Language Experience Impacts L2 Scope Computation in English

LeeAnn Stover

Department of Linguistics, CUNY Graduate Center

August 18, 2022

Abstract

The current study examines how the bilingual experience of Heritage Speakers (HSs) impacts the computation of doubly-quantified constructions in English such as *'**Every** shark is attacking **a** pirate'*, which are believed to have two possible readings in English due to quantifier raising but are argued to be unambiguously scope-rigid in Mandarin Chinese. Research questions probe 1) whether HSs pattern more in line with the well-documented *Processing Scope Economy (PSE)* or an emergent *Avoidance of Ambiguity (AA)* strategy, and 2) if group differences will arise based on bilingual experience between HSs and their late second-language acquiring Mandarin-English counterparts. Forty-three highly proficient bilingual participants completed a language background questionnaire and an online experiment using novel PC Ibex software, where they completed a forced choice task which involved hearing an aural stimulus, selecting which of two pictures best matches the sentence they heard, and then rating the picture they selected for appropriateness. Results show mixed support for the *PSE* and *AA*, suggesting that HSs can compute inverse scope (and more than L2s) in line with the *PSE* but may also prefer to avoid it in line with the *AA* strategy when a surface scope interpretation is available.

Language Experience Impacts L2 Scope Computation in English

Despite its prevalence in virtually all human decision making, ambiguity parsing is a notably complex phenomenon. A successful resolution of a structurally ambiguous sentence which has more than one potential meaning requires interaction between the syntax-semantics-pragmatics interface, which incurs a large cognitive load. Syntactic ambiguity processing induces a processing cost (Bornkessel et al., 2004), and even constrains working memory (MacDonald et al., 1992). In fact, because of its involvement of the interfaces and cognitive difficulty, ambiguity parsing has been used as a tool to understand the basic principles underlying basic human comprehension (see Fodor et al., 1974; Frazier & Fodor, 1978; Frazier & Rayner, 1982).

This complexity in parsing ambiguity is potentially compounded among bilinguals with more than one competing grammar. Late-learners of a language may have difficulty acquiring syntactic ambiguity in that language (see Dussias, 2003; Frenck-Mestre & Pynte, 1997). To minimize the cognitive demands on the bilingual processor, it has been proposed that heritage speakers, a group of bilinguals characterized by their dominance in a societally-dominant L2 over their L1 home language, might disprefer ambiguity in both their L1 and L2 (Polinsky & Scontras, 2020). However, not much is known about this possible strategy employed by this group of bilinguals.

The current study explores syntactic ambiguity parsing among two groups of bilinguals using structurally ambiguous English constructions with two quantifier phrases. I probe an understudied proposal (Avoidance of Ambiguity) utilizing novel remote software (PC Ibex) and a non-standard methodology that (1) directly compares two bilingual groups rather than relying on monolingual comparisons and (2) sometimes forces a choice between two dis-preferred readings.

Quantifier scope ambiguity

A doubly quantified phrase has two landing sites at logical form (LF), and if the quantified expression *a pirate* (in 1) raises to a position higher than the other DP *every shark*, it scopes over *every shark* and a (1b) interpretation is possible via quantifier raising (Kratzer & Heim, 1998; May, 1978). Similarly in (2), *every mouse* could raise to a position higher than *a cat* which allows for both surface (narrow) and inverse (wide) scope interpretations. Both word orders give rise to surface-inverse scope ambiguity.

(1) *Every shark is attacking a pirate*.

a. SURFACE SCOPE $\forall x\ [\text{shark}(x)\ \&\ \exists y\ [\text{pirate}\ (y) \text{->} \text{attacking}\ (x,y)]]$

For every shark, there is a (different) pirate that the shark is attacking.

b. INVERSE SCOPE $\exists x\ [\text{pirate}(x) \text{->} \forall y\ [\text{shark}(y)\ \&\ \text{attacking}\ (y,x)]]$

There is a single pirate that every shark is attacking.

(2) *A cat is chasing every mouse*.

a. SURFACE SCOPE $\exists x\ [\text{cat}(x)\ \&\ \forall y\ [\text{mouse}(y) \text{->} \text{chasing}\ (x,y)]]$

There is a single cat that is chasing every mouse.

b. INVERSE SCOPE $\forall x\ [\text{mouse}(x) \text{->} \exists y\ [\text{cat}(y)\ \&\ \text{chasing}\ (y,x)]]$

For every mouse, there is a (different) cat that is chasing each mouse.

For native speakers of English, both the surface scope and inverse scope readings have been found to be available in sentences containing two quantifiers. There is a robust preference for the surface/narrow scope (1a and 2a) which has been found using a variety of methodologies and tasks both explicit (see Kurtzman & MacDonald, 1993) and implicit (see Anderson, 2004;

Filik et al., 2004; Paterson et al., 2008), but it is also widely attested that the inverse scope is computable by English speakers of many ages (Lidz & Musolino, 2002). Crucially, a wide scope interpretation is not deemed possible in languages such as Mandarin (Aoun & Li, 1993; Huang, 1982). For Mandarin speakers, sentences of this type are argued to only have surface scope as a possible interpretation. This scope rigidity is described in Huang's *Isomorphic principle*: the only scope relationship possible at the surface level is the c-command relationship between the two quantifiers/scope-bearing elements.

While both (1) and (2) are considered ambiguous with two possible meanings in English, the difference between the two structures is non-trivial. The inverse interpretation of (1), where the quantifier *every* scopes over *a* at surface structure, entails the surface interpretation. In other words, for one pirate to be attacked by every shark, this also entails that every shark attacked a pirate. Because of this, Scontras et al (2017) argue that it is a poor test case to probe the availability of inverse scope. Additionally, some have argued that every>a sentences may give rise to ambiguity in Mandarin (Wu et al., 2017; Zhou & Gao, 2009). While English and Mandarin still clearly have representations of scope in their grammars that differ from one another, the current study includes both sentence types in analysis to probe the possible effect of quantifier position that may lead to conflicting scope rigidity in Mandarin or differentially affect HSs (though I have no specific predictions for HS behavior based on quantifier position).

Two Alternate Explanations: PSE vs Avoidance of Ambiguity

The inverse scope (1-2b) is more costly to process than surface scope (1-2a), possibly due to a *Processing Scope Economy (PSE)* principle (Anderson, 2004). According to the *PSE,* individuals prefer to parse a construction with the fewest syntactic operations. Computing a more complex operation (such as inverse scope) is possible, but this computation incurs a processing cost. Since the surface scope does not require further movement operations, it is a simpler and thus preferred parse. The *PSE* has been supported across monolinguals, and has also been argued for L2 learners (Chu et al., 2014) as a potential explanation for why they did not accept inverse scope. Since the inverse scope is more costly to compute, these intermediate and advanced L2 learners had yet to acquire it.

An alternate, emergent hypothesis of *Avoidance of Ambiguity (AA)* among HSs has been proposed by Polinsky and Scontras (2020). In a seminal paper about understanding and modeling heritage grammars, they note that phenomena involving ambiguity suffer in heritage grammars. That is, HSs struggle with one-to-many mappings between form and meaning. Polinsky and Scontras demonstrate this argument with two different cases: Japanese topics (Laleko & Polinsky, 2013, 2016, 2017) and Chinese scope (Scontras et al., 2017). The researchers argue that in situations of language contact, heritage bilinguals “lose the ability to successfully generate and resolve ambiguities” (Polinsky & Scontras, 2020, p. 12). In both the heritage and dominant languages HSs seem to simplify scope grammar, preferring one-to-one mappings and eliminating ambiguity even when it is available in the baseline grammar. Polinsky and Scontras attribute this *AA* to either amplified cognitive pressures in language contact situations or reduced experience with the full range of baseline meaning options. However, they also acknowledge that this *AA* hypothesis (and the others in their paper) is just a starting point toward a model of heritage grammars, and that much more empirical work must be conducted to progress in this endeavor.

The *PSE* predicts that English-dominant HSs and, to a lesser extent, highly proficient L2s should be able to compute a more complex configuration (inverse/wide scope), even though they will incur a greater processing cost than surface/narrow scope. In contrast, the *AA* principle predicts that, as an end result of changes in heritage grammars, English-dominant HSs will disallow inverse scope in both their heritage and second language. In some ways, these two strategies predict different outcomes (HSs computing or not computing inverse scope in their dominant language). However, they need not contrast entirely from each other. From my understanding, the *AA* principle does not strongly claim that the inverse scope is not computable at all for HSs. Additionally, if an L2 learner has not acquired inverse scope, it does not mean that they will never be able to (and the *PSE* could still align with this outcome). While studies may support one strategy more than the other, these two theories need not be mutually exclusive.

Effects of Bilingualism

While the *PSE* and *AA* are both possible explanations for how HSs parse constructions with more than one scopal DP, there are additional effects of bilingualism that impact all aspects of cognition and processing. These effects are also important to consider when comparing HSs to other groups of bilinguals, namely their L2 late bilingual counterparts.

Heritage Speakers

Heritage speakers, bilinguals whose first-learned language (L1) is acquired naturalistically at home but are educated, work, and often communicate with peers in their second-learned language (L2), are a difficult group to classify yet increasingly important to understand. There are estimated to be more than 5.5 million heritage bilinguals (HSs) in New York state alone (Nagano, 2015), and about 38% of undergraduate students at CUNY are native in a language other than English. Furthermore, an estimated 30% of NYC's 565,000 Chinese Americans were born in the US (Asian American Federation, 2019), making the study of heritage Chinese and the way HSs construct their grammars a vital topic to our community.

There are a few existing studies that have demonstrated difficulty among HSs to resolve ambiguities. HSs show difficulties resolving ambiguities when it comes to Japanese topics

(Laleko & Polinsky, 2013, 2016, 2017), quantifier and negation scope (Wu & Ionin, 2019), and distinctions between plural definites and bare plurals in Romance generics (Kupisch, 2012; Montrul & Ionin, 2010). Scontras et al. (2017) examined the same structure being studied presently: doubly-quantified constructions in English (and Mandarin). They found that, like native speakers of Mandarin, HSs lack inverse scope in Mandarin. The study also found that HSs lack inverse scope in their dominant English, while their native English-speaking counterparts did not.

The current study is an indirect expansion of Scontras et al (2017), who compared acceptability judgments of Mandarin HSs in their heritage Mandarin and dominant English with judgments of native Mandarin and English speakers when computing doubly quantified sentences. Using a web-based platform called ExperigenRT, Scontras et al (2017) gave all three groups an acceptability judgment task where participants saw an image on the screen, clicked to hear an aural sentence, and then answered whether the audio appropriately described the picture using a 7-point Likert scale. The picture corresponded to either a surface or inverse interpretation of the aural stimulus. This study found that Mandarin HSs rated the inverse image significantly lower than native English speakers (Mandarin HSs = 2.55, native English = 4.46). HSs also rated the inverse image in the Mandarin experiment higher than the native Mandarin baseline (HSs = 2.79, native Mandarin = 1.56) which the authors attributed to a yes-bias. They argue that, since dominant-language transfer is unlikely due to the lack of availability of inverse scope in the English experiment, the lack of inverse scope is more likely due to a more default and less encumbered scope-calculation system.

Late Bilinguals

There are many different terms for first-generation immigrants who acquire a second language later in life as they become immersed in a society that speaks a language that differs from their L1. I will refer to these individuals as late second language bilinguals (L2s). This does not make any specific claims on when or how these speakers learned/acquired their L2, but rather simply denotes that these bilinguals continue to be dominant in their first language that is not the societal majority language. Of note here as a comparison group to HSs is that these late L2 acquirers are the time apparent-parents of HSs, and that L2ers are often classified by their age of acquisition and/or their ultimate attainment in the later-learned language. In other words, the emphasis is on L2 acquisition and proficiency rather than L1 attainment (like HSs). While HSs tend to be dominant in the societal majority language (in this case English), L2s are more likely to be dominant in their first-learned language (Mandarin for the current study). However, while the end state grammars of late L2 acquiring bilinguals can vary widely, it is entirely possible for these bilinguals to become highly proficient in their second-learned language (see Hopp, 2010).

Cross-linguistic differences between an L1 and an L2 can notoriously cause difficulties for learners in second language acquisition (SLA), particularly those phenomena within an interface between semantics, syntax, and pragmatics such as quantifier scope ambiguity. Since native English adult speakers can seemingly access both surface and inverse scope readings while Mandarin native speakers only allow surface scope, possible L1 transfer would lead adult Mandarin-native English-language L2s to initially not be able to access inverse scope readings. Several studies have analyzed scope ambiguity acquisition among L2 learners who do not have this feature available in their L1 (Marsden, 2009; Scontras et al., 2017; Wu & Ionin, 2019). While it has been found that L1 Mandarin speakers acquiring L2 English can successfully learn inverse scope (Wu & Ionin, 2022), this finding is mixed (Chu et al., 2014) and it is nevertheless difficult for them to generalize the availability of inverse scope (Wu & Ionin, 2019, 2022).

Current study

This study aims to further explore the availability of inverse scope interpretations among bilinguals, with a particular emphasis on HSs and, as a comparison group, L2s. Since monolinguals do not have competing grammar systems leading to even a possibility of avoidance to accommodate this, L2s were chosen as a comparison group rather than the standard monolingual comparison. Mandarin-English bilinguals living in the anglophone United States completed an online experiment in English where they heard aural stimuli such as (1) and (2), chose between two images that matched either a surface, inverse, or non-corresponding distractor image, and then rated their selection for appropriateness. The novelty of this study is threefold: 1) study a bilingual population which little is known about in terms of their scope processing in their dominant L2 using 2) an online software that allows for participants to complete the study in a location of their choosing, with 3) bilingual time-apparent parent comparisons as opposed to standard monolingual comparisons. Additionally, I utilize a forced choice task between two images sometimes forcing a choice between an inverse or a non-corresponding distractor image, which was not examined in Scontras et al (2017) or other previous studies. The two research questions are as follows:

1. Do Mandarin heritage speakers activate L2 inverse scope in accordance with the *Processing Scope Economy* principle (measured through RTs, goodness ratings, and picture selection), or do they avoid inverse scope computation in line with the *Avoidance of Ambiguity* strategy?

2. Will bilingual language experience (HL vs L2) differentially impact the ability to compute inverse scope interpretations of doubly-quantified constructions in L2 English?

The non-standard methodology of this study will help to answer these research questions for several reasons. The remote administration of this experiment reduces potential performance effects when being directly studied, and allows for participants that are not the standard college undergraduate participant pool. Showing a pair of pictures rather than just one (as Scontras et al, 2017 did) allows for direct comparisons of picture selection, particularly when the two images are argued to be dispreferred. The goodness measure will still allow a direct exploration of how appropriate each interpretation feels to the participants. Lastly, RTs provide a bit of online evidence for processing difficulty (despite not having implicit measures such as eye-tracking).

To help guide predictions, Table 2 in the Method section provides what the *PSE* and *AA* strategies would predict for the Picture Set independent variable and the three dependent variables of Picture Selection, Goodness Rating, and Response Time (RT). These variables are also further explained in the next section. In line with the proposal of Polinsky and Scontras (2020), I hypothesize that HSs will not activate the inverse scope interpretation any more than a non-corresponding distractor image, supporting the emergent *AA* principle over the well-documented *PSE*. This does not necessarily mean that the inverse scope is unavailable to HSs or that it cannot be pointed out to them, but rather that they prefer to avoid its activation (more so than monolingual comparisons, though this is not directly tested in this study), which would affect selection, goodness ratings, and RTs across the board. This hypothesis would mean that HS results would look similar in measurement to late-acquired L2 bilinguals whose first language is scope rigid. However, while L2s may not have the inverse scope available to them at all, I predict that HSs will have access to the inverse scope but will avoid its activation due to their language background experience. This would be demonstrated with faster RTs from HSs when directly comparing the two groups, and an increased selection of the inverse image when it is presented with a non-corresponding distractor image. Additionally, I predict that HSs will not be affected by which quantifier (*every* or *a*) precedes the other, but that L2s will due to the possible mismatched availability of inverse scope readings in one but not the other structure in Mandarin.

Method

Participants

Fifty-eight Mandarin-English bilingual adults residing in the US were given financial compensation to participate in this study. Participants were recruited through CUNY-internal emails, Prolific, and emails to internal groups at Google. Due to the nature of recruitment and lack of rigid screening, fifteen participants were removed because they either did not complete the entire study or were disqualified based on their background questionnaire responses. There was also an issue with spamming, but luckily this was caught before receiving payment. This left forty-three participants included in analysis. All participants self-reported as fluent in both Mandarin and English, with Mandarin as their first-learned language. Participants were classified as heritage bilinguals (HS, *n* = 21) or late bilinguals (L2, *n* = 22) based on criteria commonly used in HS studies (Benmamoun et al., 2013). HS participants were either born in the US or arrived before age 5, while L2 participants immigrated to the US after age 12 (most in adulthood). Language background information was collected using the Bilingual Language Profile (BLP: Birdsong et al., 2012). Mean participant characteristics by group for four variables of interest, along with significance results from independent samples t-tests, are summarized in Table 1.

Table 1

Demographic information for HS and L2 participants M(SD)

	HS (*n* = 21)	L2 (*n* = 22)	Group difference
Current age	30.14 (5.39)	28.27 (4.56)	*p* = 0.23
Age of Arrival	*n* = 4, 3.75 (1.26) *n* = 17, *US-born*	22.00 (5.25)	*p* < .01
Mandarin comprehension	4.14 (1.11)	5.68 (0.48)	*p* < .01
English comprehension	6.00 (0)	4.54 (0.96)	*p* < .01

The BLP focuses on four modules: language history, current language use, language proficiency, and language attitudes. These four modules compile to output individual relative dominance scores, though the current study does not implement this score since I focus on group comparisons. For the Mandarin and English comprehension questions used in Table 1, participants were asked "How well do you understand Mandarin?" and "How well do you understand English?". Answers range from a 0-6 point Likert scale, with 0 indicating "not at all" and 6 indicating "very well". Follow up studies will do a deeper dive into these BLP questionnaire scores at an individual level, but this is outside the scope of the current study.

Stimuli and Design

The design of this study is 2x3, with the Sentence Type (1, *every>a*) vs. (2, *a>every*) crossed with Picture Set (Surface+Inverse; Surface+Distractor; Inverse+Distractor, Figure 1 & 2). Participants listen to (1) or (2), select one image from the pair, and then rate how appropriate their chosen image is using a 5-point Likert scale. In other words, the three measured variables are Picture Selection, Goodness Rating, and Response Time (RT). While participants are only shown two images per stimulus, there are three possible images dispersed in pairs among three lists. One image corresponds to the surface scope interpretation (S), a second corresponds to the inverse scope interpretation (I), and the third image is a distractor (D) that corresponds to neither (and thus should always be deemed unacceptable). The S+I and S+D picture pairs allow a comparison among individuals in each bilingual group, assessing if the inverse scope image will be activated more than a non-corresponding distractor when the presumed picture selection preference is for the surface image. The third picture pair condition (I+D) allows for a group comparison to see how each group reacts when the two non-preferred images are presented together.

Figure 1

Images used to create pairs for Sentence (1) Every shark is attacking a pirate.

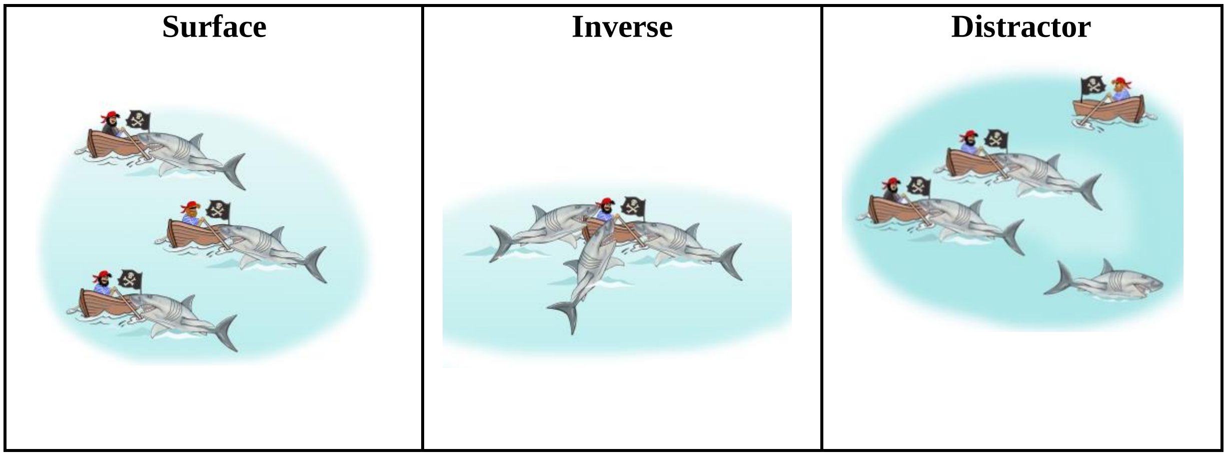
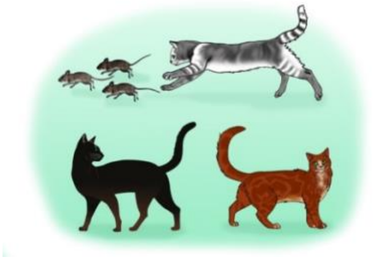
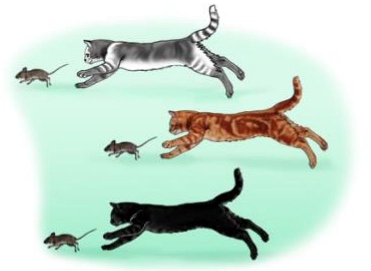
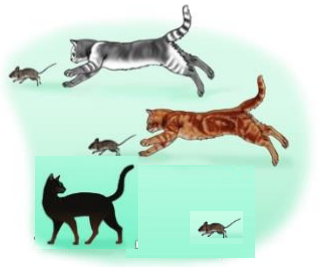


Figure 2

Images used to create pairs for Sentence (2) A cat is chasing every mouse

Surface	Inverse	Distractor
		

There are 12 experimental stimuli items that are counterbalanced among three lists, allowing each stimulus to have each combination of picture sets. There are six items per condition for sentence type, and four item per condition for picture set. In addition, participants answer 12 filler questions for a total of 24 items.

Procedure

Participants completed the experiment at a location of their choosing using PCIbex software (Zehr & Schwarz, 2018). First, participants read and signed a virtual consent form. Next, the online experiment was completed. Participants were told that they would see two pictures and hear a sentence. Their task was to evaluate which image best matches the audio. For example, if the left image matches more, participants would click on the left image. Then, participants were asked to select how appropriate the image is from a scale of 1 (does not match) to 5 (matches). Participants were also told that response time is measured and asked to answer questions as quickly as possible. One practice trial was given, and then participants answered all 24 questions in a single sitting. A catch trial in Mandarin was also included halfway through stimuli presentation to ensure Mandarin proficiency. Lastly, participants completed the BLP to obtain language background information. The duration of the entire experiment was less than 30 minutes.

Data analysis

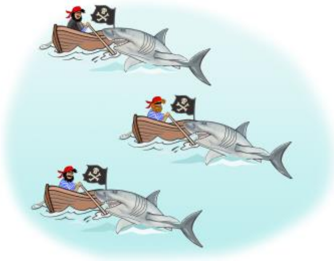
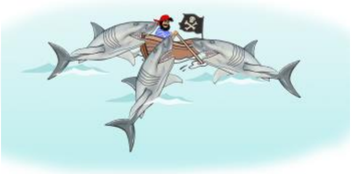
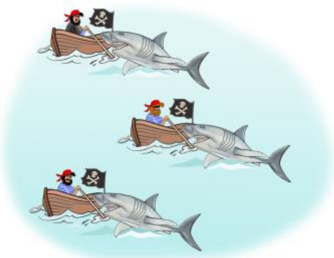
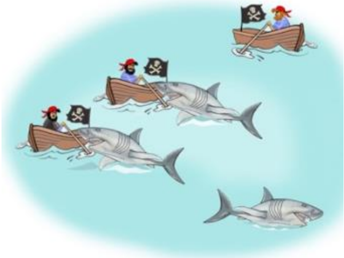
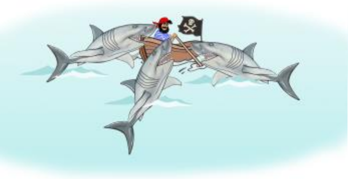
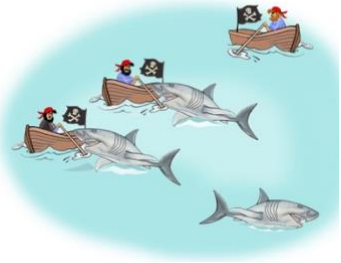
Data were separated by bilingual Group (HS and L2), Picture Set (S+I, I+D, S+D), and Sentence Type (every>a, a>every). Descriptive means and standard deviations were calculated in R for Picture Selection, Goodness Rating, and RT for both the HS and L2 groups. Picture Selection is represented on a 1-5 ordinal scale, where a score of 1 indicates “does not match” and 5 indicates “matches”. Goodness Ratings are depicted as a percentage of selecting the more expected answer. In both the S+I and S+D conditions, this score shows the percentage that participants selected the surface image. For the I+D condition, this score depicts the percentage that participants selected the inverse image. Finally, RT is presented in milliseconds, with outlier times longer than 30,000 milliseconds removed from analysis (less than 1% of the data). Comparisons of the HS vs L2 group are presented with figures that contain each group’s mean and 95% confidence interval bars by condition for each of the three dependent variables. Separate figures are presented for the I+D condition, since this condition directly measures inverse scope interpretations with non-corresponding distractor images while the other two picture sets (S+I, S+D) compare these two in relationship to an assumedly preferred surface image.

For each bilingual group, inferential models were run for each of the three dependent variables (picture selection, goodness, and RT). The I+D condition is removed for by-group modeling, since of interest within each group is the comparison between the S+I and S+D picture sets. Logistic mixed-effects models were run for both picture selection and goodness using the lme4:glmer() function (Bates et al., 2014). Due to ceiling effects, goodness ratings are coded as a binary score of good (a score of 5) and bad (a score of 1-4). Log-transformed RTs were modeled by a linear mixed effects model using the lme4::lmer() function (Bates et al., 2014). All models have sum-coded fixed effects of picture set (S+I and S+D), type of sentence (a>every and every>a), and their interaction. All models also contain by-participant and by-item random slopes when the AIC score is better to include them (Bozdogan, 1987).

For group comparison modeling, picture selection and goodness are modeled with logistic mixed-effects models while log-transformed RTs are modeled with a linear mixed-effects model. Each model has fixed effects of sum-coded bilingual group (HS as +1), sum-coded sentence type (a>every as +1), and their interaction in addition to by-participant and by-item random slopes. For the group comparison modeling, the only picture set used is the I+D condition since this picture set directly compares inverse scope and a non-corresponding distractor, allowing us to compare inverse scope activation across groups.

Table 2

Predictions made by PSE and AA strategies by Picture Set for each dependent variable

Picture Set (every > a condition)	PSE predictions		AA predictions	
SURFACE (S + I) INVERSE  	**Picture Selection**	S picked at high rate I < D	**Picture Selection**	S picked at high rate I = D
SURFACE (S + D) DISTRACTOR	**Goodness Rating**	High acceptability I < D	**Goodness Rating**	Highest acceptability I = D
 	**RT**	Low RTs I > D	**RT**	Low RTs I = D
INVERSE (I + D) DISTRACTOR	**Picture Selection**	I selected more HS > L2	**Picture Selection**	I & D equally likely HS = L2
 	**Goodness Rating**	Medium acceptability HS > L2	**Goodness Rating**	Low acceptability HS = L2
	RT	Slower than S HS < L2	**RT**	Slow RTs HS = L2

Results

Results are first presented for each bilingual group, with means and models for picture selection, goodness rating, and RT among HSs and then L2s. Then, group comparison results are presented with group mean differences and models with bilingual group as a fixed effect for each of the three measured variables. Finally, I provide a brief summary of all findings. Table 2 on the previous page can be used to reference the predictions that each strategy would make for the every>a sentences by Picture Set for all three measured variables.

HSs

Table 2 provides a summary of the predictions that each processing strategy makes by Picture Set. For the HS group, I predicted that these bilinguals would disprefer the inverse image to a similar degree as the non-corresponding distractor image, indicating an avoidance of ambiguity in their dominant language. This could be supported in several ways. First, similar means could be found across all instances of the S+I and S+D measurements. If HSs choose the surface image over both the inverse and the distractor image at a similar percentage rate (Picture Selection) with comparable goodness ratings in about the same amount of time (RT), this would indicate that the inverse interpretation is not activated more than the non-corresponding distractor image when the surface interpretation is available. A lack of activation of the inverse scope over the distractor image can also directly be measured when those are the two images available, in the I+D condition. Longer RTs and lower goodness ratings could indicate a struggle with these stimuli items, which would illustrate that neither image seems accessible for HSs. Their ultimate selection of the inverse image is predicted due to exhaustiveness, but lower selection rates of the inverse image could also be indicators. I do not predict a meaningful difference of sentence type, as this is unrelated to inverse scope activation and there is no evidence suggesting that HSs would interpret these constructions differently depending on which quantifier is higher in the construction.

Descriptives: HSs

Means and standard deviations for the three measured variables of Picture Selection, Goodness Rating, and RT are shown in Table 3. A line separates the S+I and S+D picture sets from the I+D picture set as a reminder that these measure two different concepts. The first two conditions compare inverse and distractor images when the more likely surface interpretation is an available choice, while the third condition forces a choice between these two (presumably) dis-preferred images.

Table 3

Descriptive Statistics: HSs.

Picture Set	every>a: “every shark is attacking a pirate”	a>every: “a cat is chasing every mouse”
SURFACE + INVERSE	Selection: 76.19% Goodness: 4.52 (0.74) RT: 4027 ms (1390)	Selection: 85.71% Goodness: 4.74 (0.59) RT: 3940 ms (1731)
SURFACE + DISTRACTOR	Selection: 97.62% Goodness: 4.76 (0.53) RT: 3778 ms (1215)	Selection: 95.24% Goodness: 4.76 (0.48) RT: 3810 ms (1485)
INVERSE + DISTRACTOR	Selection: 64.71% Goodness: 3.88 (1.15) RT: 5209ms (2502)	Selection: 97.62% Goodness: 3.88 (1.21) RT: 5632ms (2354)

Overall, the surface image is selected at a higher rate in the a>every condition (86% and 95%) than the every> a condition (76% and 98%). The surface image is also selected at a higher numerical rate in the S+D condition (98% and 95%) than in the S+I condition (76% and 86%). When the inverse and distractor images are presented together HSs have a preference for the inverse image, which is higher in the a>every condition (98%) than the every>a condition (65%). Goodness ratings are similar across all conditions where the surface image is presented, with goodness ratings higher than 4.5/5 for all four means. Goodness ratings are lower in the I+D condition (3.88), while there is no difference by sentence type for this picture set. RTs are similar across sentence type for the first two picture sets, and HSs respond slightly faster on average when choosing between the surface and distractor images compared to when choosing between the surface and inverse images. RT averages are overall higher for the D+I picture set (over 5 seconds each compared to around 4 seconds for the other four conditions), and HSs are slightly faster in the every>a sentence type than the a>every sentences.

Inferentials: HSs

Logistic and linear mixed effects models for the three measurements of picture selection, goodness rating, and RT are found in Table 4 and Table 5. These inferential models seek to confirm whether the trends found in the descriptive data can be generalizable to a larger population outside the participant sample. The I+D condition is not included in these models. For the HS modeling, the maximal models for Picture Selection and Goodness Rating included a random slope for item but not participant, while the maximal model for log-transformed RT includes both by-participant and by-item random slopes.

Table 4

Logistic mixed-effects models for HS picture selection and goodness ratings

DV	Fixed Effect	Estimate	Std. Error	z-value	p-value
	(Intercept)	3.23	0.66	4.89	<.001***
Picture	Condition	-0.01	0.38	-0.01	.99
Selection	**Picture Set**	**1.19**	**0.41**	**2.88**	**.004****
	Cond:Pic	-0.42	0.38	-1.11	.27
	(Intercept)	1.87	0.57	3.29	<.001***
Goodness	Condition	0.23	0.24	0.99	.32
Ratings	Picture Set	0.36	0.24	1.48	.14
	Cond:Pic	-0.36	0.24	-1.48	.14

Table 5

Linear mixed-effects model for HS RTs

Fixed Effect	Estimate	Std. Error	t-value	p-value
(Intercept)	8.19	0.06	129.38	<.001***
Condition	-0.02	0.04	-0.62	.55
Picture Set	-0.02	0.03	-0.84	.40
Cond:Pic	0.01	0.03	0.21	.83

Across all three models, the only significant fixed effect is the picture set during the picture selection task. The surface image is selected at a significantly higher rate when presented with a distractor image as compared to a S+I pair where participants chose between the surface and inverse images. No fixed effects or interactions were found in the goodness ratings or RT models.

L2s

Similar to the HS group, I predicted that L2 late bilinguals would disprefer the inverse image to a similar degree as the non-corresponding distractor image. However, whereas HS may disprefer the inverse interpretation due to avoidance of ambiguity, L2s would disprefer the inverse image because of L1 transfer as Mandarin is more or less scope-rigid. While both of these hypothetical causes would lead to similar results through the current measured variables (similar means across all instances of the S+I and S+D measurements or longer RTs and lower goodness ratings in the I+D condition), I predicted the L2 group would be more affected and show even more similarity across all three measures. With the L2 group, a meaningful difference is also predicted to be found based on Sentence Type due to possible L1 transfer (assuming that there is indeed a possibly imbalanced availability of the inverse scope in Mandarin), such that every>a may lead to increased activation of the inverse scope compared to the a>every sentences.

Descriptives: L2s

Means and standard deviations for the three measured variables of picture selection, goodness rating, and RT are shown in Table 6. Once again, a line separates the S+I and S+D picture sets from the I+D picture set as a reminder that these measure two different concepts. The first two conditions compare inverse and distractor images when the more likely surface interpretation is an available choice, while the third condition forces a choice between these two (presumably) dis-preferred images.

Table 6

Descriptive Statistics: L2s

Picture Selection	every>a: *Every shark is attacking a pirate*		a>every: *A cat is chasing every mouse*	
SURFACE + INVERSE	Selection:	88.64%	Selection:	86.36%
	Goodness:	4.73 (0.82)	Goodness:	4.77 (0.60)
	RT:	5068 ms (3691)	RT:	4642 ms (2223)
SURFACE + DISTRACTOR	Selection:	100%	Selection:	95.45%
	Goodness:	4.95 (0.21)	Goodness:	4.73 (0.82)
	RT:	3931*ms* (1777)	RT:	5097*ms* (3586)
INVERSE + DISTRACTOR	Selection:	70.27%	Selection:	84.09%
	Goodness:	3.97 (0.99)	Goodness:	2.07 (1.39)
	RT:	5782*ms* (2709)	RT:	8096*ms* (5474)

Selection of the surface image over the inverse is similar in both sentence type conditions (89% and 86%), but is lower overall than surface image selection when the distractor image is the second choice (100% and 95%). The inverse image is selected over the distractor at a higher rate for the a>every condition (84%) than every>a sentences (70%). Goodness ratings for conditions with an available surface image are similarly high (above 4.7/5 for all four), while they are lower for the I+D picture set. There is a noticeable difference in the goodness rating for the I+D picture set, such that the rating for the a>every sentences is much lower (2.07) than the every>a stimulus items (3.97). RTs are fastest in the every>a and S+D condition (under four seconds), while the other three RT means with a surface image available are more or less comparable. Interestingly, participants responded faster in the a>every sentences than the every>a sentences when presented with a surface and inverse image. In the I+D condition, L2s on average answer much faster during every>a constructions (5.7 seconds) than a>every constructions (8.1 seconds).

Inferentials: L2s

Logistic and linear mixed effects models for the three measurements of picture selection, goodness rating, and RT are found in Table 7 and Table 8. These inferential models seek to confirm whether the trends found in the descriptive data can be generalizable to a larger population outside the participant sample. The I+D condition is not included in these models. For the L2 modeling, the maximal model for Goodness Rating included a random slope for participant but not item, while the maximal model for log-transformed RT includes both by-participant and by-item random slopes. The maximal model for Picture Selection that converges includes both by-item and by-participant random slopes, but with a near singular fit with no by-item variance. It is fairly common for complex mixed-effects models to result in singular fits, which means that one or more dimension of the variance-covariance matrix is estimated to be zero. Singular models are statistically well-defined, but may indicate an overfitted model with poor power (see Barr et al., 2013; Bates et al., 2015; Matuschek et al., 2017).

Table 7

Logistic mixed-effects models for L2 picture selection and goodness ratings

DV	Fixed Effect	Estimate	Std. Error	z-value	p-value
	(Intercept)	9.92	143.74	0.07	.95
Picture	Sent Type	-6.51	143.74	-0.04	.96

Selection	Picture Set	7.13	143.74	0.05	.96
	Type:Pic	-6.25	143.74	-10.04	.97
Goodness Ratings	(Intercept)	2.93	0.60	4.88	<.001***
	Sent Type	-0.37	0.29	-1.27	.21
	Picture Set	0.37	0.29	1.27	.21
	Cond:Pic	-0.35	0.29	-1.21	.23

Table 8

Linear mixed-effects model for L2 RTs

Fixed Effect	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	2.20	0.32	6.81	<.001***
Sent Type	-0.32	0.30	-1.07	.29
Picture Set	0.32	0.26	1.20	.23
Type:Pic	-0.28	0.26	-1.07	.28

There are no significant fixed effects or interactions in either of the generalized linear models for picture selection and goodness rating. No differences are found across picture sets or sentence type. For the linear mixed effects model of log-transformed RT, there are also no fixed effects or interactions between Sentence Type and Picture Set. None of the L2 models reveal any significant predictors, indicating no meaningful difference between the Inverse and Distractor images among any of the three measured variables.

Comparison of HS and L2

This third and final result section differs slightly in format from the previous two. While the first two sections described measurements of Picture Selection, Goodness Rating, and RT at the group level, I now compare these three dependent measurements across the two groups. Predictions for each strategy can be revisited in Table 2. I also predict that if Mandarin has an imbalance of inverse scope depending on Sentence Type, L2s will activate the inverse image more in the every>a condition than the a>every constructions to a larger degree than HSs. Support of this would be shown by an interaction in the inferential models between the two predictor variables of Bilingual Group and Sentence Type.

Descriptives: group comparisons

Figures 3, 4, and 5 show means and 95% confidence interval bars by group for each of the six unique conditions. Once again the I+D picture set is differentiated from the S+I and S+D picture sets to illustrate the difference in what this picture set is measuring. In this group comparison, I focus on the I+D set where there is no surface image available in order to examine whether HSs and L2s differentially respond to this forced choice among two presumably dis-preferred images.

Figure 3

Picture Selection by Group for Surface+Other (left) and Inverse+Distractor (right)

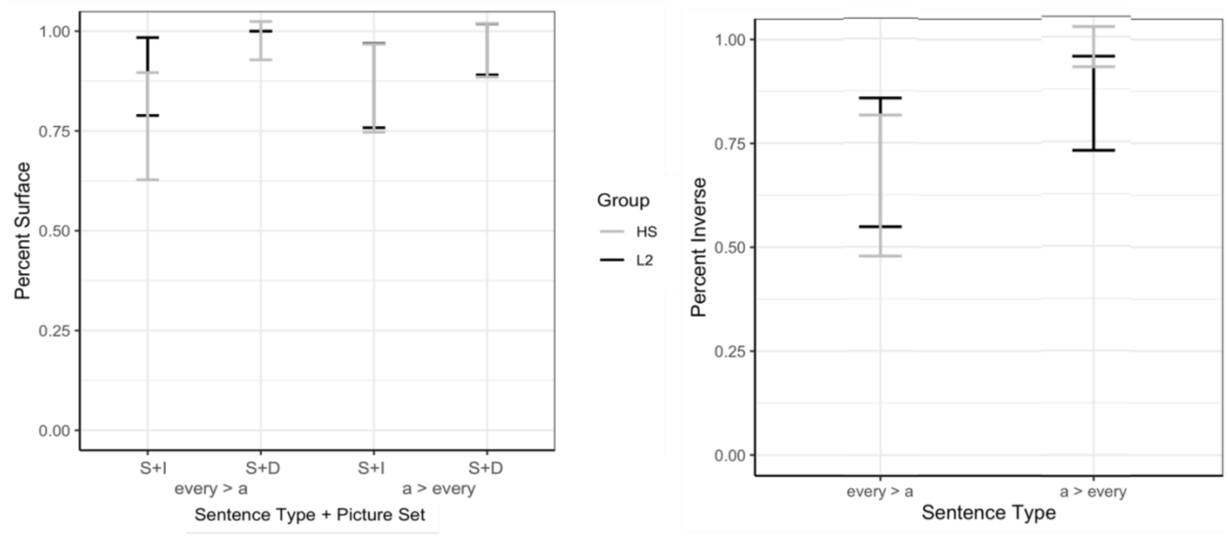


Figure 4

Goodness Ratings by Group for Surface+Other (left) and Inverse+Distractor (right)

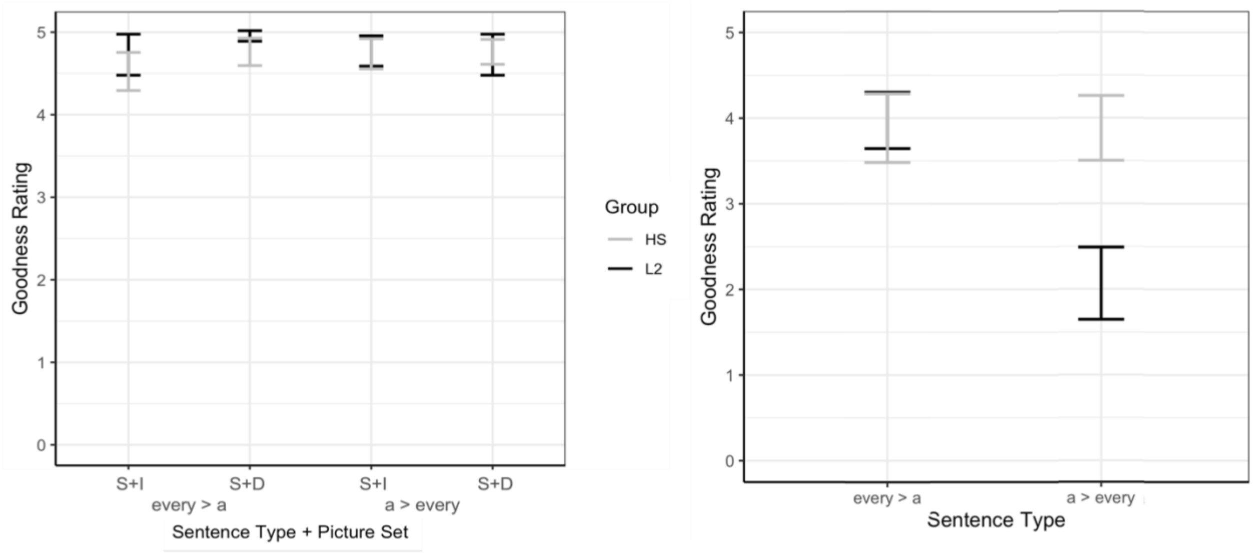
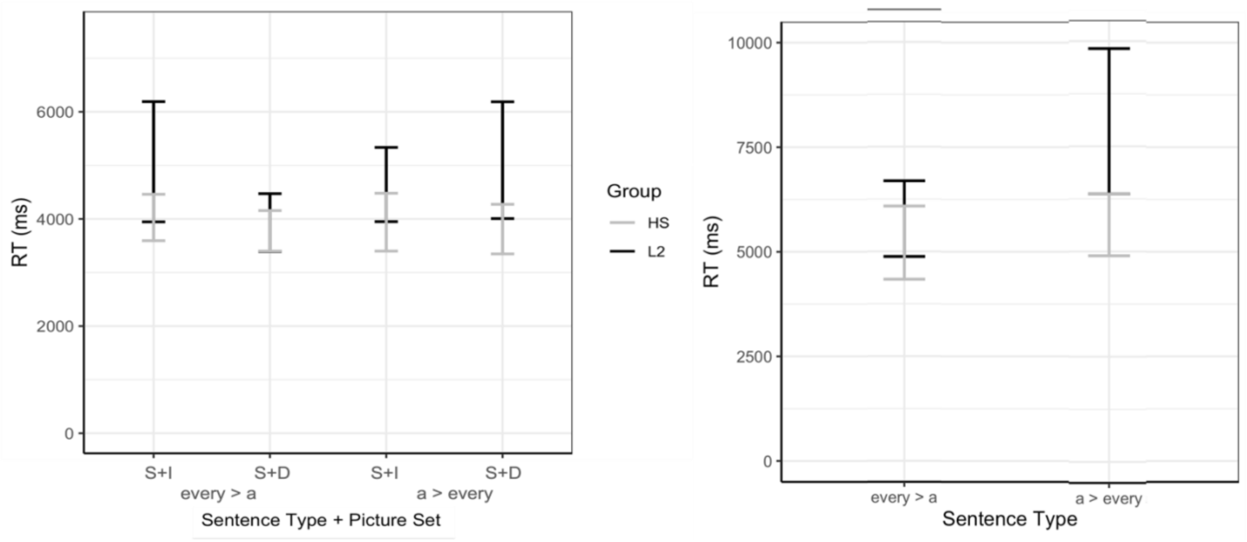


Figure 5

RTs by Group for Surface+Other (left) and Inverse+Distractor (right)



There are no meaningful differences between the HS and L2 groups for any of the three measured variables when a surface image is available. In all four conditions, both the HS group and the L2 group chose the surface image at a similarly high rate. Goodness ratings have ceiling effects across both groups. L2s have higher variance in their RTs as evidenced by the larger CI bars, but these bars overlap with the 95% CIs of the HSs so they do not seem to differ significantly from each other. For the right-hand figures which depict the I+D conditions, a group difference is found in a>every sentences for both goodness rating and RT. While both groups similarly select the inverse image ultimately over the distractor image, L2s rate the a>every sentences lower and take longer to respond than the HSs. The two bilingual groups do not differ in goodness ratings or RTs for D+I picture sets when participants hear an every>a construction.

Inferentials: group comparisons

Logistic and mixed effects models for the three measurements of picture selection, goodness rating, and RT are found in Table 9 and Table 10. These inferential models seek to confirm whether the trends found in the descriptive data can be generalizable to a larger population outside the participant sample. The fixed effects are different for this group comparison, these models look for predictor variables of bilingual Group and Sentence Type. The only picture set used for these models is the I+D condition, since it potentially reveals group-level differences in how participants respond when two dispreferred images are the only ones available.

Table 9

Logistic Mixed-effects Models for Picture Selection and Goodness Ratings

DV	**Fixed Effect**	**Estimate**	**Std. Error**	**z-value**	**p-value**
Picture Selection	(Intercept)	2.17	0.71	3.08	.002**
	Group	**0.69**	**0.34**	**2.06**	**.04***
	Sent Type	1.19	0.70	1.71	.09
	Group:Type	0.60	0.33	1.81	.07
Goodness Ratings	(Intercept)	-1.38	0.33	0.53	.003**
	Group	**0.88**	**0.27**	**1.79**	**.01***
	Sent Type	-0.44	0.32	-2.89	.22
	Group:Type	**0.64**	**0.23**	**3.21**	**.01***

Table 10

Linear Mixed-effects Model for RTs

Fixed Effect	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	8.60	0.07	120.18	<.001***
Group	-0.10	0.06	-1.73	.09
Sent Type	0.09	0.05	1.84	.10
Group:Type	-0.03	0.03	-0.86	.39

For picture selection proportion model, a significant main effect of bilingual Group is found. HSs select the inverse image in the I+D pair at a significantly higher rate than the L2s. The maximal Goodness Ratings model also shows a significant main effect of Group, where HSs rate the image as more appropriate than L2s do. There is also a significant interaction between these two fixed effects, driven primarily by the low ratings given to the a>every stimuli by L2s. No significant effects or interactions are found in the linear mixed effects model for log-transformed RTs.

Summary of findings

Table 11

Summary of findings by strategy prediction

Picture Set	PSE	AA
S + I	Picture Selection: ✓	Picture Selection: ✗
S + D	Goodness Rating: ✗	Goodness Rating: ✓
(measure HS as a group)	RT: ✗	RT: ✓

I + D (compare HS and L2s)	Picture Selection: ✓ Goodness Rating: ✓ RT: ✗	Picture Selection: ✗ Goodness Rating: ✗ RT: ✓

Evidence for which scope strategy HSs employ is mixed. Overall, out of three measures—picture selection, goodness rating, and RTs, HSs only show a significant difference between the inverse and distractor images only in one, namely, picture selection. HSs choose the surface image at a higher rate when presented with a distractor image compared to when presented with an inverse image. They also choose the inverse image over the distractor image in the I+D group at a high rate, but no other measures suggest more activation of the inverse interpretation over a non-exhaustive distractor. These two measures suggest a PSE strategy implementation, but the lack of difference between the inverse image and the distractor across goodness ratings and RTs suggest more of an avoidance of ambiguity strategy, since the inverse image is not being rated higher or answered any faster than a non-corresponding distractor. Meanwhile, L2s treat the inverse and distractor images similarly across all measures. That is, there is no clear support that the L2s are preferring/activating an inverse scope interpretation apart from their direct selection of the inverse image when forced to choose in the I+D condition.

Group comparisons show that HSs and L2s behave differently when it comes to inverse scope activation. HSs choose the inverse image at a higher rate and rate the goodness of the inverse image higher, albeit at a non-significantly different response time, compared to L2s. Interactions are driven by the L2s strong dis-preference for the inverse image in the a>every condition, suggesting that perhaps there is a mismatch in scope rigidity among these two sentence types in Mandarin.

Discussion

In the present study, I have investigated which strategies Mandarin HSs seem to employ when parsing structurally ambiguous doubly quantified constructions in English, probing a possible reduction of ambiguity in the dominant language based on their language experience as HSs. I also studied how late-acquired, highly proficient L2s respond to scope ambiguity in their second-learned language when their L1 Mandarin is scope rigid and does not allow this ambiguity. Lastly, I compared the two groups when forced to choose between an inverse and non-corresponding distractor interpretation to probe the availability of an inverse interpretation when a surface interpretation is not provided. I hypothesized that HSs would avoid ambiguities in their dominant L2, L2s would not have an inverse interpretation available to them, that HSs would respond faster and rate the inverse image better than L2s when presented alongside a non-corresponding distractor image, and lastly that L2s would show a difference of Sentence Type/quantifier position due to possible differences among these structures in their L1 Mandarin.

The by-group analyses revealed that neither HSs nor L2s treat the inverse and distractor images differently from each other when presented a doubly-quantified sentence in English alongside a surface image interpretation, with one exception of HSs selecting the inverse image at a significantly higher rate when presented with a surface image than in the S+D condition. This indicates no evidence of activation of inverse scope among L2s, and some activation but overall dis-preference among HSs. In the group comparisons, on the other hand, results show higher rates of Picture Selection and Goodness Ratings of the inverse image among HSs compared to L2s, suggesting that the inverse interpretation is indeed more available to HSs than it is to late bilingual L2 counterparts. This might be due to L1 transfer for L2s that does not impact HSs as much due to their English dominance.

The first research question, probing whether HSs employ a *Processing Scope Economy (PSE)* or an *Ambiguity Avoidance (AA)* strategy when computing scope, yielded interestingly mixed results. When using conditions similar to those found in Scontras et al., (2017) where participants are presented with a picture set that includes surface and inverse interpretations, the *AA* strategy is supported over the *PSE*. While HSs do select the inverse image at a higher rate than the distractor image, they do not show any differences in goodness ratings or RTs. However, when forced to choose between the inverse and distractor images directly, HS processing strategy aligns more with the *PSE* than the *AA*. This is seen through HSs selecting the inverse image at a higher rate and rating the image appropriateness higher overall than their first-generation L2 counterparts. This suggests that the *PSE* and *AA* can indeed coexist among HSs, but at this point neither can confidently be supported among these highly proficient bilinguals. It is also unclear how the *PSE* interacts with the L2 group, as it could be argued to be a reason why L2s cannot access an inverse scope (because it is more costly) but also why an advanced enough learner should be able to access this interpretation (albeit at a higher cost).

The *AA* strategy relies on the assumption that HS grammars are unique from other bilingual grammars. However, this is not a universally accepted belief. It is possible that differences in parsing structural ambiguity stem from factors such as use, exposure, and proficiency that affect HSs and L2s differentially, or these differences may be more salient at the individual level. Further exploration is crucial to tease apart group and individual effects on ambiguity interpretations.

The second research question reveals novel results that, to my knowledge, previous studies have not empirically tested. When only provided an inverse and distractor image (in other words, a surface reading is not available as a selection), HSs behave differently than L2s. HSs show more activation of the inverse image than L2s do, which suggests that bilingual experience plays a role in scope phenomenon beyond low/high proficiency. Both of these bilingual groups are highly proficient in both Mandarin and English, yet clear group differences arise. This difference could be due to differential processing strategies (i.e., a *PSE* allowing HSs to compute inverse scope at a higher processing cost where it is not available to L2s), to cross-linguistic transfer that differentially affects the L2 group more than the HS group, or to language dominance differences. Further measures and probing should be conducted to explore these possibilities and tease out the reasons behind these differential behaviors.

Another interesting finding of this study is the lack of main effect found between every>a and a>every constructions among L2s and HSs, but a strong interaction between sentence type and bilingual group when comparing the two groups in both Picture Selection and Goodness Rating. Scontras et al (2017) tested both of these constructions with their HSs, and even though they focused their discussion only on a>every constructions they found a higher acceptability rating for the inverse image in every>a sentences than a>every. I did not replicate this finding, as HSs show no effect of sentence type across any of the three DVs. L2s also do not show a differential effect of Sentence Type in the S+I and S+D conditions. However, in the I+D condition, L2s rate the inverse image higher when given an every>a construction than a>every. The evidence of a differential effect of quantifier position among L2s and not HSs warrants further exploration and a deeper dive into this structure in Mandarin.

Lastly, another potential confound in this methodology is the potential effect of priming and possible yes-biases that HSs may employ (see Benmamoun et al., 2013; Laleko & Polinsky, 2013; Orfitelli & Polinsky, 2017) . When forced to choose between two dispreferred interpretations, participants may want an answer to be correct to the extent that they then project correctness onto one of the interpretations. In this case, they may have been primed to select the inverse image simply because it is exhaustive while the distractor image is not, even if the inverse interpretation is not one they agree with. This possible confound is mitigated by the lower goodness ratings given during this condition, but also cannot be removed from consideration. Furthermore, yes-bias has been found among HSs in their heritage language but has never been explored in the dominant L2 so it is likely not a strong factor in the current study.

Implications and future directions

This study is beneficial to academia's grand pursuit of knowledge for a number of reasons. First of all, the implications of this study contribute to a theoretical understanding of how heritage grammars parse ambiguity. Little is known on how HSs treat ambiguity in either of their languages, and more broadly not much is known about how heritage language backgrounds impact the parsing of dominant L2 grammars. Secondly, I explore a relatively novel methodology which allows for participants to take the experiment from any location of their choosing without rigorous screening or conversational back-and-forth. This is helpful for efficiency and timeliness of data collection, and this methodology could also create new processing signatures that have not been found in controlled lab settings. Third, I add a direct forced choice between two dis-preferred interpretations which reveals a new behavior pattern for both bilingual groups. Lastly, this study utilizes a group comparison of two bilingual groups rather than relying on monolingual comparisons, which has been considered good practice in the field of bilingualism (Madsen, 2018). As a field we widely acknowledge that bilinguals are highly heterogeneous as individuals and that they should not be compared to monolinguals, yet we have not sufficiently begun to explore these type of studies.

Data collection is ongoing for this study. Next steps include an implicit, eye-tracking study in both English and Mandarin. This will allow a comparison of L2 processing with scope ambiguity parsing in the heritage language. Furthermore, future analyses will utilize a continuous scale of bilingualism, allowing us to capture the gradience of the bilingual experience and to compare a bilingual individual across both of their languages rather than relying on categorizing individuals into buckets of bilingual groups that are well-attested to be heterogeneous (Luk & Rothman, 2022). This individual analysis will allow an exploration of the roles of factors such as proficiency and use in scopal interpretations among bilinguals without a need to chunk individuals into bilingual groups. Lastly, a deeper dive is needed into the role of quantifier order (every>a, a>every) to better understand bilingual behaviors with doubly quantified structures.

References

Anderson, C. (2004). *The structure and real-time comprehension of quantifier scope ambiguity* [PhD Thesis]. Northwestern University Evanston, IL.

Aoun, J., & Li, Y. A. (1993). *Syntax of scope* (Vol. 21). Mit Press.

Asian American Federation. (2019). *Profile of New York City's Chinese Americans*. Data derived from analysis by the Asian American Federation Census Information Center. aafederation.org

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure in mixed-effects models: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *ArXiv Preprint ArXiv:1506.04967*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv: 1406.5823*.

Benmamoun, E., Montrul, S., & Polinsky, M. (2013). Heritage languages and their speakers: Opportunities and challenges for linguistics. *Theoretical Linguistics*, *39*, 129–181.

Birdsong, D., Gertken, L. M., & Amengual, M. (2012). Bilingual language profile: An easy-to-use instrument to assess bilingualism. *COERLL, University of Texas at Austin*.

Bornkessel, I. D., Fiebach, C. J., & Friederici, A. D. (2004). On the cost of syntactic ambiguity in human language comprehension: An individual differences approach. *Cognitive Brain Research*, *21*(1), 11–21.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Chu, C.-Y., Gabriele, A., & Minai, U. (2014). Acquisition of quantifier scope interpretation by Chinese-speaking learners of English. *Selected Proceedings of the 5th Conference on Generative Approaches to Language Acquisition North America (GALANA 2012)*, 157–168.

Dussias, P. E. (2003). Syntactic ambiguity resolution in L2 learners: Some effects of bilinguality on L1 and L2 processing strategies. *Studies in Second Language Acquisition*, 529–557.

Filik, R., Paterson, K. B., & Liversedge, S. P. (2004). Processing doubly quantified sentences: Evidence from eye movements. *Psychonomic Bulletin & Review*, *11*(5), 953–959.

Fodor, J., Bever, A., & Garrett, T. G. (1974). *The psychology of language: An introduction to psycholinguistics and generative grammar*.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210.

Frenck-Mestre, C., & Pynte, J. (1997). Syntactic ambiguity resolution while reading in second and native languages. *The Quarterly Journal of Experimental Psychology A*, *50*(1), 119–148.

Hopp, H. (2010). Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. *Lingua*, *120*(4), 901–931.

Huang, C. J. (1982). Move WH in a language without WH movement. *The Linguistic Review*, *1*(4), 369–416.

Kratzer, A., & Heim, I. (1998). *Semantics in generative grammar* (Vol. 1185). Blackwell Oxford.

Kupisch, T. (2012). Specific and generic subjects in the Italian of German–Italian simultaneous bilinguals and L2 learners. *Bilingualism: Language and Cognition*, *15*(4), 736–756.

Kurtzman, H. S., & MacDonald, M. C. (1993). Resolution of quantifier scope ambiguities. *Cognition*, *48*(3), 243–279.

Laleko, O., & Polinsky, M. (2013). Marking topic or marking case: A comparative investigation of heritage Japanese and heritage Korean. *Heritage Language Journal*, *10*(2), 178–202.

Laleko, O., & Polinsky, M. (2016). Between syntax and discourse: Topic and case marking in heritage speakers and L2 learners of Japanese and Korean. *Linguistic Approaches to Bilingualism*, *6*(4), 396–439.

Laleko, O., & Polinsky, M. (2017). Silence is difficult: On missing elements in bilingual grammars. *Zeitschrift Für Sprachwissenschaft*, *36*(1), 135–163.

Lidz, J., & Musolino, J. (2002). Children's command of quantification. *Cognition*, *84*(2), 113–154.

Luk, G., & Rothman, J. (2022). Experience-based individual differences modulate language, mind and brain outcomes in multilinguals. *Brain and Language*, *228*, 1–4.

MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology*, *24*(1), 56–98.

Madsen, C. N., II. (2018). *De-centering the monolingual: A psychophysiological study of heritage speaker language processing* [PhD Thesis]. The Graduate Center, CUNY.

Marsden, H. (2009). Distributive quantifier scope in English-Japanese and Korean-Japanese interlanguage. *Language Acquisition*, *16*(3), 135–177.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

May, R. C. (1978). *The grammar of quantification.* [PhD Thesis]. Massachusetts Institute of Technology.

Montrul, S., & Ionin, T. (2010). Transfer effects in the interpretation of definite articles by Spanish heritage speakers. *Bilingualism: Language and Cognition*, *13*(4), 449–473.

Nagano, T. (2015). Demographics of adult heritage language speakers in the United States: Differences by region and language and their implications. *The Modern Language Journal*, *99*(4), 771–792.

Orfitelli, R., & Polinsky, M. (2017). When performance masquerades as comprehension: Grammaticality judgments in experiments with non-native speakers. In *Quantitative approaches to the Russian language* (pp. 197–214). Routledge.

Paterson, K. B., Filik, R., & Liversedge, S. P. (2008). Competition during the processing of quantifier scope ambiguities: Evidence from eye movements during reading. *Quarterly Journal of Experimental Psychology*, *61*(3), 459–473.

Polinsky, M., & Scontras, G. (2020). Understanding heritage languages. *Bilingualism: Language and Cognition*, *23*, 4–20. https://doi.org/10.1017/S1366728919000245

Scontras, G., Polinsky, M., Tsai, C.-Y. E., & Mai, K. (2017). Cross-linguistic scope ambiguity: When two systems meet. *Glossa: A Journal of General Linguistics*, *2*(1), 1–28.

Wu, H., Larson, R., Liu, Y., Liu, L., & Mar, G. (2017). Rethinking quantifier scope in Mandarin. *Poster Presented at the 48th Annual Meeting of the North East Linguistic Society (NELS 48), University of Iceland, Reykjavík, Iceland*, 27–29.

Wu, M.-J., & Ionin, T. (2019). L1-Mandarin L2-English speakers' acquisition of English universal quantifier-negation scope. *Proceedings of the 43rd Annual Boston University Conference on Language Development*, 716–729.

Wu, M.-J., & Ionin, T. (2022). L1-Mandarin L2-English learners' acquisition of English double-quantifier scope. *Generative SLA in the Age of Minimalism: Features, Interfaces, and beyond. Selected Proceedings of the 15th Generative Approaches to Second Language Acquisition Conference*, *67*, 93.

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832

Zhou, P., & Gao, L. (2009). Scope processing in Chinese. *Journal of Psycholinguistic Research*, *38*(1), 11–24.