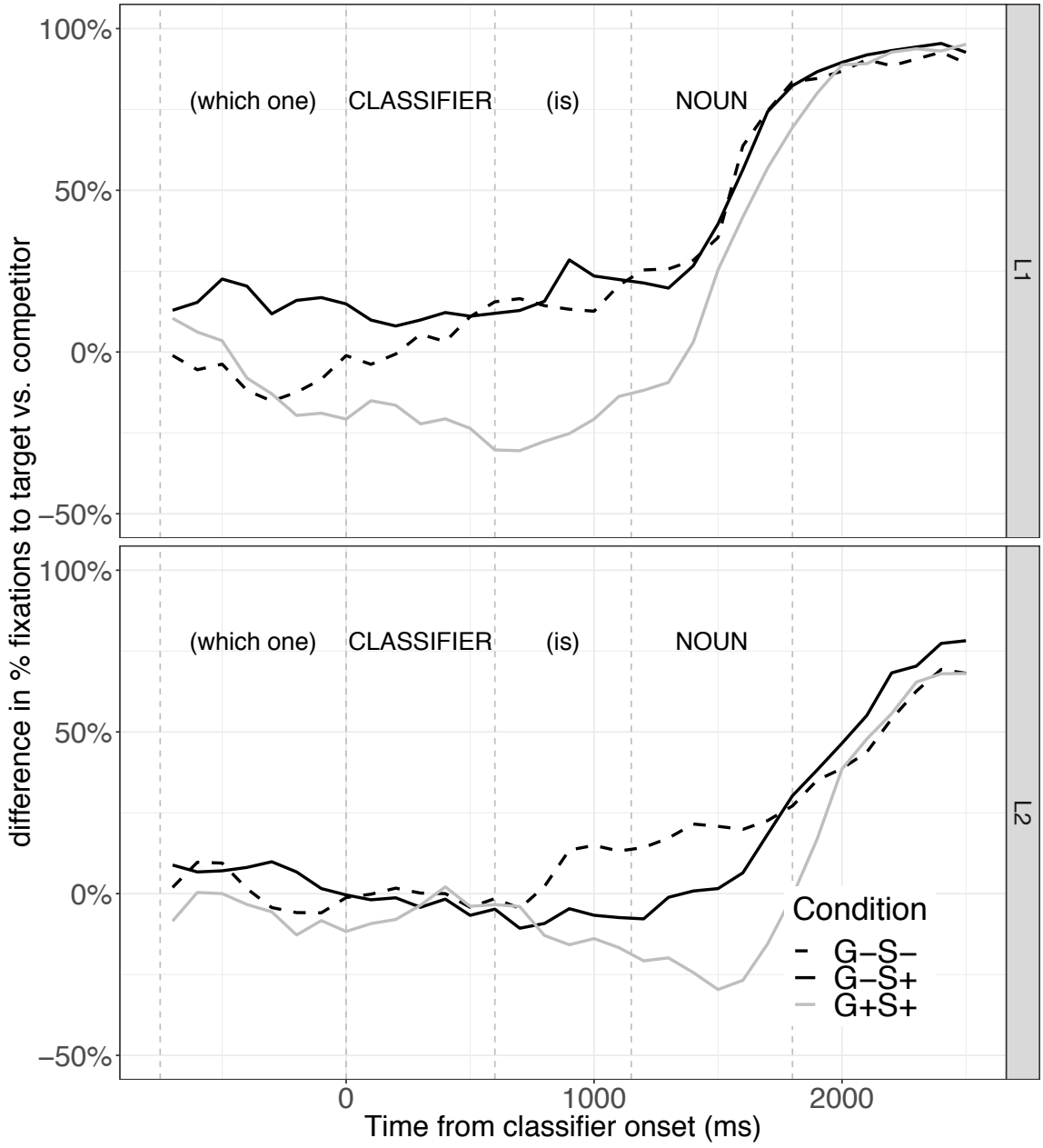


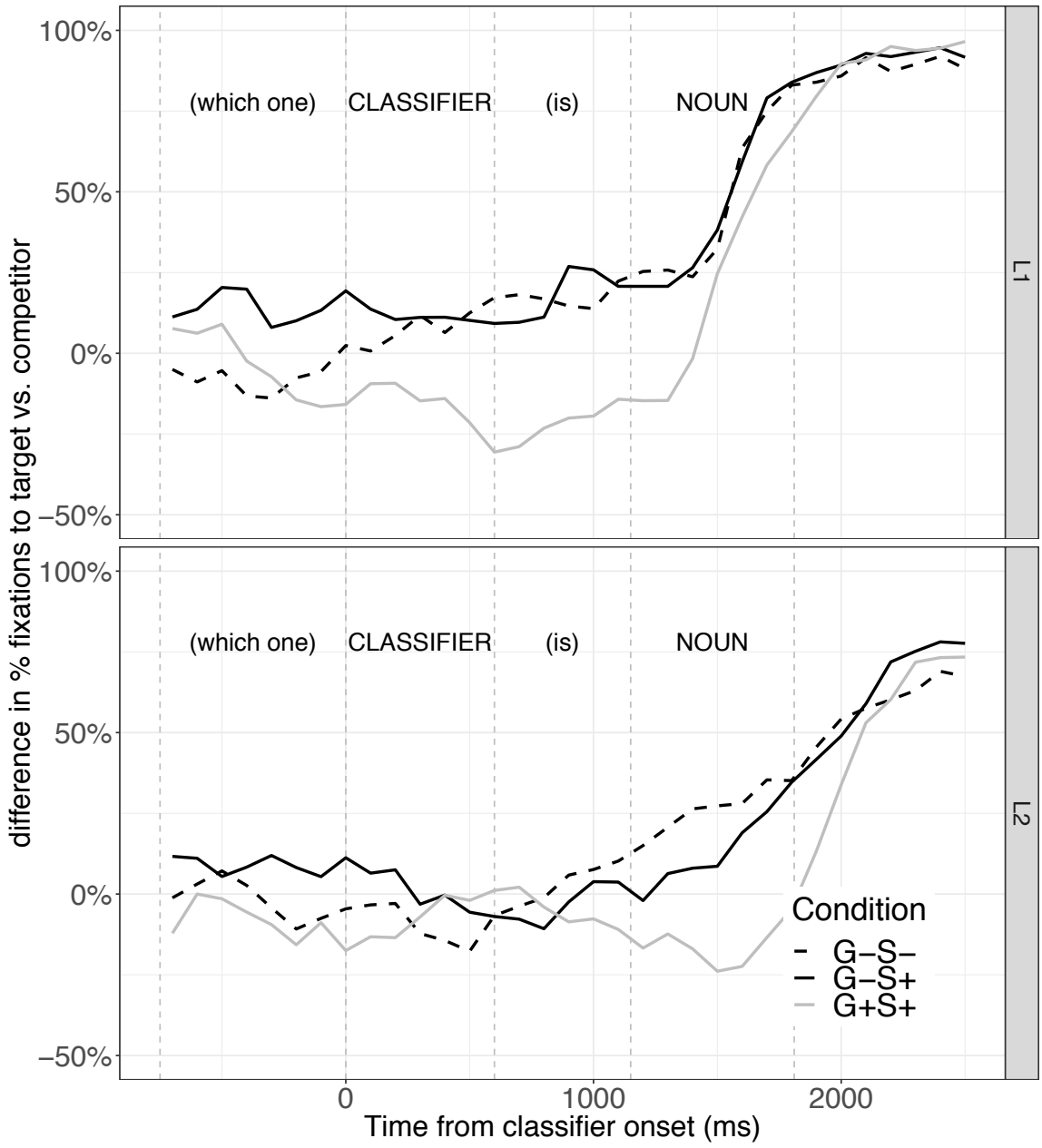
G-S+



G+S+




100%
50%
0%
−50%
(which one)
CLASSIFIER
(is)
NOUN
L1
L2
difference in % fixations to target vs. competitor
Condition
G−S−
G−S+
G+S+
0
1000
2000
Time from classifier onset (ms)


100%
50%
0%
−50%
(which one)
CLASSIFIER
(is)
NOUN
L1
L2
difference in % fixations to target vs. competitor
Condition
G−S−
G−S+
G+S+
0
1000
2000
Time from classifier onset (ms)

Endnotes

[i] A preliminary short report of this study appeared in the BUCLD proceedings (Grüter, Lau, and Ling, 2018) while data collection was still on-going. The small sample of L2 participants at that point, which also included some heritage speakers (now excluded), did not afford the statistical power needed for the planned analyses (reported here, Analysis 2), nor did this format allow for presentation of materials, including norming and rating procedures, sufficient for replicability, or full discussion of theoretical implications.

[ii] The in-house listening comprehension task consisted of an adaptation of the listening part of the Hanyu Shuiping Kaoshi (HSK or Chinese Standard Exam) Level Two sample test available on the Confucius Institute website (http://www.confuciusinstitute.manchester.ac.uk/hsk/hsk-learning-resources/past-papers-hsk-2/).

[iii] Note that unlike in T&C's Cantonese stimuli, the classifier in our Mandarin stimulus sentences did not immediately precede the noun, but preceded both the copula *shì* and the noun. This word order was selected to increase the critical time window between classifier and noun onset, considering the possibility that potential effects of prediction might arise more slowly in L2 listeners (Kaan, 2014).

[iv] T&C's experiment included visual scenes with four objects: a target, a competitor, and two distractors. In order to maximise potential looks to competitors, visual scenes in the present study included only one distractor, i.e., three objects in total. If the Mandarin translation of both Cantonese distractor items fit our criteria, Distractor 1 was selected.

[v] Two trials with no recorded fixation data during the critical time window were removed.

[vi] For example, if on a given trial, of the 58 bins comprised by this window, 20 contained looks to the target and 15 contained looks to the competitor, the TargetAdvantage score for this trial would equal 5. This dependent measure was chosen over weighted empirical logits (Barr, 2008) because it was more normally distributed.

[vii] This led to the removal of 37 trials (L1: 6/283, L2: 31/443) in which an early initiated fixation continued throughout the critical time window.

[viii] All models reported contain random intercepts only. Models with random slopes were attempted, but did not consistently converge. For the sake of comparison, we therefore modelled all data with random intercepts only. When fuller models converged, anova() comparisons indicated that they were not a significantly better fit for the data. The patterns of results did not change in fuller models.