The Role of Gaze and the Semantics of Demonstratives in Referent Selection

Crystal H. Y. Chen
University of Toronto, Canada

Lyn Tieu
University of Toronto, Canada
MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia
Macquarie University, Sydney, Australia

Ana T. Pérez-Leroux
University of Toronto, Canada

Abstract

Demonstratives (*this/that*) situate objects in space with the aid of gestures and a proximal-distal contrast. However, it is unclear how these cues interact to aid the listener in referent selection. The current paper presents a referent selection task where listeners choose an object out of a group of objects based on a physical and verbal cue provided by a speaker. Results indicate that listeners are sensitive to a variety of cues, but only integrate the minimum amount of information necessary for referent selection, with physical cues being prioritized over the semantic contributions of the demonstrative.

2012 ACM Subject Classification Theory of computation → Semantics and reasoning

Keywords and phrases Demonstratives, spatial language, proximal-distal contrast, referent distance, joint attention coordination, gesture, deixis, referent selection, experimental semantics

Digital Object Identifier 10.4230/LIPIcs.COSIT.2024.20

Category Short Paper

Funding *Crystal H. Y. Chen*: Vanier CGS.
Lyn Tieu: SSHRC Insight Development Grant.
Ana T. Pérez-Leroux: Department of Spanish and Portuguese (University of Toronto).

1 Introduction

Demonstratives (*this/that*) are expressions used to situate objects in space by indicating their distance, often with the use of gestures (see (1)). In some contexts, they are used purely to direct a listener's attention while in other contexts, they point out an object's position relative to a reference point. What is unclear is the degree to which gesture and referent distance play a role in the interpretation of demonstratives and how these cues might interact with each other.

(1) I want *this cookie* (points left), not *that cookie* (points right).

The current paper presents an experiment investigating three questions: (1) *Do listeners rely on the speaker's gaze in interpreting demonstratives?* (2) *Do listeners consistently apply the proximal-distal contrast when choosing a demonstrative's referent?* (3) *How do these cues interact?* To anticipate, the results suggest that people are sensitive to both cues, but employ them hierarchically to identify a unique referent.

2 Background

Demonstratives are often accompanied by gestures [21, 13, 22, 18], which may come in the form of eye gaze, lip/chin pointing and touching [16, 13, 14, 23]. Often, these gestures seem to serve a more social interactional purpose than a spatial one [16, 13, 14, 22, 15]. The connection to gesture is unsurprising given claims that demonstratives coordinate joint attention of interlocutors towards a shared referent [17, 9, 10]. This is apparent even in early uses of demonstratives, as children first produce demonstratives only after acquiring non-verbal joint attention coordinating strategies such as eye gaze and pointing gestures [5, 6]. Early demonstratives also behave like linguistic finger-pointing and lack more complex semantic properties [24].

Demonstratives later develop spatial associations, with the proximal demonstrative *this* referring to referents nearby and the distal demonstrative *that* referring to referents further away. Together, they form a proximal-distal contrast which has been argued to be a language universal [8, 11]. Furthermore, experimental results suggest that demonstratives divide a speaker's perceptual space into peripersonal (near) and extrapersonal (far) space [7].

The shift from being language used purely for social interaction to spatial language raises the question of how much gesture and the proximal-distal contrast contribute to the interpretation of demonstratives in adult speakers. Furthermore, experimental investigations of the proximal-distal contrast have often excluded the speaker's gaze/gestures in order to avoid their influence on interpretation. But as suggested above, both are integral components of demonstratives. Their omission therefore reduces the naturalness of the demonstratives in these experiments. It is also unclear how the different cues interact with each other in the interpretation of demonstratives.

3 Method

The ethical aspects of this study were approved by the University of Toronto Social Sciences, Humanities and Education Research Ethics Board.

3.1 Participants

Thirty-one self-reported native English speakers with normal/corrected to normal vision completed the experiment. Participants were recruited through Prolific (`www.prolific.com`) and were paid at an average rate of £9.54/hour for the task (average completion time: 12m35s).

3.2 Procedure

The experiment was implemented in Gorilla Experiment Builder (`www.gorilla.sc`). Participants first provided informed consent and then completed the experimental task (i.e., referent selection task) where they had to choose an object from among a set based on instructions provided to them in each trial. Afterwards, participants provided feedback and completed a demographic survey that collected information regarding their age, gender, vision, whether they previously accessed speech services, and language background.

3.3 Materials

The experiment involved 64 trials (4 training, 12 filler, 48 critical) framed as an interactive story in which participants helped an alien character named Waba-Waba (WW) cook dinner. On each trial, WW stood in the middle of the screen surrounded by three objects on each side.

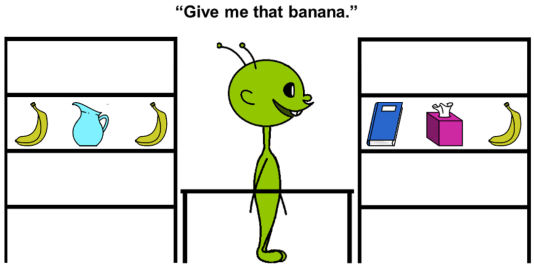


Figure 1 Example of critical trial.

He turned left or right and uttered "Give me [description of object]." Then, participants were prompted to click on an object via the instruction "What would you give to Waba-Waba?" After responding, WW turned back to thank the participant before moving on to the next trial.

Critical trials involved three different non-fruit distractors and three identical fruit targets. We manipulated three factors: Determiner, Gaze, and SingleFruitPosition (SFP). Determiner was a property of WW's linguistic cue and refers to the determiner in his instruction (Determiner = "Give me a/this/that [fruit name]"). Gaze was the physical cue provided by WW, and is defined as the direction towards which his body turned in a given trial (Gaze=towards a single fruit/a pair of fruits). Though this is not how eye gaze is typically defined, this definition approximates an informative gesture while maintaining a clear and simple visual scene. To mitigate the possibility that participants entirely ignored WW's linguistic cue in favour of his physical cue, Gaze was also designed so that some ambiguity still persisted in terms of the exact object that WW was looking at. Finally, SFP was a property of WW's environment, referring to the position of the single fruit (SFP = near/far) (see Figure 1). We employed a $3 \times 2 \times 2$ design with four repetitions per condition and two dependent variables: whether the selected fruit was in the same direction as WW's gaze and the position of the selected fruit.

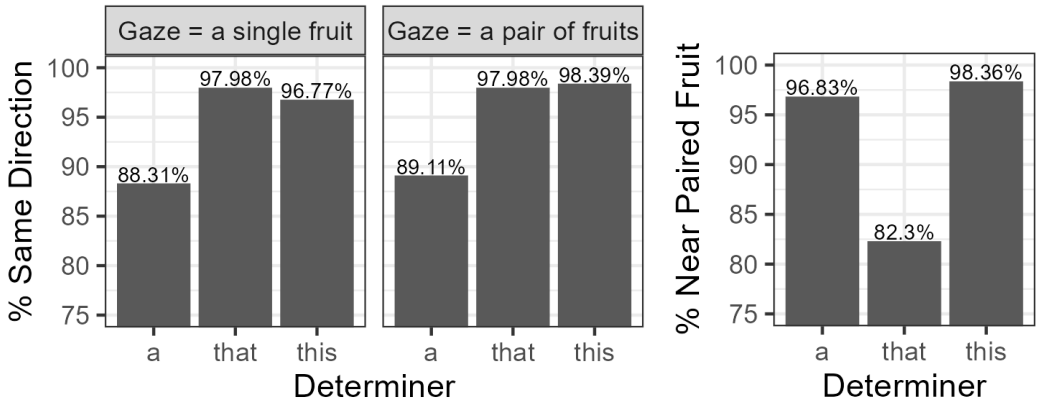
In training trials, participants were familiarized with the kind of array that would be used in critical trials (i.e., a single fruit on one side, a pair of fruits on the other), but were free to choose any fruit as WW uttered "Give me a fruit." In filler trials, the continual attention and comprehension levels of the participant were tested by having instructions involve a description matching only one object in the array (e.g. "Give me a blue book"). The experiment began with training trials, followed by critical and filler trials in a randomized order.

4 Results

4.1 The Role of Gaze

The first research question explored in this study was: *do listeners rely on the speaker's gaze in interpreting demonstratives?* Here, we analyzed the frequency with which participants chose fruits in the direction of WW's gaze. The experiment involved three independent variables: Determiner, Gaze, and SFP. Figure 2a shows that participants strongly preferred choosing fruits in the direction of WW's gaze regardless of Determiner and Gaze type, but this preference was higher for demonstratives compared to the indefinite article.

Using R (v4.3.1) and the LMER4 package [20, 2], the data from the critical trials were fitted with a mixed-effects logistic regression model with dummy variable coding. The dependent variable was whether participants chose fruits in the direction of WW's gaze; fixed



(a) Subfigure A. **(b)** Subfigure B.

Figure 2 Percentage of selected fruits in the direction of WW's gaze for Determiner × Gaze (Subfigure A) and percentage of near paired fruits selected in the direction of WW's gaze when Gaze = a pair of fruits (Subfigure B).

effects were Determiner (reference level: "a"), Gaze (reference level: a pair of fruits) and SFP (reference level: near), along with all interactions; Participant was included as a random effect. A Type III ANOVA was conducted on the model, revealing a significant intercept ($\chi^2 = 23.21, df = 1, p < 0.05$). Specifically, the odds of choosing a fruit in the direction of WW's gaze over a fruit in the other direction while he looked at a pair of fruits and uttered an indefinite phrase was 59.84 ($z = 4.82, p < 0.05$). There was a significant main effect of Determiner ($\chi^2 = 19.43, df = 2, p < 0.05$): relative to the indefinite, the distal demonstrative significantly increased the odds of participants choosing a fruit in the direction of WW's gaze by a factor of 14.51 ($z = 3.43, p < 0.05$), while the proximal demonstrative significantly increased these odds by a factor of 23.88 ($z = 3.59, p < 0.05$). All other factors were not significant.

4.2 Inferring Referent Distance Based on Demonstrative Type

A secondary question of interest was: *do listeners consistently apply the proximal-distal contrast when choosing a demonstrative's referent?* Here, we focus on the position of the selected fruit, with separate analyses conducted on the set of responses involving fruits in the same direction as WW's gaze and on those in the opposite direction of his gaze. Both Determiner and SFP were factors of interest.

Among fruits selected in the same direction as WW's gaze, participants categorically selected the single fruit if WW looked at a single fruit. When WW looked at a pair of fruits, Figure 2b shows an observable drop of 16.06% in participants' preference for the near paired fruit when encountering the distal demonstrative compared to the proximal demonstrative.

A mixed-effect logistic regression model with dummy variable coding was fitted on responses involving fruits in the direction of WW's gaze while he looked at a pair of fruits. The dependent variable was whether the selected fruit occupied a near position; fixed effects were Determiner (reference level: 'a') and SFP (reference level: near) along with their interaction. Participant acted as a random effect. A Type III ANOVA revealed a significant intercept ($\chi^2 = 30.38, df = 1, p < 0.05$). Specifically, when WW looked at a pair of fruits, with the remaining single fruit occupying a near position, and his instructions contained an indefinite

article, the odds of choosing a near fruit over a far fruit was 304.14 ($z = 5.51, p < 0.05$). There was a significant main effect of Determiner ($\chi^2 = 22.88, df = 2, p < 0.05$): compared to the indefinite article, the distal demonstrative significantly decreased the odds by a factor of 0.11 ($z = -3.47, p < 0.05$), but when comparing between the indefinite article and the proximal demonstrative, the odds did not change significantly ($z = 1.38, p = 0.17$). Lastly, compared to the proximal demonstrative, the decrease in odds for the distal demonstrative was also significant ($z = -4.12, p < 0.05$). All other factors were not significant.

Out of all responses to critical trials, there were only 78 selections (∼5.24%) of fruits in the opposite direction of WW's gaze. When he looked at a pair of fruits and participants did not follow his gaze, they categorically chose the single fruit in the opposite direction. Among the 42 responses where WW looked at a single fruit and participants chose fruits in the opposite direction (i.e., chose from among the pair of fruits), 29 selections were in response to the indefinite article, with 79.31% selecting the near fruit; 5 selections were in response to the distal demonstrative, with 80% selecting the near fruit; and 8 selections were in response to the proximal demonstrative, with all responses selecting the near fruit. Pairwise comparisons of the percentage of near fruit selections were conducted using three Fisher's tests with no significant difference across the determiners (all $p > 0.05$).

4.3 Interactions Between Gaze and Proximal-Distal Contrast

The final question of this study was: *how do gaze and the proximal-distal contrast interact?* Here, only the selections made in response to demonstratives are relevant. We compared responses to concordant demonstrative trials with those of discordant demonstrative trials. On both types of trials, WW looked at a single fruit. But on concordant demonstrative trials, WW looked at a near single fruit while uttering "this X" or looked at a far single fruit while uttering "that X", so that both cues should direct the participant towards the same fruit. On discordant demonstrative trials, WW looked at a far single fruit while uttering "this X" or at a near single fruit while uttering "that X", so the two kinds of cues would direct the participant towards different fruits. Participants' responses in discordant demonstrative trials would indicate which cue they prioritized in referent selection; if they chose the single fruit that WW looked at but whose position did not match the uttered demonstrative, then we could infer that they were prioritizing WW's gaze. If they chose the paired fruit whose position matched the uttered demonstrative's semantics, but which WW did not look at, then they would be prioritizing the linguistic cue. Furthermore, if there was no difference in the responses between discordant and concordant trials for a given demonstrative, this would indicate that the mismatch in semantics did not interfere with referent selection. In other words, the semantics did not further contribute to referent selection in such conditions.

Across both determiners, there was a near categorical preference (all above 95%) for fruits in the same direction as WW's gaze for both demonstratives on discordant and concordant trials. An Exact Fisher's test revealed no significant difference between the percentage of single fruits being selected on concordant versus discordant trials for the proximal demonstrative (95.97% vs. 97.58% respectively, $p = 0.72$). There was also no significant difference in percentage of single fruits selected for the distal demonstrative on concordant and discordant trials (97.58% vs. 98.39%, $p = 1$).

5 Discussion

This experiment set out to investigate how speaker gaze and the proximal-distal contrast contribute to a listener's referent choice. But in providing both physical and linguistic cues to listeners, this experiment also investigated how the two cues interact with each other,

providing a more complete view of how listeners interpret demonstratives. The results indicate that participants used speaker gaze to guide referent choice, as reflected in the overall high percentage of fruits selected in the direction of WW's gaze, even in response to an indefinite article, which has no inherent association with gesture. This preference for following the speaker's gaze increased significantly (and at relatively equal levels) for both demonstratives, supporting claims that demonstratives as a whole are involved in joint attention coordination. Participants also consistently applied the proximal-distal contrast in response to demonstratives. When faced with the choice between two potential referents, participants selected the near fruit at near categorical levels for the proximal demonstrative, but this preference dropped significantly for the distal demonstrative. Moreover, this contrast was present only in selections of fruits in the same direction as WW's gaze which in turn was inadequate in identifying a singular referent.

In terms of interactions between the two cues, we consider two possibilities. The first is that each cue independently outlines a set of potential referents and the optimal referent (i.e., the entity satisfying all cues) lies within the intersection of the sets (i.e., $A \cap B$). The second possibility is that cues are integrated hierarchically, such that one cue identifies a set of potential referents and another cue is used to subset this set, with the iterative process continuing until the set contains a single entity corresponding to the optimal referent (i.e. $B \subseteq A$). Both would give rise to the same optimal referent in the current experiment, but the sets of potential referents would differ. Under independent integration, entities satisfying only cue A or B would still be a potential referent. But under hierarchical integration, potential referents must always satisfy cue A, even if they do not satisfy cue B.

In our experiment, interlocutors appeared to employ hierarchical integration, as indicated by the presence of the proximal-distal contrast in fruits selected in the direction of WW's gaze and the absence of the contrast in fruits selected in the opposite direction. If participants integrated each cue independently, then the contrast should still be present even when participants ignored WW's gaze. Furthermore, speaker gaze seems to be integrated before the proximal-distal contrast as suggested by the near categorical preference for fruits in the direction of WW's gaze on trials where the position of said fruits did not match the demonstrative's semantics. If the proximal-distal contrast was prioritized over gaze, then participants should have chosen fruits whose position *did* match the demonstrative's semantics, even if WW was not looking at said fruits. The finding that people did not differentiate between trials with matching and mismatching cues suggests that the proximal-distal contrast did not further contribute to referent selection, provided gaze was enough to pick out a unique referent. All these findings suggest that listeners are economical in choosing referents; while speakers may provide a variety of information indicating a desired referent, listeners will only use the minimum amount of information necessary to choose the referent.

Lastly, participants were affected by a strong proximity bias, a type of affordance bias [4, 3]. This explains the general preference for near grouped fruits even for the distal demonstrative, which is associated with far referents, and the indefinite article, which has no association with referent distance. This lack of spatial associations for the indefinite suggests a proximity bias.

The current experiment has certain limitations. One limitation pertains to the validity of using the alien's body turn as a stand in for gaze. Although it did not limit the results, future studies could explore potential differences between using stylized cartoon images vs. naturalistic videos. In terms of sampling, an anonymous reviewer raises questions about potential limitations of recruiting through Prolific. It is true that demographics are generally self-reported on online platforms such as Prolific, with little to no interaction with

investigators required. However, a number of studies have investigated the quality of data collected through online platforms for behavioural research, and have reported that Prolific generally yields high quality data, particularly compared to other data collection platforms (see, for example, [19, 12, 1]).

6 Conclusion

Despite their structural simplicity, demonstratives are semantically rich and involve a complex referent identification process which is not solely limited to choosing a referent based on distance. The current paper investigated how listeners integrated physical and linguistic cues in demonstrative referent identification and how these cues might interact with each other. Listeners seem to be aware of the various cues available in the context, but only employ certain ones for referent selection. Response patterns suggest hierarchical integration of multiple cues, with physical cues and visual world biases taking priority over the semantic contributions of the demonstratives. More generally, the present findings suggest that referent selection is a far more complex process than formal theories might assume.

References

1 Derek A. Albert and Daniel Smilek. Comparing attentional disengagement between prolific and mturk samples. *Scientific Reports*, 13(1):20574, 2023. doi:10.1038/s41598-023-46048-5.

2 Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi:10.18637/jss.v067.i01.

3 Craig Chambers. The role of affordances in visually situated language comprehension. *Visually situated language comprehension*, 12:205–226, 2016.

4 Craig G. Chambers, Michael K. Tanenhaus, and James S. Magnuson. Actions and affordances in syntactic ambiguity resolution. *Journal of experimental psychology: Learning, memory, and cognition*, 30(3):687–696, 2004. doi:10.1037/0278-7393.30.3.687.

5 Eve V. Clark. From gesture to word: On the natural history of deixis in language acquisition. *Human growth and development*, pages 1–38, 1976.

6 Eve V. Clark and CJ Sengul. Strategies in the acquisition of deixis. *Journal of child language*, 5(3):457–475, 1978. doi:10.1017/S0305000900002099.

7 Kenny R. Coventry, Bernice Valdés, Alejandro Castillo, and Pedro Guijarro-Fuentes. Language within your reach: Near–far perceptual space and spatial demonstratives. *Cognition*, 108(3):889–895, 2008. doi:10.1016/j.cognition.2008.06.010.

8 Holger Diessel. Demonstratives: Form, function, and grammaticalization, typological studies. *Language*, 42, 1999.

9 Holger Diessel. Demonstratives, joint attention, and the emergence of grammar. *Cognitive Linguistics*, 17(4):463–489, 2006. doi:10.1515/COG.2006.015.

10 Holger Diessel. Demonstratives, frames of reference, and semantic universals of space. *Language and Linguistics Compass*, 8(3):116–132, 2014. doi:10.1111/lnc3.12066.

11 Robert M.W. Dixon. Demonstratives: A cross-linguistic typology. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 27(1):61–112, 2003. doi:10.1075/sl.27.1.04dix.

12 Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720, 2023. doi:10.1371/journal.pone.0279720.

13 Nick J. Enfield. Demonstratives in space and interaction: Data from lao speakers and implications for semantic analysis. *Language*, 79(1):82–117, 2003. doi:10.1353/lan.2003.0075.

14 Mats Eriksson. Referring as interaction: On the interplay between linguistic and bodily practices. *Journal of Pragmatics*, 41(2):240–262, 2009. doi:10.1016/j.pragma.2008.10.011.

15 Sonja Gipper. Pre-semantic pragmatics encoded: a non-spatial account of yurakaré demonstratives. *Journal of Pragmatics*, 120:122–143, 2017. doi:10.1016/j.pragma.2017.08.012.

16 John Hindmarsh and Christian Heath. Embodied reference: A study of deixis in workplace interaction. *Journal of pragmatics*, 32(12):1855–1878, 2000. doi:10.1016/S0378-2166(99)00122-8.

17 Stephen C. Levinson. Deixis and pragmatic. *The Handbook of Pragmatics*, pages 97–121, 2004.

18 Stephen C. Levinson. Introduction: demonstratives: patterns in diversity. In *Demonstratives in cross-linguistic perspective*, pages 1–42. Cambridge University Press, 2018. doi:10.20759/elsjp.98.0_168.

19 Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. Data quality of platforms and panels for online behavioral research. *Behavior research methods*, 54(4):1643–1662, 2022. doi:10.3758/s13428-021-01694-3.

20 R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL: https://www.R-project.org/.

21 Craige Roberts. Demonstratives as definites. *Information sharing: Reference and presupposition in language generation and interpretation*, pages 89–196, 2002.

22 Anja Stukenbrock. *Deixis in der face-to-face-Interaktion*, volume 47. Walter de Gruyter GmbH & Co KG, 2015.

23 Leonard Talmy. *The targeting system of language*. MIT Press, 2018.

24 Roger Wales. Deixis. *Language acquisition*, 2:401–428, 1986.