SUPPLEMENTARY MATERIALS

Rating Study

The purpose of the rating study was to assess whether our selection of items for the target (G+S-) and the three competitor categories (G+S+, G-S+, G-S-) was appropriate. In particular, the rating study served to answer the following question: Are S+ items (G+S+ and G-S+ competitors) rated as better semantic fits for their respective classifier classes than S- items (G+S- targets, G-S- competitors, and G-S- distractors)? Since previous research on grammatical gender has suggested that knowledge of noun class membership can affect (native) speakers' judgment of the semantic properties of objects (Boroditsky, Schmidt, & Phillips, 2003), the rating study was conducted with speakers who had no knowledge of Chinese.

Participants

Thirty-two members of the University of Hawai‘i student community completed the rating questionnaire in exchange for partial course credit. The criterion for inclusion was no knowledge of Chinese; one participant was excluded for not meeting this criterion. Data from the remaining 31 participants (mean age: 21 years; 17 females) were included in the analysis. Most ($N = 25$) identified as native speakers of English; the remaining 6 listed Italian, Japanese, Korean ($N = 2$), Spanish, or Tagalog as their native language.

Materials and Procedure

All 60 images that served as targets (k = 12), competitors (G-S-: k = 12; G-S+: k = 12; G+S+: k = 12), or unrelated distractors (k = 12) in the visual-world experiment were included in the rating questionnaire. For each image, participants were asked to (i) label the object, and (ii) rate on a scale of 0-4 to what extent they felt given descriptors applied to this type of object. A sample item is provided in Figure A1. Each image appeared with the descriptors associated with the classifier in the linguistic stimulus with which the image appeared in the main visual-world experiment. For example, *wristwatch* acted as a G-S+ competitor for a target from the *tiáo* class; thus the descriptors for which it was to be rated were the semantic attributes associated with the classifier *tiáo*. Descriptors, listed in Table A1, were taken from Gao and Malt (2009, Appendix A).

72. **Label the object.** *

73. **To what extent do you feel that the following descriptions apply to this type of object?** *

(0 = not at all; 4 = perfectly)
Mark only one oval per row.

	0	1	2	3	4
slender	⬭	⬭	⬭	⬭	⬭
long-shaped	⬭	⬭	⬭	⬭	⬭
flexible	⬭	⬭	⬭	⬭	⬭

Figure A1. Sample item from rating questionnaire.

Table A1. *Semantic attributes associated with classifiers.*

Classifier	Meaning associated with the classifier, as listed in Gao & Malt (2009)	Descriptors listed in rating questionnaire
tiáo	"a slender, long-shape thing, often flexible"	- slender - long-shaped - flexible
zhāng	"to spread open/flat"	- spread-open - has a flat surface
zhī	"a stick-like long thing"	- stick-like - long

Analysis and Results

Each participant's ratings for each item were averaged over the 2 (for *zhāng* and *zhī*) or 3 (for *tiáo*) individual responses provided. For example, if a participant rated *wristwatch* as 2 for "slender", 3 for "long-shaped", and 2 for "flexible", the rating for this participant-item combination that entered the analysis was 2.33.

Table A2 shows the mean ratings for each item type. These descriptive statistics show that, on average, the two S+ items were rated the highest (G+S+: 2.99, G-S+: 2.70), the two G-S- item types the lowest (competitors: .82, distractors: .71), with ratings for the G+S- target nouns (1.79) falling in between.

Table A2. *Mean ratings by item type*

Item Type		*Mean (SD)* rating
G+S-	(target)	1.79 (1.19)
G+S+	(competitor)	2.99 (1.02)
G-S+	(competitor)	2.70 (1.07)
G-S-	(competitor)	0.82 (1.07)
G-S-	(distractor)	0.71 (0.86)

In order to test for differences between item types, ratings were entered into a linear mixed-effects model, using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) in R (R Core Team, 2018), with ItemType as a fixed effect, and participant and item as random effects, using the maximal random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013). The 5-level ItemType variable was treatment-coded with G-S+ as the reference level in order to test for the contrasts of key interest. Table A3 shows the fixed effect statistics from the model output.

Table A3. *Fixed effects from the linear mixed-effect model of item ratings*

	Estimate	*SE*	*t*	*p*
(Intercept)	2.70	0.20	13.29	<.001 ***
ItemType G+S-	-0.91	0.24	-3.78	<.001 ***
ItemType G-S-	-1.88	0.25	-7.49	<.001 ***
ItemType G+S+	0.29	0.24	1.21	0.23
ItemTypeDistractor	-1.99	0.25	-7.85	<.001 ***

Note: Formula: rating ~ ItemType + (1+ItemType | participant) + (1 | item))

The model output shows significant negative estimates for all three S- items types (G-S- competitors and distractors, as well as G+S- targets), indicating that they were all rated as poorer semantic fits than the G-S+ competitors. Meanwhile, the difference between G+S+ and G-S+ items was not significant. These results support the key assumption underlying the Visual World experiment, namely that (G-/G+)S+ items are better aligned with the semantics of the classifier than (G-/G+)S- items.

Given the higher average ratings for G+S- targets viz-a-viz G-S- competitors and distractors, the model was rerun with different reference levels for ItemType in order to further explore differences among S- items, which were not predicted. The output from these additional models showed no significant differences between G-S- competitors and distractors (b = .11, p = .6). However, both were rated significantly lower than the G+S- targets (bs > .9, ps < .001). We interpret these results as a reflection of the fact that the G+S- target items are NON-PROTOTYPICAL, rather than entirely accidental, members of their classifier class. In terms of the interpretation of eye gaze patterns in the visual world experiment, these findings indicate that a preference to look at G+S- targets over G-S- competitors or distractors may not be driven purely by formal knowledge of class-membership, but could be supported, at least partially, by semantics. It is thus only a preference to look at G+S- targets over G-S+ competitors that can provide unequivocable evidence that listeners prioritize knowledge of class membership over classifier semantics during online comprehension.

References

Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax and semantics. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 61-79). Cambridge, MA: MIT Press.

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

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

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