

Schooling and language usage matter in heritage bilingual processing: Sortal classifiers in Mandarin

Second Language Research

1–26

© The Author(s) 2024



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/02676583241270900

journals.sagepub.com/home/slr**Jiuzhou Hao** 

UiT The Arctic University of Norway, Norway

Maki Kubota

UiT The Arctic University of Norway, Norway; University of Bergen, Norway

Fatih Bayram 

UiT The Arctic University of Norway, Norway

Jorge González Alonso 

UiT The Arctic University of Norway, Norway; Nebrija University, Spain

Theres Grüter 

University of Hawai'i at Mānoa, USA

Muhan Li

UiT The Arctic University of Norway, Norway

Jason Rothman

Lancaster University, UK; UiT The Arctic University of Norway, Norway; Nebrija University, Spain

Abstract

Mandarin sortal classifiers simultaneously encode semantic and grammatical form class cues. Building on a second language (L2) study of Grüter et al. we used the same visual world eye-tracking experiment, designed to examine the relative use of the two cues, testing Mandarin

Corresponding author:

Jiuzhou Hao, Department of Language and Culture, UiT The Arctic University of Norway, Hansine Hansens veg 18, Tromsø, 9019, Norway

Email: jiuzhou.hao@uit.no

heritage speakers (HSs) living in an English-speaking environment. Given the importance of understanding individual differences, the present study also examined if/how individual HSs may systematically differ as a function of experience with Mandarin. As a group, HSs – like the first language (L1) group in Grüter et al.'s study – showed a clear reliance on the grammatical form class cue. Nevertheless, individual HSs with Mandarin schooling and with more Mandarin social exposure/use showed more reliance on semantics, the L2 pattern observed in Grüter et al. We discuss why this latter pattern might have obtained (formal and informal literacy), while highlighting the value of individual difference approaches to understanding HS processing.

Keywords

classifier, eye-tracking, heritage language processing, individual differences, Mandarin Chinese

I Introduction

A heritage language (HL) is a minority language acquired by a subset of people in contexts where a different language dominates as the common, default societal one (Fishman, 2014; Montrul, 2008, 2016; Rothman, 2009; Valdés, 2000). This means that HLs are often, yet not exclusively, associated with immigrant contexts. Heritage speakers (HSs) are early bilinguals who acquire their HL naturalistically from birth. Although HSs typically become dominant in the societal majority language by adolescence, it is important to keep in mind that they are nevertheless native, first language (L1) users of the HL (Rothman and Treffers-Daller, 2014). HSs can be simultaneous or sequential bilinguals, the former typically occurring when only one parent is a speaker of the HL and/or both parents are themselves (second generation) bilinguals of the HL, dominant in the societal majority language.

At the macro-group level, HSs' linguistic competence is typically characterized by considerable differences from homeland L1-dominant users although each is an early childhood acquirer of the same language (Dominguez et al., 2019; Montrul, 2008, 2016; Polinsky, 2008). At the same time, HSs may perform differentially compared to late second language (L2) learners of the same language, even when matched for relative proficiency and despite both being dominant bilinguals in the same majority language (Montrul et al., 2013; Montrul and Potowski, 2007). The observed differences between HSs and homeland L1-dominant users and/or L2 learners (L2ers) are, however, unsurprising given that their contexts for acquiring the language differ dramatically at crucial points in development across several dimensions, e.g. exposure to, education in and domains and opportunities for using the HL. As telling as it can be, focusing (solely) on such comparative approaches reduces group-level differences and/or similarities to one factor or a highly reduced set of factors (Bayram et al., 2019; Kupisch and Rothman, 2018; Pascual y Cabo and Rothman, 2012), e.g. timing of acquisition, input quantity, etc. Moreover, doing so also fails to acknowledge at all, or chooses to place minimal importance on, the considerable variability among HSs, both within and between studies, as evidenced in reported measures of dispersion and participant-level random effects in multilevel modelling. Unpacking what factors give rise to individual differences among HSs offers a window into the underlying causes for why each individual HS performs the

way they do. Under an individual differences approach, the central question is, then, not what factors make HSs perform more or less like L1-dominant users and/or L2ers but rather what factors help explain variability among HSs themselves.

Additionally, the vast majority of such comparative work has adopted offline comprehension and/or production tasks. These tasks are often facilitated or at least affected performatively by various factors such as degree of metalinguistic knowledge and affective factors. This could result in under-/over-estimation of linguistic representation in general. Meanwhile, online processing methods measure how participants respond to linguistic stimuli in real time. While such methods can also be subject to various performance-level variables, they reduce the (potential for) application of metalinguistic knowledge and affective filters. Insofar as these factors are especially challenging for HL empiricism (Polinsky, 2018), methods that reduce them are especially welcome. In fact, recent studies focusing on comparisons between HSs vs. L1-dominant users and/or L2ers have observed task effects at the group-level (e.g. Hao et al., 2024; Regulez and Montrul, 2023). For example, online processing studies show that HSs often adopt the same (qualitatively) online processing strategies used by homeland speakers, a pattern that would not be anticipated given preexisting work showing HSs display differential offline comprehension accuracy and production patterns (Di Pisa et al., 2022; Fuchs, 2022a, 2022b; Hao et al., 2024; Jegerski, 2018a, 2018b; Jegerski et al., 2016; Luque et al., 2023). Perhaps such discrepancies are unsurprising since offline and online methods only partially overlap in terms of what they (are designed to) tap into. Our point is not to suggest that one is superior to the other. Rather we highlight that there continues to be a dearth in online HL processing studies and that discrepancies between the methods as highlighted in recent work bring to the surface that both are needed to understand our object of study more completely. Moreover, in recent years psycholinguistic work regressing sociolinguistic variables that proxy for various types of HL exposure/engagement has been emerging, offering promising insights into the magnitude of HS individual differences and, indeed, what underlies them (Bayram et al., 2019; Daskalaki et al., 2022; Flores et al., 2017; Hao and Chondrogianni, 2023; Hao et al., 2024; Lloyd-Smith et al., 2020; Soto-Corominas et al., 2022).

Given discrepancies between HSs' online and offline comprehension (and production) and, following recent calls to move away from monolingual comparative normativity as the baseline/benchmark for (HS) bilingual performance (De Houwer, 2023; Ortega, 2020; Paradis, 2023; Rothman et al., 2023), processing-based HL studies that sidestep or de-emphasize comparisons to homeland L1-dominant users and/or L2ers are needed. The present study seeks to understand which and how different bilingual language experience factors modulate HL processing using the visual-world eye-tracking paradigm. Specifically, we examine the processing of sortal classifiers in adult HSs who speak Mandarin as the HL and live in a society where English is the dominant language. Although the main goal of the present study is to take up an intra-HS-group individual differences approach, we will make comparative reference to pre-existing data with homeland dominant users and L2ers from Grüter et al. (2020), the study whose methodology we replicate herein.

Mandarin sortal classifiers are free-standing morphemes that categorize nouns they co-occur with into classes according to the inherent properties of the object, e.g. shape,

natural kind, and function (Erbaugh, 2012). And, yet, there are (exceptional) cases where the classifier–noun pairing does not accord with the prototypical semantic category association, for example, when the semantics of the classifier is *thin, long, often flexile* (*tiao* classifier), but the classifier co-occurs with something that is not, such as *dog*. Such cases are instances of item-specific grammatical form class mappings (represented in the mental lexicon) between nouns and classifiers (for more details, see Section II). In other words, Mandarin sortal classifiers encode both semantic and grammatical form class cues to the co-occurring nouns. Importantly, although (some) Mandarin classifiers are produced from the age of 3 years – suggesting the syntactic structure is in place early – typically-developing homeland L1 children do not fully acquire the breadth of the complex distribution of classifiers until around age 10 years (Erbaugh, 2012). Such a developmental pattern suggests the need for significant amounts of input over time and/or a role for schooling/literacy in fully developing the classifier system. As such, the processing of sortal classifiers – both in terms of if and how different cues are used during online processing – stands out as a property for which we might expect variability across HS participants, predictable (i.e. regressed statistically) based on differences in how individuals have come to acquire and use their HL.

II Mandarin sortal classifiers: Properties, acquisition and processing

I Properties of Mandarin sortal classifiers

Within the morphosyntax of the nominal system, languages can differ significantly. For example, many of the world’s languages have overt inflection (and different degrees of agreement across the NP/DP) for gender and number. Languages can also differ in terms of whether they have nominal classifier systems. While Mandarin, like many Asian languages, does not have grammatical gender (assignment/agreement) or even obligatory number inflection, it does have a rather robust classifier system. For example, as in (1) and (2) below, to express ‘one rope’ and ‘three ropes’, Mandarin, unlike English, requires a classifier – *tiao* – between the numeral *yi* ‘one’ and *san* ‘three’ and the noun *shengzi* ‘rope’. Crucially, such classifiers do not provide any number information (as can be seen, whether there are one or three ‘rope’, the classifier and the noun are the same). Rather, they offer categorizing information related to the inherent properties of the noun, e.g. shape, natural kind and/or function. Grammatically, they are obligatory if the noun is modified by a numeral or a demonstrative, as shown by the ungrammaticality (*) when the classifier is omitted in (1) and (2). This type of classifier that divides the inventory of count nouns into semantic classes has been termed as sortal classifier (Erbaugh, 2012), the focus of the current study.

- | | | | | |
|-----|-----|---|--------------------------|--------------|
| (1) | — | * | (条) | 绳子 |
| | yi | * | (tiao) | shengzi |
| | one | | CL _{LONG-THING} | rope |
| | | | | ‘a/one rope’ |

- (2) 三 * (条) 绳子
 san * (tiao) shengzi
 three CL_{LONG-THING} rope
 ‘three ropes’

The mapping between sortal classifiers and nouns based on semantic features alone (object categorization) is (1) gradient and (2) not always guaranteed. To start, classifier–noun goodness of fit is determined by degree of exemplar prototypicality within a categorical hierarchy (Gao and Malt, 2009). That is, not every member of a category is as prototypical as others, whereby members of the same category lie on a spectrum. In the case of the classifier *tiao*, for example, while both *shengzi* ‘rope’ and *gou* ‘dog’ co-occur with the *tiao* classifier, *shengzi* is a prototypical member for the *tiao* classifier – slender, long, often flexible – whereas *gou* is simply an exception (i.e. most dogs are not long, slender, and flexible). As a result, the grammatical form class of *gou* – its idiosyncratic classifier–noun entry in the mental lexicon – is in a sense (semantically) irregular. Thus, like irregular past forms in English and grammatical gender of inanimate nouns in many gender-marking languages, for example, the mapping must be learned as a grammatical property of an individual lexical item. Moreover, having the attributes associated with the semantic features of a given classifier does not entail that all such nouns take it. For example, although *shoubiao* ‘wristwatch’ is an object that one might describe as ‘slender, long and flexible’, it does not co-occur with the *tiao* classifier. Rather it takes *kuai*, a classifier semantically associated ‘a lump-shape thing’. While a wristwatch is also a lump-shape object – so unlike dog with *tiao*, it is not a semantic exception for *kuai* per se – the fact that it cannot collocate with *tiao* highlights how semantics alone cannot be the whole story. Semantic information, while extremely important for acquisition and processing, cannot be the only criterion underlying the Mandarin classifier system. While it is true that for many, potentially most, classifier–noun pairs the grammatical form class and the semantic information encoded by the classifier itself fully align with each other, the rather non-trivial existence of the above scenarios suggests that the abstract grammatical property of the form class (within the mental lexicon) plays a non-trivial role, at least for the ultimately attained systems of L1-dominant users.

2 The acquisition and processing of Mandarin sortal classifiers

The sortal classifier system is a late acquired property for monolingual children and (even more for) child HSs (e.g. Erbaugh, 2012, for monolingual children; Jia and Paradis, 2015, for child HSs). Although monolingual children produce classifiers following the grammatical constraints early on, i.e. classifiers are obligatory after numeral/demonstrative, and do not co-occur with one another (double classifiers), child HSs produce ungrammatical classifier–noun pairs for a longer period of time. In addition, child HSs tend to overproduce the general classifier when monolingual children prefer specific sortal classifiers (for similar results in Cantonese–English child HSs, see Kan, 2019; Li and Lee, 2001).

In terms of the processing of sortal classifiers, Huettig et al. (2010) tested Mandarin-speaking L1-dominant adults in a visual world eye-tracking study. The results showed that linguistic stimuli that had a classifier compatible with the target noun in the visual scene facilitated processing (look at the target faster) relative to linguistic stimuli without a classifier. However, as the study included semantically-matched classifier–noun pairs only, it does not permit one to examine the potentially differential use of the grammatical form class information and/or semantic information in classifiers.

The study by Li et al. (2021), on the other hand, included four conditions in an electroencephalography (EEG) study with adult Mandarin-speaking L1-dominant speakers and Mandarin HSs living in Malaysia (Malay as the societal dominant language) to examine the processing of classifier–noun pairs at the phrasal level, i.e. isolated classifier–noun pairs. The important contrast, for the purpose of this study, is between classifiers and nouns that co-occur with each other and the ones that do not. Overall, the authors found that the HSs used classifiers to predict and integrate the upcoming nouns to the same extent as the L1-dominant speakers. Specifically, a larger N400 amplitude was observed in implausible classifier–noun pairs (classifier–noun mismatch) than in plausible classifier–noun pairs (classifier–noun match). Crucially, the magnitude of the N400 effect was the same across groups, suggesting that the two groups performed similarly in terms of how they process classifiers and their sensitivity to the classifier–noun mismatches.

Grüter et al. (2020) followed Experiment 3 from Tsang and Chambers (2011), where Cantonese L1 speakers were tested by explicitly manipulating the matching between classifiers and nouns in both grammatical form class and semantic properties. Also using the eye-tracking visual world paradigm, Grüter et al. (2020) tested Mandarin L1-dominant speakers and (mostly) English-speaking adults who learned Mandarin as a second language (L2ers). Herein, we have adopted the exact same experiment(s) and terminology from Grüter et al. (2020) upon which we will now elaborate.

In a visual scene, an object, e.g. *gou* ‘dog’, that co-occurs with a specific classifier, e.g. *tiao*, but does not prototypically belong to the semantic class the classifier is associated with (*tiao* being semantically associated with long, thin, often flexible objects), served as the target along with a competitor (of various types as explained below) and a wholly irrelevant distractor in the same visual scene. Following the terminology used by Tsang and Chambers (2011) and Grüter et al., the target object in experimental trials could always be characterized as G+S–, meaning that it matched the classifier in the spoken sentence in grammatical form class (G+), but did not provide a (prototypical) match in semantic features for that classifier class (S–). Meanwhile, objects that can be characterized as G–S–, e.g. *pingguo* ‘apple’, G–S+, e.g. *shoubiao* ‘wristwatch’, and G+S+, e.g. *shengzi* ‘rope’, served as competitors in three different conditions, respectively.

While looking at the visual scenes, participants heard sentences in the format of ‘Which one + classifier + is + target noun (G+S–)’, as in (3). The rationale was that if the processing of classifiers is driven primarily by grammatical form class, objects/nouns that do not co-occur with the classifiers (G–) should induce similar looking patterns irrespective of whether the nouns share the semantic features associated with the classifier (S+) or not (S–). Conversely, if the processing of classifiers is primarily driven by semantic features, objects/nouns that have semantic features associated with that classifier (S+) should receive more looks relative to the ones that do not (S–). Grüter et al.’s results showed that for both L1-dominant speakers and L2ers, the G+S+ condition

induced more looks to the competitor than the G–S+ condition. However, the G–S+ condition induced more competition than the G–S– condition only for the L2ers. The authors interpreted the latter result as L2ers’ heavier reliance on semantic information.

- (3) 哪一 条 是 狗?
 Na-yi tiao shi gou?
 Which one CL is dog?
 ‘Which one is dog?’

It is worth noting here that the underlying nature of processing Mandarin classifiers in the L1-dominant speakers remains an open question, i.e. if it is primarily guided by the semantic property, the grammatical form class or both. In Li et al.’s (2021) EEG-based study, classifier–noun mismatches induced larger N400 effects, an event-related-potential signature reasonably associated with the processing of semantic integration (see also Kwon et al., 2017 in Mandarin L1-dominant users). On the other hand, the lack of performance difference between the G–S– condition and the G–S+ condition among the L1-dominant speakers found in Grüter et al. (2020) seems to suggest that they were not, or to a lesser extent than the L2ers, recruiting the semantic information encoded in the classifiers (see also Tsang and Chambers, 2011 in Cantonese). However, the question of whether the L1-dominant speakers employed the semantic information at all cannot be directly answered by the eye-tracking study given its design. Specifically, because the competitor nouns in both the G–S– condition and the G–S+ condition cannot co-occur with the classifiers given in the experiment, the difference between the two does not correspond to any difference observed between classifier–noun match vs. mismatch conditions in other studies. For now, and for the purpose of the current study, it suffices to note that performance difference between the G–S– condition and the G–S+ condition indexes how likely it is that one still engages in the use of semantic information when the classifier and noun do not co-occur with each other.

A few studies targeting production of classifiers in Mandarin HL have been undertaken and show that several language background factors modulate HSs’ performance. Specifically, more HL exposure and use (Kan, 2019), later Age onset of Acquisition (AoA) of the majority language (Jia and Paradis, 2015; Kan, 2019), and having HL schooling (Saturday schools; Jia and Paradis, 2015) have separately predicted production performance in HSs. Adopting a confirmatory (statistical) approach, the current study builds on these previous findings, targeting these specific language background factors to examine their predictive value in the domain of comprehension and, crucially, online processing of Mandarin classifiers. We explore this by using the main visual world experiment from Grüter et al. (2020).

III The present study

The present study adopts the methods used by Grüter et al. (2020), as described above, to examine the processing of Mandarin sortal classifiers in Mandarin adult HSs living in a society where English is the dominant language. We sought to address and answer two research questions:

- Research question 1: At the group level, what are the performance patterns of Mandarin–English HSs on an eye-tracking task examining Mandarin sortal classifiers? More specifically, what is the status of classifiers as grammatical and semantic cues in HSs’ mental grammars and sentence processing?
- Research question 2: How do individual level bilingual language experience factors modulate HSs’ performances?

Although no previous study, to our knowledge, has examined classifier processing in HL Mandarin at the sentence level (for the phrasal level, see Li et al., 2021), in the context of research question 1, three reasonable data outcomes exist. It is possible that our HS group processes classifiers primarily as a grammatical form class cue, which would be similar to what Grüter et al. (2020) found for L1 Mandarin-dominant users. Alternatively, it is possible that our HSs will show evidence of primarily semantically driven processing, more akin to the L2ers in Grüter et al. (2020). Finally, it is possible that the HSs show processing patterns that cannot be clearly identified as being primarily grammatical form class- or semantics-driven. Our prediction is that the group results will lean more towards the first possibility (in line with the findings of Li et al., 2021).

The present study is in fact primarily interested in what the data reveal beyond the aggregate trend. Indeed, we anticipate that the group data will obscure important individual differences that we unpack in addressing research question 2. For research question 2, two possibilities exist, i.e. (1) individual HSs simply do not differ from each other in a systematic way or (2) they do. If individual HSs do not differ from one another, it should follow that none of the language background factors we regress matter for the processing of sortal classifiers, i.e. all participants either do or do not process Mandarin sortal classifiers in the same way irrespective of differences they have with exposure to and engagement with Mandarin. Alternatively, individual HSs could differ in processing patterns as a function of their experience/engagement with Mandarin (and English). In particular, it would be reasonable to expect that increased exposure to and usage of Mandarin as well as formal training in Mandarin (literacy) may matter. Our prediction is that different language background factors will modulate individual HSs’ processing of classifiers. This is supported by previous behavioural research on Chinese classifier production in HL contexts, which identifies HL exposure and use (Kan, 2019), AoA of English (Jia and Paradis, 2015; Kan, 2019), and HL schooling (Jia and Paradis, 2015) as modulating factors.

IV Methodology

I Participants

A total of 60 Mandarin–English HSs took part in the study online. Data from three participants were excluded from further analysis due to data loss and data from one participant were excluded due to poor eye-tracking quality. The final sample consists of 56 participants (27 female, mean age: 25.21 years, SD: 5.46). Within the 56 participants, 31 resided in Singapore, 13 were in the UK, while the other 12 were in the US. All HSs were exposed to Mandarin from birth at home and to English before the age of 5 years: our

participants were either born and raised ($n=44$) or immigrated to their current location before the age of 5 ($n=11$). Additionally, 18 HSs report knowing Cantonese, Hokkien, and/or Teochew (all classifier languages). Critically, the social reality for all HSs in the current study is that English is the sole language of their primary education and the dominant and preferred language of the larger society. Thus, by any defined qualification of heritage speaker-ness, all participants in this study unambiguously qualify. We acknowledge, however, that it is possible that location, i.e. Singapore vs. the UK vs. the US, could contribute more or less variability at the aggregate level. For example, the ranges for individuals' language usage may be significantly larger or smaller by location. To make sure this is not the case in our sample – not ignoring the potential that in a different sampling it very well could be – we ran statistical analyses to quantify this (see Section IV.3 and Section V).

2 Materials and procedure

The visual world experiment and the vocabulary test from Grüter et al. (2020) were included as the main experimental tasks, along with an independently-developed elicited production task. The results of the production task are not included in the present analyses. The in-house listening proficiency task Grüter et al. adapted from the listening part of the *Hanyu Shuiping Kaoshi* (HSK) Level 2 was also included to measure participants' Mandarin proficiency. We also administered an adapted version (details below) of the Language and Social Background Questionnaire (LSBQ; Anderson et al., 2018) to all participants to collect their (language) background information.

All participants took part in the study at their homes. We implemented all tasks with Gorilla on a webpage. Gorilla Experiment Platform is a Graphical User Interface (GUI)-based experiment builder software (Anwyl-Irvine et al., 2020), which utilizes WebGazer.js (Papoutsaki et al., 2016) to run webcam-based eye-tracking. Each participant participated in all the tasks in two separate sessions and the entire experiment lasted approximately 60 minutes. The first session included the LSBQ, the vocabulary test, and the elicited production task while the second consisted of the main Visual World experiment and the proficiency task.

To minimize any carry over effect from the vocabulary test to the Visual World experiment as the target nouns were the same, all participants were only given the link to the second session of the experiment at least 7 days after having completed the first session. The study was approved by the institutional ethics committee. In accordance, all participants were informed of their ethical rights of participation in written form, prior to the experiment. Before any tasks, participants were asked to check boxes on the webpage to give consent for their participation.

3 Language background

The LSBQ is a comprehensive, validated questionnaire that measures bilingual proficiency, language usage in different contexts inclusive of code-switching behaviours. For measures of Mandarin usage patterns, the LSBQ calculator provides two relevant factor scores, i.e. Mandarin Home Use (MHU) and Proficiency and Mandarin Social Use

Table 1. Summary of the language background factors across Location (means, standard deviations, minimum and maximum values).

	Singapore (<i>n</i> =31)	UK (<i>n</i> =13)	USA (<i>n</i> =12)	Test statistics
MHU	8.68 (3.70; 1.94–15.24)	10.40 (6.78; 0.61–19.18)	10.39 (6.50; 2.43–20.25)	$H(2)=0.64, p=.72$
MSU	12.47 (8.82; −0.59–35.71)	10.73 (6.24; 0.71–24.54)	13.07 (11.17; −2.22–31.88)	$H(2)=0.26, p=.87$
AoA E	0.77 (1.68; 0–5)	1.84 (1.81; 0–5)	1.16 (1.80; 0–5)	$H(2)=5.22, p=.07$
Programme	$N_{Yes}=6$	$N_{Yes}=5$	$N_{Yes}=2$	$p=.39$

Notes. MHU=Mandarin Home Use; MSU=Mandarin Social Use; AoA E=Age onset of Acquisition of English (in years), Program=Language Programme.

(MSU). The factor scores are calculated by multiplying the standard score ((Observed Score-Mean)/ Standard Deviation) by the variable's weight and then summing all the variables that load onto specific factors. For the specific questions/standard scores loaded onto these factor scores, we refer the readers to the LSBQ calculator or the LSBQ manual. We directly used the factor score for MSU. However, we adapted the factor score MHU and Proficiency by excluding standard scores related to proficiency to proxy for MHU more directly. For both scores, a higher value indexes more exposure and use of Mandarin in home contexts (MHU) and in social contexts (MSU). It is worth noting here that the two scores are not directly comparable to each other; given the way these scores are calculated, the two scores have different minimum and maximum values. Apart from the two factor scores, we also collected information about the participants' AoA of English (AoA E) and experience with formal Mandarin schooling by asking at what age the participant learned English and if they had attended any Saturday Schools or similar courses in Mandarin. AoA E was operationalized in years and formal Mandarin schooling was coded as a binary factor, i.e. Yes (1) or No (0). To test if these factors differ across locations, we adopted the Kruskal–Wallis test for numeric factors and the Fisher's exact test for categorical variables. These tests were used because our numerical data are not normally distributed and/or do not have equal variances between groups and because our categorical data have a small contingency table. These non-parametric tests are more robust in these situations (Nahm, 2016). Table 1 summarizes the distribution of these factors as Means (SDs; ranges) and presents the test results showing that these factors do not differ across locations.

4 Visual world experiment

Participants were shown visual scenes with colour clipart images of three objects (Figure 1): a target, a competitor, and a distractor. The experimental trials included 12 targets paired with three target sortal classifiers, i.e. *tiao* '∼long, flexible'; *zhi* '∼stick-like, long'; and *zhang* '∼flat, spread open'. The 12 target nouns were selected such that although they should co-occur with only one of the three classifiers (grammatical form class; G+), they do not have the prototypical semantic features typically associated with that classifier (class; S−). For example, the noun, *gou* 'dog', co-occurs with the classifier, *tiao*, despite the classifier



Figure 1. Examples of visual scenes for target noun *gou* ‘dog’ and classifier *tiao* for each condition.

being generally associated with objects that are long, slender, and flexible: features not straightforwardly applying to ‘dog’. Three types of competitor objects were included, i.e. three conditions. The G+S+ condition included competitor objects that matched with the target classifier in both grammatical form class and semantic features, e.g. *shengzi* ‘rope’ for the *tiao* classifier when the target was *gou* ‘dog’. The G–S+ condition included competitor objects that matched with the target classifier in semantic features but not in grammatical form class, e.g. *shoubiao* ‘wristwatch’. The G–S– condition included competitor objects that neither matched with the target classifier in grammatical form class nor semantic features, e.g. *pingguo* ‘apple’ for the *tiao* classifier. As for the distractors, they were selected by the same criteria as G–S–competitors (see Figure 1).

To avoid repetition, items were counterbalanced across three lists such that each target item appeared only once, in one of the three conditions (G+S+, G–S+, G–S–), resulting in four items per condition (as in Tsang and Chambers, 2011). The order of items was pseudo-randomized and interspersed with 24 fillers. There were three types of fillers, consisting of items where (1) the two unmentioned objects were of the same class, (2) objects differed by colour only and (3) objects shared perceptual or semantic properties but were not from the same grammatical form class.

Given that we used the same tasks as in Grüter et al. (2020), it is worth mentioning that they conducted an independent rating study and a corpus analysis on their experimental materials. They found that:

- in terms of semantic compatibility, S+ objects (i.e. G+S+ and G–S+ competitors) were rated to be significantly more aligned with their classifier’s associated semantic features, compared to S– objects (i.e. G+S– targets, G–S– competitors), and distractors; and
- the three competitors did not differ from each other in their overall frequency.

Additionally, all target, competitor and distractor nouns were compatible with only one of the classifiers used in the experiment and were likely to be familiar to 3rd-year learners of Mandarin (based on inspection of textbook vocabulary). Accompanying audio stimuli for the visual scenes were all in the format of *nayi* ‘which one’ + classifier + *shi* ‘is’ + target noun ‘Which one is TARGET’. The duration of the parts before the

target noun onset was held constant across all items. Silence was added after the classifier and after *shi* ‘is’ such that the duration from classifier onset to noun onset was exactly 1,150ms in each experimental item, constituting the critical region for analysis. For a link to the experimental materials including the images, audio recording, etc., used in the study and Grüter et al. (2020), see the data availability statement below.

5 Vocabulary test

The vocabulary test included the 12 classifier and target noun pairs used in the experimental trials in the Visual World experiment, together with 38 filler trials. The main purpose of the vocabulary test was to examine whether participants have the item-specific classifier–noun pairing representations of the target classifiers to the target nouns. This is to ensure that participants know that target nouns co-occur with the target classifiers despite the fact that they do not match in semantic categorization and to account for potential variability in idiosyncratic association between classifiers and nouns. As such, items in which participants did not select the target classifier–noun pairing in the vocabulary test were removed from analyses in the Visual World experiment.

Different from Grüter et al. (2020), the task was adjusted to a listening experiment to account for the fact that not all HSs would have experience with reading/writing in Mandarin. In the task, participants heard, in one trial (for examples, see 4 and 5), four items (phrases) that differed only in one morpheme and were asked to choose the item that sounded the most natural among the four. More specifically, the morphemes that differed across items within the same trial were classifiers for the classifier–noun trials (4) and lexical or functional morphemes for the filler trials (5). Another difference between the task in the current study and that of Grüter et al. (2020) was that there was no English translation to the phrases provided. In the examples, the underlined morphemes correspond to the morpheme choices given in Grüter et al. (2020), and the bolded items are considered the optimal choices.

- (4) a. **yi tiao yu;**
 b. yi zhang yu;
 c. yi zhi yu;
 d. yi tai yu
 ‘a/one fish’
- (5) a. yao mei shui;
 b. **hai mei shui;**
 c. neng mei shui;
 d. suo mei shui
 ‘still hasn’t slept’

V Results

For both mouse-click accuracy and eye-gaze data, generalized linear mixed effect regressions with logistic link function (GLMERs) were carried out with the *lme4* package

(Bates et al., 2015) in R (R Core Team, 2018). We included the maximal random effects justified by the design where possible (Barr et al., 2013), i.e. by-participant and by-items random intercepts, as well as by-participant random slopes for Condition. When the maximal model failed to converge, we tried different optimizers first where possible, using the *afex* package (Singmann et al., 2020), and then iteratively simplified random effect structures until convergence was achieved, i.e. removing random effect(s) accounting for the least variance.

For model selection, both confirmatory and exploratory approaches were used, albeit for different reasons. Specifically, our statistical modelling, whenever possible, follows a confirmatory approach that is subjective and theory-driven (McElreath, 2020; Winter, 2019). Under such an approach, we included all predictors of interest in the models. Typically, these predictors include Condition (*treatment coded*; **G-S-**, **G-S+** and **G+S+**; note that unless stated otherwise, the bolded level is the reference level entered in the model), Mandarin Home Use (MHU; *centred*), Mandarin Social Use (MSU; *centred*), AoA of English (AoA E; *centred*), and Language Programme (*treatment coded*; **Yes** and **No**). Given the quantity of data at our disposal, we did not include interaction terms in the models among these background factors. Interactions between Condition and individual language background factors were included. A forward stepwise selection approach using maximal likelihood ratio tests was adopted to examine if Location (*treatment coded*; **Singapore** vs. **UK** vs. **US**) would matter for participants' performance, even though participants from the three locations did not differ in language background factors.

For post hoc analyses, Bonferroni pairwise comparisons were conducted if the optimal models included significant interaction terms, along with models with all possible combinations of reference levels for all variables. It is worth noting that, as all categorical variables are treatment coded, all effects reported in the statistical tables reflect simple effects instead of main effects, i.e. the effect of a variable in one level relative to the baseline/reference level. Additionally, when summarizing statistical outcomes for specific comparisons (between levels of one variable), estimate, SE, *z*, and *p* statistics will not be repeated if they are recoverable/apparent in the model output summarized in the relevant table(s); model statistics only available from post hoc analyses will be spelled out explicitly in text. We refer the readers to the R Scripts at <https://osf.io/qcxnh> for more information.

1 Vocabulary test

Two participants had below chance accuracy and were excluded from further analyses. After the exclusion, across all 50 items, i.e. including fillers ($n=38$), the mean accuracy was 0.84 (range=0.44–0.98, SD=0.09). For the target 12 classifiers the mean accuracy was 0.73 (range=0.33–1, SD=0.17).

2 Visual world experiment

a Mouse-click accuracy. Overall accuracy reached 0.96 (Range=0.75–1, SD=0.08) across conditions, but also varied numerically as a function of Condition (Mean_{G-S+}=1, Range_{G-S+}=1–1, SD_{G-S+}=0; Mean_{G-S-}=0.99, Range_{G-S-}=0.75–1, SD_{G-S-}=0.03; Mean_{G+S+}=0.91, Range_{G+S+}=0.75–1, SD_{G+S+}=0.12). However, an effect of Condition

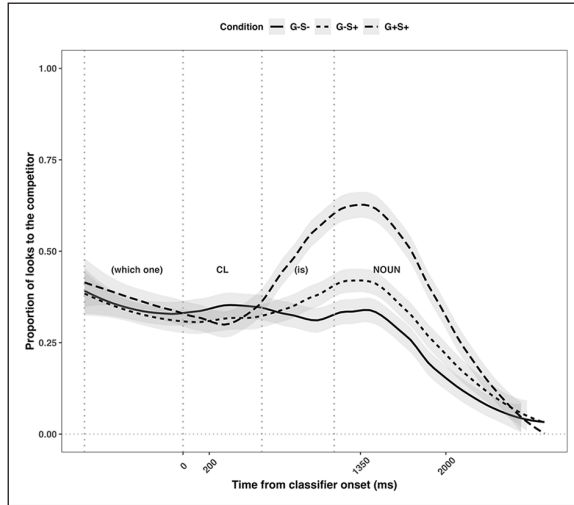


Figure 2. Proportion of looks to the competitors across conditions.

Note. For an analogous figure including looks to the targets, see supplementary material.

was not attested statistically using GLMERs. When looking closely at the descriptive data, this might be because performance across all conditions was essentially at ceiling such that individual performance was clustered closely. Given the overall high accuracy across conditions, we excluded all trials with inaccurate mouse-click responses from the eye-tracking data analysis (21 out of 628).

b Eye gaze. In addition to excluding incorrect mouse-click responses, we removed trials where participants chose the incorrect classifier–noun pairs in the vocabulary test, analogous to Analysis 2 in Grüter et al. (2020). This was because if inaccurate responses in the vocabulary test were caused by lack of knowledge of specific classifier–noun pairs, their processing could not be guided by semantic or grammatical form class information. Conservatively, then, we only included trials for which an individual’s performance on the vocabulary task indicated that they had knowledge of a specific classifier–noun mapping.

After the above exclusions, a total of 433 out of 607 trials remained. Figure 2 illustrates participants’ looking behaviours from the onset of classifiers in millisecond as the x-axis and the proportion of looks to the competitor as the y-axis. Visual inspection suggests that after the onset of the classifiers and before the onset of the nouns, participants’ looking patterns begin to diverge, suggesting they were using classifiers proactively to identify the upcoming noun. Importantly, Figure 2 suggests that their eye-gaze patterns were modulated by condition. Specifically, more looks to the competitor were observed in the G+S+ condition than in the G–S+ condition, which, in turn, induced more looks to the competitor relative to the G–S– condition.

Similar to Grüter et al. (2020), we included the time window from 200 ms post-classifier onset (marked as 200 in Figure 2) to 200 ms post-noun onset (marked as

Table 2. The model with Condition (G–S–, **G–S+** vs. G+S+) as the fixed effect.

Term	Estimate	CI	SE	Statistic	p-value
Intercept	–0.15	[–0.69, 0.40]	0.28	–0.52	.60
G–S–	–0.43	[–1.18, 0.31]	0.38	–1.15	.25
G+S+	0.87	[0.11, 1.63]	0.39	2.25	.02

Notes. $\text{glmer}(\text{comp_AOI} \sim \text{Condition} + (1|\text{Participant}) + (1|\text{Trial}), \text{data} = ., \text{family} = \text{binomial})$.

1,350 in Figure 2) as the critical region of interest (1150ms in total). For the dependent variable, the binary outcome of ‘if there was a look to the Competitor (coded as “1”) or not (coded as “0”)’ across the whole critical window was selected. This is different from the ‘TargetAdvantage’ score (the difference between the number of 20 ms bins containing looks to the competitor and the number of bins containing looks to the target) used in Grüter et al. (2020). Since the current study was conducted with web-based eye-tracking, individual participants had varying sampling rates and even the same participant had different sampling rates per item. A ‘TargetAdvantage’ score would be difficult to calculate consistently in light of such sampling differences. With the above in mind, to make the analyses/results more comparable to Grüter et al. (2020), we excluded looks to distractors and other parts of the screen. As such, all analyses essentially reflect the dynamic relationship between looks to the Competitors (‘1’s) and looks to the Targets (‘0’s), conceptually sharing the rationale behind the ‘TargetAdvantage’ score.

We, first, ran models to investigate if Location had an effect on the relative status of semantics vs. grammatical form class (Location and Condition with interaction) or on general looking patterns (Location and Condition without interaction). However, the model with the interaction between Location and Condition did not improve model fit from the model without the interaction ($\chi^2(4) = 7.40, p = .11$), which did not improve model fit from the model with Condition as the only fixed effect ($\chi^2(2) = 0.70, p = .71$). This suggests that Location did not affect the relative status of semantic vs. grammatical form class cue nor general looking patterns among the participants in the current study.

c Group level analysis (research question 1). Given the above, we fitted the model to examine any effect of Condition (G–S–, **G–S+** vs. G+S+) across all HSs from different locations (research question 1). Table 2 presents the model output. This model, together with models with different reference levels for Condition, revealed:

- a significant difference between the G+S+ condition and the G–S+ condition;
- no difference between the G–S+ condition and the G–S– condition; and
- a significant difference between the G+S+ condition and the G–S– condition (Estimate = 1.31, 95% CI [0.55, 2.07], $SE = 0.39, z = 3.37, p < .001$).

d Individual differences analysis (research question 2). To operationalize individual differences in the relative use of semantic over grammatical form class information (or vice

Table 3. The model with Condition (G–S–, **G–S+** vs. G+S+) as the fixed effect interacting with Mandarin Home Use (MHU), Mandarin Social Use (MSU), AoA E and Language Program in modulating looks to the competitor.

Term	Estimate	CI	SE	Statistic	p-value
(Intercept)	–0.17	[–0.68, 0.33]	0.26	–0.66	.51
G+S+	0.88	[0.15, 1.61]	0.37	2.37	.02
G–S–	–0.13	[–0.84, 0.59]	0.36	–0.34	.73
MHU	–0.01	[–0.04, 0.01]	0.01	–0.96	.33
MSU	0.01	[–0.01, 0.02]	0.01	0.19	.85
AoA E	0.01	[–0.07, 0.10]	0.04	0.35	.73
LP (yes)	0.26	[–0.10, 0.63]	0.19	1.42	.16
G+S+:MHU	0.02	[–0.02, 0.06]	0.02	1.00	.31
G–S–:MHU	0.01	[–0.04, 0.05]	0.02	0.23	.82
G+S+:MSU	0.01	[–0.01, 0.03]	0.01	1.20	.23
G–S–:MSU	–0.04	[–0.06, –0.01]	0.01	–3.22	.001
G+S+:AoA E	–0.10	[–0.22, 0.01]	0.06	–1.73	.08
G–S–:AoA E	–0.10	[–0.22, 0.02]	0.06	–1.62	.11
G+S+:LP (yes)	–0.26	[–0.74, 0.22]	0.25	–1.07	.29
G–S–:LP (yes)	–1.35	[–1.90, –0.80]	0.28	–4.81	<.001

Notes. LP=Language Programme; $glmer(\text{comp_AOI} \sim \text{Condition} \times (\text{MHU} + \text{MSU} + \text{AoA E} + \text{language_programme}) + (1|\text{Participant}) + (1|\text{Trial}), \text{data} = ., \text{family} = \text{binomial})$.

versa) in processing classifiers, we selected performance differences between the G–S– condition and the G–S+ condition and between the G–S+ condition and the G+S+ condition respectively. Precisely, more looks to the G–S+ condition over the G–S– condition was taken to indicate that participants used the semantic information whereas more looks to the G+S+ condition over the G–S+ condition was taken to reflect the use of grammatical form class. Therefore, the reference level for Condition was set to G–S+. To account for the factors modulating performance differences between conditions, interaction terms between Condition and language background factors were the critical information.

We fitted the maximal model that included all language background factors of interest to interact with Condition, i.e. $\text{Condition} \times (\text{MHU} + \text{MSU} + \text{AoAE} + \text{language_programme})$. Convergence was achieved for the maximal model (Table 3).

As we are (primarily) interested in how different factors modulate the greater or lesser use of semantic (difference between the G–S– and the G–S+ conditions) and grammatical form class (difference between the G+S+ and the G–S+ conditions) cues, we focus on the interaction terms in Table 3 instead of any simple effects. As also evident in Figure 3 which visualizes the interaction terms, an interaction with Condition was significantly attested for MSU and Language Programme. More specifically, participants' performance showed a larger difference between the G–S– and the G–S+ conditions if they had higher MSU scores or attended HL programmes.

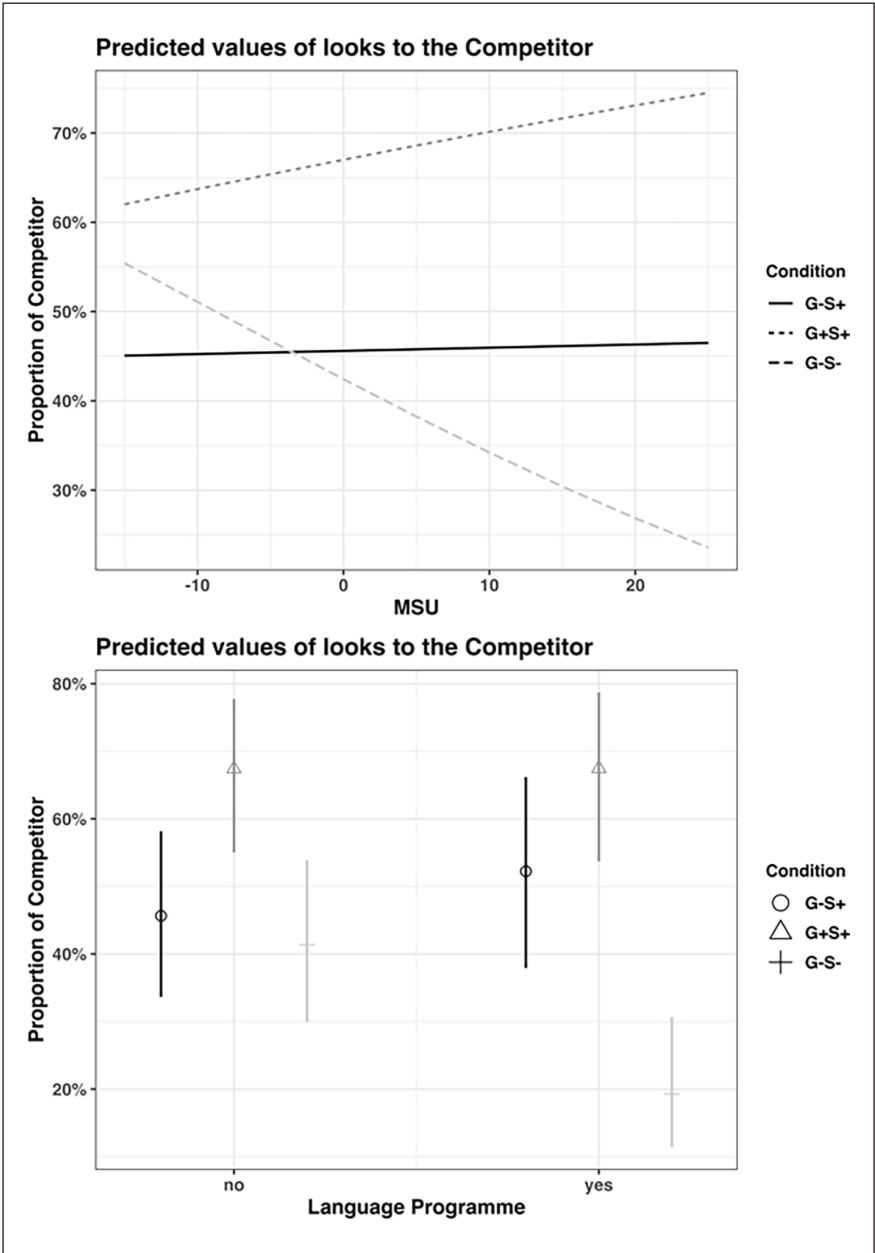


Figure 3. Visualization of interaction between Condition and Mandarin Social Use (MSU; top), and between Condition and Language Programme (bottom).
Note. MSU and is centred such that 0 is the mean.

VI Discussion

The present study adopted the visual world eye-tracking experiment from Grüter et al. (2020) to examine the performance patterns of Mandarin–English HSs. Beyond the group level (research question 1), we address the qualitative nature of individual differences in how HSs use sortal classifiers during real time comprehension, to probe if and how bilingual language experience factors modulate individual HS performances (research question 2). Given that we used Grüter et al.’s experimental materials, albeit in a web-based rather than a lab-based eye-tracking paradigm, it is both natural and interesting to begin our discussion of the current data with reference to how the groups therein performed on the same task. However, as our focus is on the independent value of HS performances (De Houwer, 2023; Ortega, 2020; Paradis, 2023; Rothman et al., 2023), reference to Grüter et al.’s group results is largely expository. This is especially true because of the differences in the modalities of tasks (online vs. in-lab), which also led to differential data analysis strategies.

Research question 1 investigated how different types of cues related to sortal classifiers are used online in real time sentential comprehension at the HS group level. Given that HSs are naturalistic, young (native) acquirers of the HL (Rothman and Treffers-Daller, 2014), we predicted that HSs would primarily use the grammatical form class to predict the upcoming nouns and would not rely on the semantic information when nouns are not selected by the classifiers in grammatical form class. To the extent that the results in the present study can be compared to those of Grüter et al. (2020), the data suggest that our HS aggregate performed similarly compared to Grüter et al.’ L1-dominant users and by extension distinctly from the late L2 bilinguals. Specifically, for our HS group, similar to the patterns observed in L1-dominant users as a group in Grüter et al. (2020), there was a significant performance difference between the G+S+ condition and the G–S+ condition (index of the use of grammatical form class) and no difference between the G–S+ condition and the G–S– condition (index of the use of semantics). Importantly, the latter was different from the L2ers in Grüter et al. (2020) who showed a significant performance difference between the G–S+ condition and the G–S–. Despite the reduced nature of input and usage patterns of Mandarin that can be assumed by virtue of the HS context, which could have favoured the use of the associated underlying semantics of classifiers to fill apparent quantitative gaps (HSs having much less experience with Mandarin), the relative similarities between the present HSs and Grüter et al.’s L1-dominant users are not so surprising. Indeed, they are in line with previous work showing that Mandarin HSs perform similarly in processing classifier–noun pairs at the phrasal level (Li et al., 2021). Moreover, the aggregate HS performance is also in line with existing work showing that HSs can adopt the same online processing strategies deployed by homeland L1-dominant users (Di Pisa et al., 2022; Fuchs, 2022a, 2022b; Hao et al., 2024; Jegerski, 2018a, 2018b; Jegerski et al., 2016; Luque et al., 2023).

With the above in mind, it is worth discussing how the present findings are partially at odds with what has been observed by Li et al. (2021) among Mandarin–Malay adult HSs who showed semantic-related processing of classifiers. We believe that this difference might reflect the distinct designs and samples in the two studies. While Li et al. (2021) tested the comprehension of classifier–nouns in isolated phrases, our

experimental stimuli have classifier–noun pairings embedded in sentential contexts, whereby syntactic cues might matter more strongly. Moreover, their nouns did not share semantic features in the mismatched classifier–noun pairs, as in the G–S+ condition in the current study. Furthermore, the societal dominant language of the HSs in Li et al. (2021), Malay, is also a classifier language. As such, it is possible that the existence of a classifier system in the societal dominant language enhanced those HSs’ sensitivity to the mismatch between classifiers and nouns in Mandarin. If so, this would be an effect of cross-linguistic influence (CLI) from the societal dominant language, which is not possible for the current HSs sample given that the societal majority language is English.

Of course, speaking another classifier language(s) or not would be another individual-level factor modulating individual differences (e.g. Paradis, 2023). Given the dataset we have at our disposal (lack of information on the usage of languages other than Mandarin and English) and the recent findings that CLI can be modulated by more speaker-centric individual-level factors, e.g. exposure and usage, age, etc. (e.g. Chondrogianni, 2023; Hao et al., 2024), we encourage future studies designed *a priori* to directly examine speaker-centric modulatory factors related to CLI in the domain of classifiers. In such studies, going beyond comparing societal dominant languages with and without a classifier system to examining the potentially differential roles of classifier languages that differ in their (dis)allowance of G+S– mapping and G–S+ mapping would provide us with finer-grained understanding of CLI in general as well as its effects on Mandarin sortal classifiers specifically.

Having established what the HS aggregate data demonstrate, a positive precursor to best contextualizing and unpacking latent individual differences within the cohort, we now turn to research question 2: What, if any, individual level factors modulate the greater or lesser reliance on semantic vs. grammatical form class cues? We predicted that individual HSs are likely to show differences and that such variation would not be random, but rather correlate to any number of dynamic variables related to HL exposure (quantity and type) and usage. Based on previous production studies with Mandarin child HSs, we anticipated HL exposure and use (Kan, 2019), AoA of English (Jia and Paradis, 2015; Kan, 2019), and HL schooling (Jia and Paradis, 2015) to modulate individual performance. As we saw in Section V, the current sample showed an effect of HL exposure and use (MSU) as well as HL schooling. The fact that AoA of English did not correlate with performance in the present sample is not very surprising when one considers that its variation in the current sample is quite small and bimodally distributed (either at 0 or 5 in years).

For the effect of HL schooling, having attended Mandarin language programmes led to greater reliance on semantic information (when the classifier and noun do not co-occur with each other). To put this into the context of the L1-dominant users vs. L2ers from Grüter et al. (2020), HSs who attended Mandarin language programmes were more likely to perform like L2ers. Indeed, performance more akin to L2ers given formal instruction is not necessarily surprising to the extent that there are certain parallels in some HSs’ experience with Mandarin that are distinct from other groups of native speakers – other HSs and L1-dominant users alike – who do not display similar behaviour. To pursue the actual causal link, should one exist, we would admittedly need to have information we

do not have access to, namely, knowledge of exactly how classifiers are taught to both L2ers and HSs in their respective formal settings. That acknowledged, it is perhaps helpful to elucidate a bit further what we have in mind and where targeted, future work could start to look. The general working hypothesis here is that HSs of Mandarin, at least in English environments, are instructed similarly and explicitly as L2ers for this domain of grammar. That is, each are likely to be instructed about the semantic feature basis of the prototypical meanings of given classifiers, whereas the average L1-dominant user and HSs who lack formal instruction in Mandarin are unlikely to be consciously aware of this. Somewhat akin to what was discussed in Grüter et al. (2012) for gender processing in L2 Spanish versus L1-dominant Spanish, the question is not whether classifiers can be acquired or form part of the mental representations of HS or L2 grammars per se but rather relates to distinctions to the conditions under which a given property is acquired/learned by different learner types. In the present case, HSs who also have formal instruction are more likely to have shared with L2ers a type of experience – we surmise explicit explanation and practice of classifiers – that leads them to greater sensitivity to semantic features when processing classifiers, especially (maybe exclusively) under experimental conditions, which themselves introduce a level of formality more comparable to an instructed setting than mere conversation would.

As for the effect of MSU, participants with more HL exposure and use in societal contexts (larger MSU scores) showed increased sensitivity to semantic features, independently of having (or not) had formal training in Mandarin. Upon examining the contributing questions/components for the MSU score in detail, it is not surprising to observe the same directionality of the effect of MSU and the effect of Schooling. The majority of questions/components included for the calculation of the MSU score targets participants' (informal) literacy practices, e.g. school, email, text, social media, movies, internet, switching on social media, TV, lists, reading. While it is certainly possible that there is no connection between the effect of MSU and formal schooling, we would like to explore the basis of an idea that underlyingly links them: the effect of MSU indexes an effect of (informal) literacy training which could result in metalinguistic awareness for grammatical patterns (Bayram et al., 2019). Specifically here, the pattern that sortal classifiers map with nouns based on semantic categorization. As is true above; to scrutinize this hypothesized link more meaningfully one would need to have information we do not have access to, specifically questions that probe the particulars of HL literacy engagement more deeply. Recall, our questions around Mandarin training simply asked whether they had been enrolled in formal classes of Mandarin, not whether they had engaged in any type of self-training, informal literacy training or any other activities that could have made the semantic dimensions of classifiers salient to them, feeding into some type of metalinguistic knowledge. Future studies may probe more deeply into what the connections, if any, are between increased informal literacy practices and formal training.

Given the link we speculated above, it would be reasonable at first glance to ponder why Mandarin L1-dominant users do not also behave this way, perhaps even more so. After all, they would have even more metalinguistic knowledge in the way we have described. That said, the Mandarin L1-dominant users in Grüter et al. (2020) did not perform in the same way, relying on semantics. We are not surprised by this, nor do we take their performance to question the potential veracity of our speculation for the following

two converging reasons. We begin by highlighting some things particular to the present findings and then move onto a more general view in terms of types of metalinguistic awareness and how they might play out differently between HSs and L1-dominant users.

In the current experiment, participants' reliance on semantics is reflected by the observation that the G-S+ condition induced more looks to the competitor than the G-S- condition. Recall that the G-S+ classifier-noun pairs are not actually licenced by the grammar of Mandarin. With this in mind, we now answer the question of why we believe more metalinguistic awareness does not lead to reliance on semantics in L1-dominant users. Two points are critical here. First, in the L1-dominant users' mental representation, G-S+ pairs are not available or perhaps while L1-dominant users do make use of semantics, the unlicensed G-S+ pairs are simply not processed. However, HSs process the G-S+ pairs given their variable knowledge of the grammatical form class licensing of classifier noun pairings.

Second, metalinguistic knowledge matters but potentially differentially so for different types of speakers and within each type of speakers, highlighting the importance of an individual differences approach. It could be the case – in fact we think it is – that informal literacy leads HSs to misleading patterns for which, unlike L1-dominant users, they cannot easily override. In the present case, informal literacy, corresponding as we surmise to increased quantities of input, might lead the HS to notice semantic patterns that are quite robust generally speaking. And while we would expect them to have higher quantities of input, this input would still be significantly reduced to that of a L1-dominant users. When the grammatical class does not correspond to the prototypical semantic cue, these are exceptions. L1-dominant users typically do not have issues with expectations, presumably because the input affords the learning of them as individual lexical items. We know that overall reduced input (and ensuing smaller lexicons) in HSs has significant consequences for exceptions, much like it does for children in L1 homeland environments who eventually receive enough input to learn exceptions in the expected adult like pattern (Fernández-Dobao and Herschensohn, 2021; Montrul and Mason, 2020; Uygun et al., 2023). For example, HSs of Spanish often apply the regular conjugation rules for past tense for irregular verbs. Doing so is logical in the context of reduced input (yet different from L1-dominant users) and actually indicates that such issues are not grammatical *per se*, but rather exist at the level of mental lexical entry. Applying the 'regular' pattern in what could be viewed as a hyper extended way is also more likely when the speaker is aware (at some level) that their language is somewhat stigmatized, as is often the case for HSs. For example, work by Bullock et al. (2014) have shown that L1-dominant users of Dominican Spanish in rural contexts that have pervasive phonological 's' deletion that is highly stigmatized often over (and wrongly) produce exponents of morphological 's' (where it has the +plural value) when they are trying to sound more formal. Such speakers are aware of the morphological rules and very aware that they are judged when they delete the 's', and so when trying to monitor or be formal they not only consciously make sure they pronounce it, but rather they can extend it to cases where it should not be used by overcompensating or hyper-correction. Our point is that conscious rules of regularity, which can obtain via instructed or informal metalinguistic awareness, can sometimes be wrongly applied as a hyper-extension. And so, we do believe in general that metalinguistic awareness can have vastly different

consequences for various types of learners: L1-dominant users rely less typically on metalinguistic knowledge in experimental contexts, despite potentially having more than other sets of speakers.

VII Conclusions

The present study aimed to examine the status and relative use of semantic and grammatical form class cues in Mandarin–English HSs’ processing of sortal classifiers. In addition, given the importance of understanding individual differences, the present study examined if/how and why individual HSs may show differences from each other as a function of their experience with Mandarin. For HSs at the group level, the present results suggested that HSs predictively used classifiers during sentence processing with a preference for grammatical form class cue over and above the semantic cue (when the nouns and classifiers do not match in grammatical form class). Crucially, at the individual level, the study showed that not all HSs were the same: linguistic experience matters. That is, Mandarin literacy engagement and more Mandarin exposure and use in societal contexts resulted in increased probability that an individual would rely more heavily on the semantic cue. We interpreted these findings to underscore the potential role for metalinguistic knowledge, be it from formal (HL schooling) or potentially informal (social literacy practices) sources, in the processing of HLs. Overall, the study showcased the utility of an individual difference approach in understanding the individual ways in which HSs process different linguistic cues. HL acquisition is not different from language acquisition for all other speaker types: it is not a uniform process for all, but rather reflects individual experience/engagement opportunities.

Acknowledgements

The authors wish to thank Professor Yow Wei Quin and her lab at Singapore University of Technology and Design for their generous help in participant recruitment. We are also grateful for valuable *Second Language Research* reviewers’ and editors’ comments that have led to many improvements of our original version.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Jason Rothman, Jiuzhou Hao, and Maki Kubota were funded via the Tromsø Forskningsstiftelse (Tromsø Research Foundation) grant number A43484: Heritage-bilingual Linguistic Proficiency in their Native Grammar (HeLPiNG) (2019–2023).


Data availability statement

Supplementary materials and the data that support the findings of this study are openly available in OSF at <https://osf.io/qcxnh> (accessed August 2024).

ORCID iDs

Jiuzhou Hao  <https://orcid.org/0000-0003-3730-0528>

Fatih Bayram  <https://orcid.org/0000-0002-5024-3612>

Jorge González Alonso  <https://orcid.org/0000-0001-5047-3226>

Theres Grüter  <https://orcid.org/0000-0001-6354-9787>

Supplemental material

Supplemental material for this article is available online.

References

- Anderson JAE, Mak L, Keyvani Chahi A, and Bialystok E (2018) The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior Research Methods* 50: 250–63.
- Anwyl-Irvine AL, Massonnié J, Flitton A, Kirkham N, and Evershed JK (2020) Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 52: 388–407.
- Barr DJ, Levy R, Scheepers C, and Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68: 255–78.
- Bates D, Mächler M, Bolker B, and Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48.
- Bayram F, Rothman J, Iverson M et al. (2019) Differences in use without deficiencies in competence: Passives in the Turkish and German of Turkish heritage speakers in Germany. *International Journal of Bilingual Education and Bilingualism* 22: 919–39.
- Bullock BE, Toribio AJ, and Amengual M (2014) The status of s in Dominican Spanish. *Lingua* 143: 20–35.
- Chondrogianni V (2023) Cross-linguistic influences in bilingual morphosyntactic acquisition. In: Elgort I, Siyanova-Chanturia A, and Brysbaert M (eds) *Bilingual processing and acquisition*. Amsterdam: John Benjamins, pp. 294–315. Available at: <https://benjamins.com/catalog/bpa.16.12cho> (accessed August 2024).
- Daskalaki E, Chondrogianni V, and Blom E (2022) Path and rate of development in child heritage speakers: Evidence from Greek subject/object form and placement. *International Journal of Bilingualism* 27: 634–62.
- De Houwer A (2023) The danger of bilingual–monolingual comparisons in applied psycholinguistic research. *Applied Psycholinguistics* 44: 343–57.
- Di Pisa G, Kubota M, Rothman J, and Marinis T (2022) Effects of markedness in gender processing in Italian as a heritage language: A speed accuracy tradeoff. *Frontiers in Psychology* 13: 965885.
- Dominguez L, Hicks G, and Slabakova R (2019) Terminology choice in generative acquisition research: The case of ‘incomplete acquisition’ in heritage language grammars. *Studies in Second Language Acquisition* 41: 241–55.
- Erbaugh MS (2012) Chinese classifiers: Their use and acquisition. In: Li P, Tan LH, Bates E, and Tzeng OJL (eds) *The handbook of East Asian psycholinguistics*. Cambridge: Cambridge University Press, pp. 39–51.
- Fernández-Dobao A and Herschensohn J (2021) Acquisition of Spanish verbal morphology by child bilinguals: Overregularization by heritage speakers and second language learners. *Bilingualism: Language and Cognition* 24: 56–68.
- Fishman J (2014) Three hundred-plus years of heritage language education in the United States. In: Wiley TG, Peyton JK, Christian D, Moore SCK, and Liu N (eds) *Handbook of heritage*,

- community, and Native American Languages in the United States: Research, policy, and educational practice. New York: Routledge.
- Flores C, Santos AL, Jesus A, and Marques R (2017) Age and input effects in the acquisition of mood in Heritage Portuguese. *Journal of Child Language* 44: 795–828.
- Fuchs Z (2022a) Facilitative use of grammatical gender in Heritage Spanish. *Linguistic Approaches to Bilingualism* 12: 845–71.
- Fuchs Z (2022b) Eyetracking evidence for heritage speakers' access to abstract syntactic agreement features in real-time processing. *Frontiers in Psychology* 13: 960376.
- Gao MY and Malt BC (2009) Mental representation and cognitive consequences of Chinese individual classifiers. *Language and Cognitive Processes* 24: 1124–79.
- Grüter T, Lau E, and Ling W (2020) How classifiers facilitate predictive processing in L1 and L2 Chinese: The role of semantic and grammatical cues. *Language, Cognition and Neuroscience* 35: 221–34.
- Grüter T, Lew-Williams C and Fernald A (2012) Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research* 28(2): 191–215.
- Hao J and Chondrogianni V (2023) Comprehension and production of non-canonical word orders in Mandarin-speaking child heritage speakers. *Linguistic Approaches to Bilingualism* 13: 468–99.
- Hao J, Chondrogianni V, and Sturt P (2024) Heritage language development and processing: Non-canonical word orders in Mandarin–English child heritage speakers. *Bilingualism: Language and Cognition* 27: 334–49.
- Huetting F, Chen J, Bowerman M, and 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.
- Jegerski J (2018a) Sentence processing in Spanish as a heritage language: A self-paced reading study of relative clause attachment: Sentence processing in Spanish as a heritage language. *Language Learning* 68: 598–634.
- Jegerski J (2018b) The processing of the object marker *a* by heritage Spanish speakers. *International Journal of Bilingualism* 22: 585–602.
- Jegerski J, Keating GD, and VanPatten B (2016) On-line relative clause attachment strategy in heritage speakers of Spanish. *International Journal of Bilingualism* 20: 254–68.
- Jia R and Paradis J (2015) The use of referring expressions in narratives by Mandarin heritage language children and the role of language environment factors in predicting individual differences. *Bilingualism: Language and Cognition* 18: 737–52.
- Kan RT (2019) Production of Cantonese classifiers in young heritage speakers and majority language speakers. *International Journal of Bilingualism* 23: 1531–48.
- Kupisch T and Rothman J (2018) Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *International Journal of Bilingualism* 22: 564–82.
- Kwon N, Sturt P, and Liu P (2017) Predicting semantic features in Chinese: Evidence from ERPs. *Cognition* 166: 433–46.
- Li F, Hong X, He Z, Wu S, and Zhang C (2021) Investigating heritage language processing: Meaning composition in Chinese classifier–noun phrasal contexts. *Frontiers in Psychology* 12: 782016.
- Li W and Lee S (2001) L1 development in an L2 environment: The use of Cantonese classifiers and quantifiers by young British-born Chinese in Tyneside. *International Journal of Bilingual Education and Bilingualism* 4: 359–82.
- Lloyd-Smith A, Bayram F, and Iverson M (2020) The effects of heritage language experience on lexical and morphosyntactic outcomes. In: Bayram F (ed) *Studies in Bilingualism*.

- Amsterdam: John Benjamins, pp. 63–84. Available at: <https://benjamins.com/catalog/sibil.60.041lo> (accessed August 2024).
- Luque A, Rossi E, Kubota M et al. (2023) Morphological transparency and markedness matter in heritage speaker gender processing: An EEG study. *Frontiers in Psychology* 14: 1114464.
- McElreath R (2020) *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*, 2nd edition. London: Chapman and Hall / CRC.
- Montrul S (2008) *Incomplete acquisition in bilingualism: Re-examining the age factor*. Amsterdam: John Benjamins.
- Montrul S (2016) *The acquisition of heritage languages*. Cambridge: Cambridge University Press.
- Montrul S, de la Fuente I, Davidson J, and Foote R (2013) The role of experience in the acquisition and production of diminutives and gender in Spanish: Evidence from L2 learners and heritage speakers. *Second Language Research* 29: 87–118.
- Montrul S and Mason SA (2020) Smaller vocabularies lead to morphological overregularization in heritage language grammars. *Bilingualism: Language and Cognition* 23: 35–36.
- Montrul S and Potowski K (2007) Command of gender agreement in school-age Spanish–English bilingual children. *International Journal of Bilingualism* 11: 301–28.
- Nahm FS (2016) Nonparametric statistical tests for the continuous data: The basic concept and the practical use. *Korean Journal of Anesthesiology* 69: 8–14.
- Ortega L (2020) The study of heritage language development from a bilingualism and social justice perspective. *Language Learning* 70: 15–53.
- Papoutsaki A, Sangkloy P, Laskey J et al. (2016) Webgazer: Scalable webcam eye tracking using user interactions. In: Kambhampati S (ed), *Proceedings of the twenty-fifth international joint conference on artificial intelligence (IJCAI'16)*, pp. 3839–45. Cambridge, MA: AAAI Press.
- Paradis J (2023) Sources of individual differences in the dual language development of heritage bilinguals. *Journal of Child Language* 50: 793–817.
- Pascual y Cabo D and Rothman J (2012) The (Il)logical problem of heritage speaker bilingualism and incomplete acquisition. *Applied Linguistics* 33: 450–55.
- Polinsky M (2008) Gender under incomplete acquisition: Heritage speakers' knowledge of noun categorization. *Heritage Language Journal* 6: 40–71.
- Polinsky M (2018) Bilingual children and adult heritage speakers: The range of comparison. *International Journal of Bilingualism* 22: 547–63.
- R Core Team (2018) *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: <https://www.R-project.org> (accessed August 2024).
- Regulez BA and Montrul S (2023) Production, acceptability, and online comprehension of Spanish differential object marking by heritage speakers and L2 learners. *Frontiers in Psychology* 14: 1106613.
- Rothman J (2009) Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. *International Journal of Bilingualism* 13: 155–63.
- Rothman J, Bayram F, DeLuca V et al. (2023) Monolingual comparative normativity in bilingualism research is out of 'control': Arguments and alternatives. *Applied Psycholinguistics* 44: 316–29.
- Rothman J and Treffers-Daller J (2014) A prolegomenon to the construct of the native speaker: Heritage speaker bilinguals are natives too! *Applied Linguistics* 35: 93–98.
- Singmann H, Bolker B, Westfall J et al. (2020) *afex: Analysis of factorial experiments*. Manual. Available at: <https://CRAN.R-project.org/package=afex> (accessed August 2024).
- Soto-Corominas A, Daskalaki E, Paradis J, Winters-Difani M, and Al Janaideh R (2022) Sources of variation at the onset of bilingualism: The differential effect of input factors, AOA, and cognitive skills on HL Arabic and L2 English syntax. *Journal of Child Language* 49: 741–73.

- Tsang C and Chambers CG (2011) Appearances aren't everything: Shape classifiers and referential processing in Cantonese. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37: 1065–80.
- Uygun S, Schwarz L, and Clahsen H (2023) Morphological generalization in heritage speakers: The Turkish aorist. *Second Language Research* 39: 519–38.
- Valdés G (2000) The teaching of heritage languages: An introduction for Slavic-teaching professionals. In: Kagan O and Rifkin B (eds) *The learning and teaching of Slavic languages and cultures*. Bloomington, IN: Slavica, pp. 375–403.
- Winter B (2019) *Statistics for linguists: An introduction using R*. New York: Routledge. Available at: <https://www.taylorfrancis.com/books/9781351677431> (accessed August 2024).