Pragmatic Inference of Intended Referents from Binomial Word Order

Anna A. Ivanova and **Roger P. Levy**
{annaiv, rplevy} @mit.edu
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

How does listeners' perceptual bias influence their interpretation of an ambiguous multiword utterance? We address this question by investigating the relationship between word order in a binomial (an expression of type "A and B") and visual properties of image pairs serving as its potential referents. We found that listeners' choices were strongly influenced by iconicity and relative salience of images within the pair: participants preferred referents where the first mentioned image was located on the left, as well as pairs where the first image was larger than the second image. The effect of image order tended to be stronger than the influence of image size, and both were modulated by participants' general visual field preferences (determined in a separate experimental condition). We further show that binomial phrase interpretation can be simulated by a Rational Speech Act model that includes both word order effects and utterance-independent preferences of the participants.

Keywords: pragmatics; binomials; reference resolution; object salience; Rational Speech Act theory

When interpreting a referring expression, listeners take into account not only its literal meaning, but also various pragmatic considerations, such as the speaker's likely communicative goals, alternative utterances, and common ground information. For example, in the presence of referential ambiguity, listeners tend to favor interpretations of reference to perceptually salient objects (Clark, Schreuder, & Buttrick, 1983). This inference might stem from a pragmatic inference that a salient object has a high chance of being in the "common ground" between the listener and the speaker, or simply from reasoning based on egocentric perspective (Keysar, Barr, Balin, & Brauner, 2000).

For reference beyond the level of the single word, word order might provide an additional source of information to listeners. Word order often reflects object salience: in spoken dialog tasks where visual scenes are in common ground, larger and more centrally located referents tend to be mentioned earlier overall in utterances, as shown by Elsner, Rohde, and Clarke (2014); Clarke, Elsner, and Rohde (2015). These authors also show that visual search follows order of mention and can facilitate performance in collaborative target-search tasks. It is less clear, however, to what extent listeners use word order as a source of information for resolving referential ambiguity. Such inferences might involve pragmatic reasoning about utterances that a speaker might have chosen but did not. While multiple studies have shown that listeners often perform this type of ad-hoc scalar pragmatic inference (e.g., Frank & Goodman, 2012; Bergen, Goodman, & Levy, 2012; Degen, Franke, & Jäger, 2013), inference on the basis of word order in multiword referring expressions has not yet been investigated.

In this study, we use binomials (multiword expressions of type "A and B", such as "salt and pepper"/"pepper and salt") to investigate listener sensitivity to word order choice in resolving referential ambiguity. It is well established that speakers' choice of binomial word order is highly sensitive to multiple linguistic and semantic constraints (McDonald, Bock, & Kelly, 1993; Benor & Levy, 2006). These constraints include perceptual markedness, according to which referents that are more salient in any of a number of manners are referred to first (e.g., proximity: "here and abroad" over "abroad and here"; Mayerthaler, 1988), and iconic sequencing, according to which referents that are conceptualized as being in a sequence are mentioned in that sequence (e.g., "takeoff and landing" over "landing and takeoff"; Cooper & Ross, 1975). In a controlled experimental setting, Gleitman, January, Nappa, and Trueswell (2007) showed that the relative positioning of two objects within an image (left vs. right) affected binomial order in speaker descriptions of these images, and so did an unconscious attention manipulation procedure. In comprehension, while it has previously been demonstrated that listeners are sensitive to violations of preferred binomial orderings in the absence of visual context (Siyanova-Chanturia, Conklin, & van Heuven, 2011; Morgan & Levy, 2016), our study is the first to our knowledge to evaluate whether and how listeners use binomial word order to resolve ambiguity among potential visual referents.

In order to test whether word order affects reference resolution in an ambiguous setting, we presented participants with a binomial phrase of the form "A and B", followed by three pairs of images depicting A and B. We manipulated iconic sequencing by varying relative location of images within the pair (left side of the screen vs. right side of the screen), as well as perceptual salience by varying their size (large vs. small). Participants then picked the image pair that, in their opinion, best matched the binomial. We hypothesize that, if word order within the binomial has no communicative effect, it would not elicit any ordering or size effects beyond listeners' general visual field preferences. However, if word order affects listeners' decisions, we would expect the referent image matching the first word of the binomial to be located on the left (reflecting an iconicity preference) and/or to be larger than the other referent image (reflecting a saliency preference). Additionally, we quantitatively model our data using several variants of the Bayesian Rational Speech Act model (Frank & Goodman, 2012), to take into account the effect of prior expectations on referent selection and to evaluate our overall ability to quantitatively predict comprehender choices[1].

[1]The data and code used for this study can be found at https://github.com/neuranna/binomial-referents

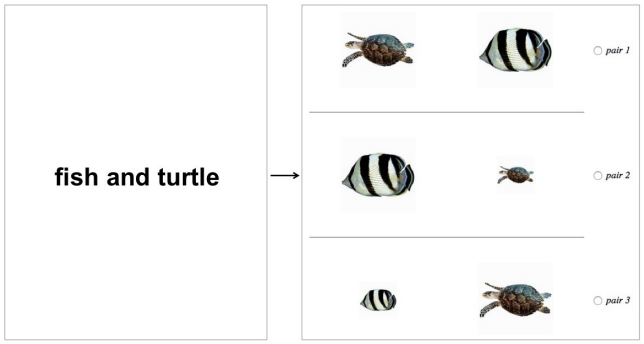

Figure 1: Setup of a single trial. Participants are presented with a binomial, followed by three sets of image pairs. Image pairs differ in relative order, size, or both.

Experiment

Method

Participants We recruited 106 participants through Amazon.com's Mechanical Turk. All workers were located in the US and had at least 95 percent approval rate. The UniqueTurker script[2] was used to ensure that participants completed the experiment only once. All participants were self-reported native speakers of English.

Stimuli We used the Pool of Pairs of Related Objects (POPORO) image set (Kovalenko, Chaumon, & Busch, 2012) as the source of our image stimuli. We selected pairs of images from the dataset based on the following criteria: (1) both images can be described by count nouns; (2) both images depict concrete objects; (3) the two objects have comparable sizes. We normed the images using a separate set of participants ($N = 36$). Each participant was presented with an image of an object and asked to provide a one-word description. We restricted the final set of image stimuli to those that were identically labeled by at least 80 percent of the participants.

Design During each trial, a participant was presented with a binomial ("A and B"), followed by three possible referent pairs. Referent pairs varied by image location (A on the left, B on the right, or vice versa), position on the screen (top, center, bottom), and relative size (both large, small A and large B, small B and large A). Word order within the binomial was counterbalanced across participants. Figure 1 shows a sample trial. In addition, in 20 percent of the trials the binomial was replaced by a letter mask. Those trials were later used to estimate baseline preferences of the participants. We refer to these trials as the *prior elicitation* condition.

Procedure The experiment was hosted online on Ibex Farm (Drummond, 2013). Each participant completed 36 trials, one for each image pair. A trial consisted of a binomial phrase presentation lasting for 1000 ms, a 500 ms pause, and then a screen with three image pairs that served as possible referents.

[2]https://uniqueturker.myleott.com

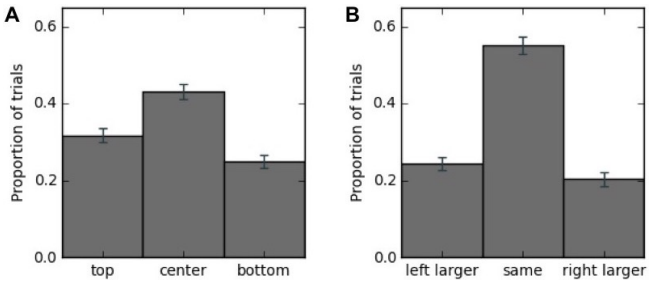

Figure 2: Position (A) and size (B) preferences of participants during the prior elicitation task.

Participants had to select a pair that best matched the referring phrase. Item presentation order was randomized.

Analysis We assessed evidence for prior and word order-based preferences through pairwise comparisons between response types using mixed logit regression analysis using `R`'s `lme4` (Bates, Mächler, Bolker, & Walker, 2015), in all cases testing the null hypothesis that two response types are equally preferred. This corresponds to testing the fixed-effects intercept (which we nevertheless refer to as β for familiarity) in a model that includes by-participants and by-items random intercepts. Statistical significance was determined using the likelihood ratio test.

Results

Prior Preferences We evaluated binomial-independent preferences of the participants' using data from the prior elicitation trials. We found that both screen position and relative image size affected people's choices (Figure 2). Specifically, they were more likely to choose the image pair in the center over the image pair at the top ($\beta = .31, p = .002$), but also more likely to select the pair at the top than the pair at the bottom ($\beta = .25, p = .019$). In addition, participants strongly preferred pairs with images of equal size over pairs where the left image was larger ($\beta = .86, p < .001$). There was no preference for image pairs with larger image on the left vs. pairs with a larger image on the right ($\beta = .21, p = .115$).

Effects of Word Order on the Choice of Referent There was a robust relationship between word order and image order: participants preferred to choose referent pairs where the image mentioned first was located on the left ($\beta = 1.88, p < .001$). Word order also affected size preferences: participants were significantly more likely to choose an image pair where the object mentioned first was larger over an image pair where the object mentioned first was smaller ($\beta = .35, p < .001$). Furthermore, they preferred an equal-sized image pair over image pairs of unequal size ($\beta = 1.17, p < 0.001$), as they did in the prior elicitation condition. Figure 3 shows the effects of order and size on participants' decisions for illustrative reference pair sets. We conclude that, when deciding on an image pair matching a binomial expression, participants tend to pick image pairs that satisfied iconicity and percep-

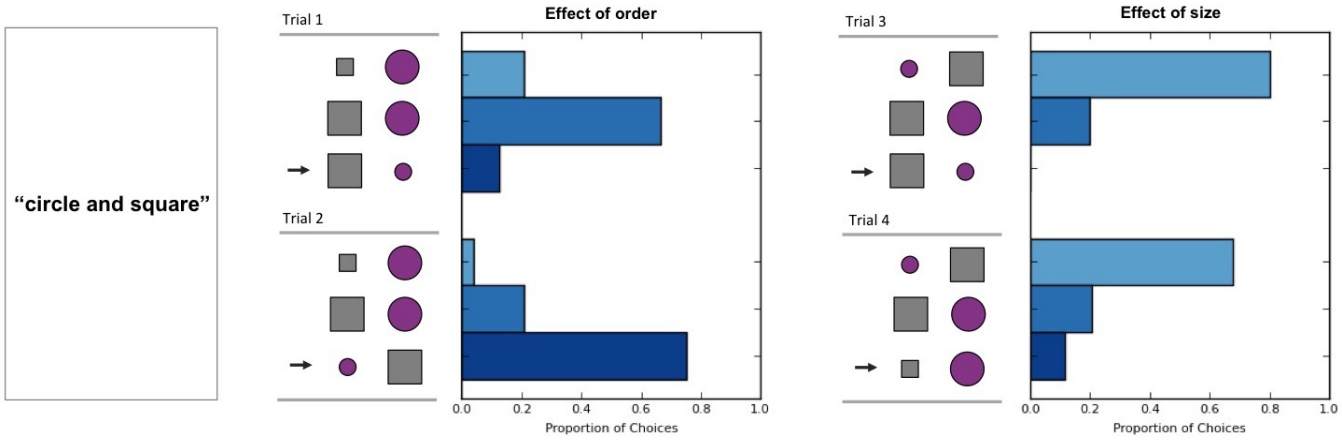

Figure 3: Sample trials demonstrating effects of image order (trials 1 and 2) and image size (trials 3 and 4) on participants' choices. The referring expression in all four trials has the form "circle and square" (actual items varied from trial to trial). Image sets in trials 1 and 2 differ only in relative order of items in the last row (marked by an arrow), yet this manipulation drastically changes the distribution of responses among participants. Similarly, image sets in trials 3 and 4 differ only in relative sizes of items in the third row. This manipulation does not have such a drastic effect, yet increasing salience of the circle by making it larger does increase the frequency with which participants chose this option.

tual salience constraints.

Discussion

We show that, when presented with a binomial expression, listeners rely on information about word order within that expression in order to identify the best matching referent. Specifically, participants in our experiment tended to choose referents in which image order matched binomial word order and in which the image mentioned first was as large or larger than the image mentioned second. Our finding indicates that not only does object ordering affect word order during binomial production (as shown by Gleitman et al., 2007), but it can also be recovered during comprehension. We also identify a separate factor affecting referent selection, namely, relative image size.

In order to further explore the relationship between word order and relative image salience, we simulated the reference resolution process using several computational models and compared their predictions with behavioral results described above.

The Model

Model Specification

In order to model the listeners' inference process, we use the Rational Speech Act (RSA) framework, first proposed by Frank and Goodman (2012). The fully specified model incorporates several levels of inference: the literal listener (L_0), who interprets the utterance according to its literal meaning, as well as the listener's prior expectations about possible states of the world; the speaker (S_1), who considers both the literal listener's interpretation and the relative cost of the utterance; and the pragmatic listener (L_1), who estimates the probability of a certain state of the world based on the utterance that the speaker chose to use. Traditional RSA models evaluate the goodness of fit between an utterance and a reference state using a categorical meaning function (the utterance either matches the referent or not). However, this approach does not apply in our case, since the binomial utterance can in principle refer to all three image pairs presented during a trial. Instead, we can model the effect of binomial-referent mismatch in two ways: (a) by replacing a categorical meaning function with an acceptability judgment that reflects the tendency for a more salient object to be mentioned first (the non-scalar model); (b) by introducing a production cost at the level of the speaker (the scalar model). We provide a formal description of the models below.

Defining the Cost Function The cost function evaluates the goodness of fit between a binomial and an image pair. The two image properties that it takes into account are: (a) relative location of images, and (b) relative size of images. We make the assumption that, if the binomial has the form "A and B", the image depicting A is likely to be on the left and the image depicting B is likely to be on the right. If that is not the case, a penalty is added to the cost function. Similarly, for the binomial of the form "A and B", a penalty will be imposed if the image depicting A is smaller than the image depicting B. This can be expressed as:

$$C(u|s) = \mu * [\![m]\!](u,s) + \sigma * [\![z]\!](u,s) \quad (1)$$

where μ is the order match parameter, $[\![m]\!](u,s)$ is the mismatch function that equals 0 if the word order matches the image order and 1 otherwise, σ is the size parameter, and $[\![z]\!](u,s)$ is the size mismatch function that equals 1 if the first mentioned image is smaller than the second image, and 0 oth-

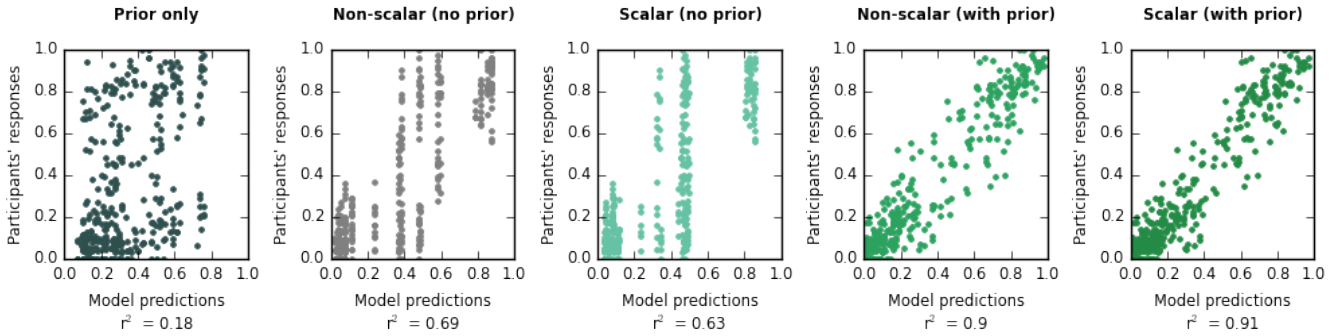

Figure 4: Comparison between model predictions and participants' choices. Each dot is an option from an image set with a particular configuration. The x axis shows the predicted probability of choosing this option, and the y axis reflects the proportion of participants who chose this option during the experiment.

erwise[3].

Prior-only Model: no effect of word order If the word order has no effect on referent selection, listeners' choices depend only on their prior preferences:

$$P_{L_0}(s|u) = P(s) \tag{2}$$

where s is the state (here, an image pair), u is the utterance (here, a binomial expression), and $P(s)$ is the prior probability of the state.

Non-scalar Model: effect at the Level of L_0 The simplest way to introduce the effect of binomial word order is by using only one level of inference (literal listener, L_0) and incorporating word order effects as an exponentiated cost function. Specifically,

$$P_{L_0}(s|u) \propto P(s) * \exp(-C(u|s)) \tag{3}$$

where s is a state (here, an image pair), u is the utterance, and $C(u|s)$ is the cost function that specifies the relationship between word order within the utterance and image pair properties. This model does not include any inference about the speaker's behavior.

Scalar Model: effect at the level of S_1 An alternative way of incorporating the cost term places it at the level of the speaker (S_1) instead of L_0. In this case, the literal listener would assign probabilities to object pairs based solely on prior preferences, without taking word order into account:

$$P_{L_0}(s|u) \propto P(s) \tag{4}$$

The speaker, however, would base their choice of utterance not only on the literal listener's interpretation, but also on the production cost of utterance given the image pair ($C(u|s)$):

$$\begin{aligned} P_{S_1}(u|s) &\propto \exp(\log(P_{L_0}(s|u)) - C(u|s)) \\ &= \exp(\log(P(s)) - C(u|s)) \\ &= \exp(\log(P(s))) \cdot \exp(-C(u|s)) \end{aligned} \tag{5}$$

where $P_{L_0}(s|u)$ is the literal listener's inference function from eq. 3. Since $P(s)$ does not depend on u, we can drop the first exponent, which leaves:

$$P_{S_1}(u|s) \propto \exp(-C(u|s)) \tag{6}$$

meaning that the only factor that the speaker is considering is production cost. The pragmatic listener would then pick a referent by inferring the speaker's perspective:

$$P_{L_1}(s|u) \propto P_{S_1}(u|s) * P(s) \tag{7}$$

Implementation We estimated prior preferences of the participants by reanalyzing data from the prior elicitation condition using the mlogit package in R (Croissant, 2012). Specifically, we constructed the model that estimated coefficients for position preference (top, center, and bottom) and relative size preference (equal, left larger, right larger). We then used these coefficients to predict prior probabilities of choosing each option ($P(s)$). The models themselves were implemented in WebPPL (Goodman & Stuhlmüller, 2014) using starter code provided by Scontras and Tessler (2017).

Comparing Behavioral Data with Model Predictions

When comparing model predictions with participants' responses, we aimed to answer three questions: (1) Does binomial word order affect relative size and order of images in a preferred image pair? (2) Do prior preferences play a role in referent selection? (3) Do order and size biases arise from listener's built-in preferences or from online inference about the speaker's production costs? In order to address the first question, we compared the performance of the baseline prior-only model with the scalar and the non-scalar models. To answer the second question, we evaluated two different versions

[3] We also considered a version of the cost function where having two images of equal size would result in higher cost than having a large image A and a small image B. However, exploratory analyses showed that this version of the cost function does not fit the data as well.

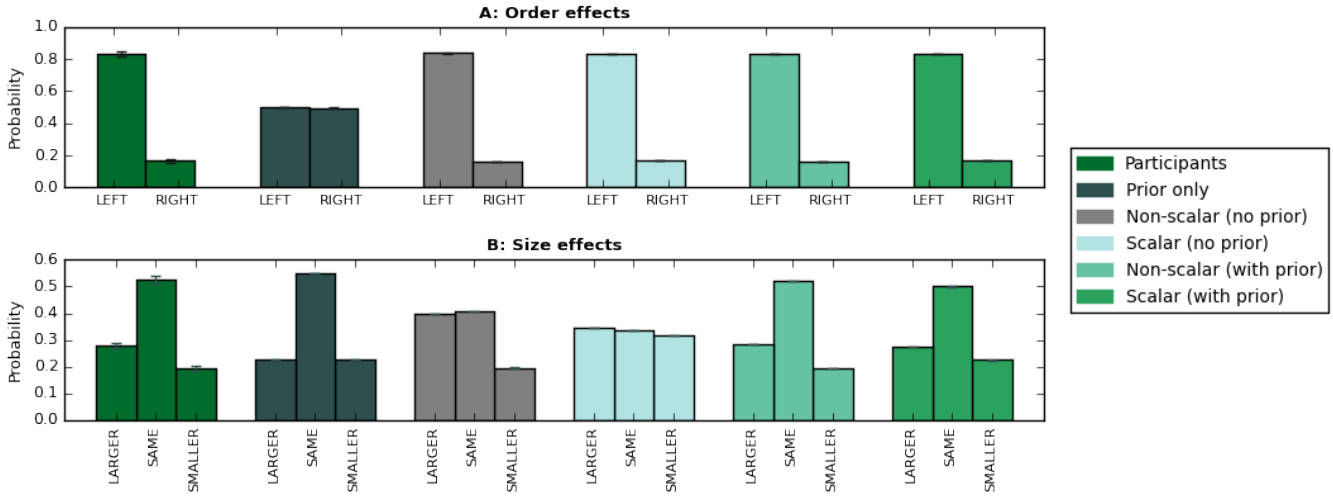

Figure 5: Comparison between behavioral data and model predictions. A: Relative location of the image mentioned first. B: Relative size of the image mentioned first (compared to the other image).

Table 1: Fitted model parameters. Parameter value is estimated as the median of the posterior distribution; values within square brackets are 95% confidence intervals.

Model	Parameters	
	order (μ)	size (σ)
Non-scalar, no prior	2.25 [2.14 - 2.40]	1.16 [1.06 - 1.29]
Non-scalar, with prior	2.43 [2.33 - 2.56]	.46 [.40 - .54]
Scalar, no prior	2.05 [1.97 - 2.20]	.21 [.11 - .38]
Scalar, with prior	2.32 [2.19 - 2.41]	.35 [.20 - .50]

of the scalar and the non-scalar models that differed in initial state preferences of the participants: initial probability of choosing a particular state ($P(s)$) was based either on chance ("no prior") or on predictions obtained from the prior elicitation condition ("with prior"). Finally, to address the third question, we attempted to determine whether the scalar and the non-scalar models yield different predictions and whether those predictions correspond to behavioral data.

We first estimated parameter values that yielded the best fit between the data and the models (see Table 1). To ensure unbiased estimates, symmetric [-5, 5] intervals were used as priors for μ and σ. Resulting parameter values are fairly consistent across models (except for the non-scalar model with no prior, which places a higher weight on size differences), and support the results of our previous analyses: image order is more likely to influence participants' choices than relative image size.

Estimated parameter values were then used to simulate the outcome of each experimental trial for all participants. The results of the comparison between participants' responses and model predictions are shown in Figure 4. We see that only the models that take into account both the prior and the cost function can capture participants' responses. The scalar model slightly outperforms the non-scalar model, with their r^2 values equal to .91 and .90, respectively. However, this difference is not large enough to establish which model represents the best fit to human behavior.

Finally, in order to visualize general trends in the data, we compared the relative order and size of object pairs picked by participants across all trials with those selected by the models (Figure 5). We can see that all models except the baseline (prior only) successfully capture the relationship between word order and image order. However, the size preference is driven both by prior preference for equally-sized images (as shown in the experimental section) and by the word order effects (image pairs where the first mentioned image is smaller are dispreferred), so the resulting pattern is a combination of these two factors. Therefore, only models that include both estimates of the prior and effects of word order are successful at capturing relative size trends, including both the preference for "same" over "larger" (caused by the prior) and the preference for "larger" over "smaller" (caused by word order effects).

Discussion

We simulated participants' behavior during a referent selection task using models based on the Rational Speech Act (RSA) framework. We show that behavioral data are best captured by models that incorporate both participants' prior preferences and a function that evaluates the goodness of fit between word order and relative object salience. Our data do not clearly discriminate between a model without scalar inference (that incorporates word order preferences directly at the level of a non-pragmatic listener) versus a model with scalar

inference (that encodes word order preferences at the level of speaker choice). Future studies may be able to discriminate between such models by using experimental conditions specifically designed for this purpose.

The mechanisms underlying the mapping between word order and relative object salience also warrant further research. In order to compare an utterance with a visual stimulus, the listener needs to convert them into compatible representational formats (Clark & Chase, 1972). In our experiment, we show that the binomial phrase exerts an influence on the subsequent referent selection task, demonstrating that the order of words within the binomial is stored (at least briefly) in working memory and can be used to determine the best matching referent on the basis of optimal alignment between word order and visual referent features (Huettig, Olivers, & Hartsuiker, 2011). Eyetracking and other continuous performance measures can be used to further elucidate the nature of language-vision interaction during referent selection (Anderson, Chiu, Huette, & Spivey, 2011).

Conclusion

We investigated the role of word order on referential ambiguity resolution by presenting participants with a binomial phrase and having them choose among a set of possible referent pairs. We found that comprehenders' choices reflect an iconicity preference (left-before-right) and a perceptual markedness preference (larger-before-smaller), both of which have been previously demonstrated in language production. As is the case in language production (Benor & Levy, 2006), we found the iconicity preference to be stronger than the perceptual markedness preference. We showed how these preferences can be jointly reconciled in a computational pragmatic model that is improved by also taking into account utterance-independent expectations regarding likely referents. Our results demonstrate that listeners infer intended meaning based not only on the contents, but also on the structure of the referring expression, highlighting the powerful role that word order can play in resolving referential ambiguity.

Acknowledgments

We gratefully acknowledge support from NSF grants BCS-1456081 and BCS-1551866 to RPL.

References

Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychologica*, *137*(2), 181–189.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, *67*(1), 1–48. Retrieved from https://www.jstatsoft.org/v067/i01

Benor, S., & Levy, R. (2006). The chicken or the egg? a probabilistic analysis of English binomials. *Language*, *82*(2), 233–278.

Bergen, L., Goodman, N., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34).

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*(3), 472–517.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, *22*(2), 245–258.

Clarke, A. D., Elsner, M., & Rohde, H. (2015). Giving good directions: order of mention reflects visual salience. *Frontiers in Psychology*, *6*, 1793.

Cooper, W. E., & Ross, J. R. (1975). World order. *Papers from the Parasession on Functionalism*, 63–111.

Croissant, Y. (2012). *Estimation of multinomial logit models in R: The mlogit package.* R package version 0.2-2. URL: http://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf.

Degen, J., Franke, M., & Jäger, G. (2013). Cost-based pragmatic inference about referential expressions. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35).

Drummond, A. (2013). *Ibex Farm.* http://spellout.net/ibexfarm/. (Accessed: 2018-1-27)

Elsner, M., Rohde, H., & Clarke, A. (2014). Information structure prediction for visual-world referring expressions. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 520–529).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the *give* and *take* between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*(4), 544–569.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* http://dippl.org. (Accessed: 2018-1-27)

Huettig, F., Olivers, C. N., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica*, *137*(2), 138–150.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.

Kovalenko, L. Y., Chaumon, M., & Busch, N. A. (2012). A pool of pairs of related objects (POPORO) for investigating visual semantic integration: behavioral and electrophysiological validation. *Brain Topography*, *25*(3), 272–284.

Mayerthaler, W. (1988). *Morphological naturalness*. Karoma Pub.

McDonald, J. L., Bock, K., & Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, *25*(2), 188–230.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384–402.

Scontras, G., & Tessler, M. H. (2017). *Probabilistic language understanding: An introduction to the Rational Speech Act framework.* https://gscontras.github.io/probLang/. (Accessed: 2018-1-27)

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *37*(3), 776–784.