Language and Cognition (2024), 16: 4, 1960–1986
doi:10.1017/langcog.2024.39

ARTICLE

Better early than late: the temporal dynamics of pointing cues during cross-situational word learning

Rachael W. Cheung[1,2], Calum Hartley[2] and Padraic Monaghan[2]

[1]Department of Health Sciences, University of York, Heslington, York YO10 5DD, UK and [2]Department of Psychology, Fylde College, Lancaster University, Bailrigg, Lancaster LA1 4FY, UK
Corresponding author: Rachael W. Cheung; Email: rachael.cheung@york.ac.uk

(Received 31 July 2023; Revised 19 July 2024; Accepted 25 July 2024)

Abstract

Learning the meaning of a word is a difficult task due to the variety of possible referents present in the environment. Visual cues such as gestures frequently accompany speech and have the potential to reduce referential uncertainty and promote learning, but the dynamics of pointing cues and speech integration are not yet known. If word learning is influenced by *when*, as well as *whether*, a learner is directed correctly to a target, then this would suggest temporal integration of visual and speech information can affect the strength of association of word–referent mappings. Across two pre-registered studies, we tested the conditions under which pointing cues promote learning. In a cross-situational word learning paradigm, we showed that the benefit of a pointing cue was greatest when the cue preceded the speech label, rather than following the label (Study 1). In an eye-tracking study (Study 2), the early cue advantage was due to participants' attention being directed to the referent during label utterance, and this advantage was apparent even at initial exposures of word–referent pairs. Pointing cues promote time-coupled integration of visual and auditory information that aids encoding of word–referent pairs, demonstrating the cognitive benefits of pointing cues occurring prior to speech.

Keywords: cross-situational word learning; endogenous cues; language acquisition; pointing gesture; referential ambiguity

Introduction

The environment surrounding the language learner is busy and multifaceted, with many sources of information that convey meaning (Holler & Levinson, 2019), such as auditory cues (e.g. sound-based information in speech) and visual cues (e.g. facial expressions and body movements). How does the language learner navigate this complexity of information to aid their learning? In this study, we investigate how the

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

temporal production of two such information sources – words and gestures – is combined by the adult language learner to disambiguate and retain novel word–referent relationships.

Learning new vocabulary involves determining how unfamiliar words relate to aspects of the environment (*referent selection*) and then encoding these pairings for later retrieval (*retention*).

Even when restricted to learning only associations between nouns and objects, there are multiple possible mappings between words and the correct object ('referent') available to the learner (Yu & Ballard, 2007). Consequently, constraints that have been proposed to address how to correctly pair words and referents have tended to focus on biases internal to the learner that guide their referent selection, such as mutual exclusivity (Halberda, 2006; Markman & Wachtel, 1988) or assuming a novel label refers to a novel object (Carey & Bartlett, 1978; Golinkoff et al., 1992). However, these strategies cannot be applied by learners in situations where all potential referents are novel.

An alternative approach is to consider how information from the wider environment can contribute to general learning processes, such as cross-situational statistics (Siskind, 1996). Cross-situational statistics refers to the aggregation of information and commonalities across several, rather than single, learning instances (Yu & Smith, 2007). Thus, a learner can acquire novel label–object pairs by tracking the co-occurrence of words and objects across multiple exposures (e.g. Fitneva & Christiansen, 2011; Monaghan & Mattock, 2012; Roembke & McMurray, 2016; Smith et al., 2011; Yu & Smith, 2007; Yurovsky et al., 2013).

However, cross-situational statistics represent only one source of environmental information that a learner can utilise when faced with multiple unknown referents. Other environmental cues, such as gaze direction, prosody and gesture cues (e.g. Hollich et al., 2000), might be combined with cross-situational word learning to facilitate mapping of word–referent pairs (Dunn et al., 2024; Hartley et al., 2020; Monaghan et al., 2017; Yu & Ballard, 2007). For instance, pointing cues (e.g. deictic gestures or gaze direction) may modulate the degree of referential ambiguity by directing learners towards the intended referent, reducing the formation of spurious word–object associations (MacDonald et al., 2017). In a cross-situational word learning study, Dunn et al. (2024) found that including reliable gaze direction as a cue to target referents for novel words increased looks to targets over foil objects compared to when gaze was less reliably coordinated with cross-situational statistics.

In adult cross-situational word learning, the presence of a visual gesture cue (implemented as an arrow pointing to the intended referent) resulted in higher accuracy (Monaghan et al., 2017), showing that learners are able to combine information from speech and gesture to constrain their formation of novel word–referent associations. In more naturalistic learning situations, deictic pointing cues in parent–infant communication have been shown to support a high degree of accuracy in identifying a word's intended referent when adults watch recordings of the interactions (Cartmill et al., 2013; Frank et al., 2013). Taken together, these studies show that auditory and gesture information can be combined to reduce referential uncertainty and support word learning.

Co-occurrence of gesture and speech during communication is prevalent in communication, both in terms of deictic gestures indicating place and iconic gestures indicating form of referents (Goldin-Meadow, 2003; Kita, 2009; McNeill, 2000). Furthermore, gestures tend to *precede* referential speech in production (Beun &

Cremers, 1998; Levelt et al., 1985; McNeill, 1985), with gesture onset seeming to be exquisitely linked in timing to the production of the referring word rather than constrained by the production requirements of the utterance (Chu & Hagoort, 2014). In a multimodal corpus study of a large number of utterances coded for co-occurring gestures, Donnellan et al. (2022) found that deictic gesture onset to a referent tended to occur approximately 370ms prior to the onset of a referential word.

Despite the numerous studies of temporal arrangement of gesture and speech production, the utility of this gesture–speech sequencing has not been studied in detail. In the Human Simulation Paradigm (HSP; Gillette et al., 1999), adult participants guess 'missing' words from parent–child interaction videos, where the target word is obscured by an auditory 'beep' (e.g. 'where's the [obscured target word]?'). Scoring participants' accuracy of guess provides a measure of how informative any surrounding cues are when identifying the target word. Trueswell et al. (2016) found that timing of gestures made by parents within parent–child interaction videos predicted the accuracy of other adult participants' guesses regarding the intended referent. Shifting the obscuring 'beep' 2–4 seconds away from actual word occurrence significantly reduced guessers' accuracy in identifying the target referent.

Nirme et al. (2020) investigated how timing of deictic gesture and speech affected judgements of naturalness for communicative acts. They found that gestures occurring 500ms before or after labelling an object resulted in no effect, except when the gesture coincided with a pause in speech which reduced naturalness ratings. Habets et al. (2011) manipulated timing of iconic gestures and referential naming and found that gesture preceding word onset by 360ms resulted in effective semantic integration of gesture and speech information, as measured by EEG N400 signals, whereas gesture preceding words by 180ms or simultaneous occurrence resulted in less efficient integration (Habets et al., 2011). Furthermore, Cavicchio and Busà (2023) found that moving an iconic gesture from co-occurring with a verb reference in English to the beginning of the sentence containing the verb resulted in slower identification of the action by English additional language learners, though there was no significant difference for first language English speakers. These results indicate that not only the presence but also the *timing* of gestural cues relative to speech may be critical for supporting word–referent mapping (Trueswell et al., 2016), though research has yet to directly demonstrate this effect in word learning.

Gesture occurring before speech, to orient attention to the intended referent, is consistent with studies of cued attention (e.g. Hauer and Macleod 2006). Such attentional cue studies distinguish *endogenous* cues (e.g. arrows or eye gaze), where attention is directed voluntarily to a target, from *exogenous* cues (e.g. flashing lights), where attention is directed automatically due to sudden salient stimuli (Jonides 1981; Posner, 1981). Naturalistic social cues during word learning, such as pointing cues, likely act as endogenous cues similar to those that are examined during attention shifting experiments (Brignani et al., 2009). There appears to be temporal sensitivity to the role of these cues in adults; whereas exogenous cues quickly shift focused attention between a cue and a target at 50ms, shifts of focused attention due to endogenous cues may take up to 500ms (Berger et al., 2005; Shepherd & Müller, 1989).

Therefore, the timing of a cue in relation to label utterance could be crucial to successful word–referent mapping and how attention is directed. Focusing attention on a referent shortly after it is labelled may be significantly less optimal for learning than focusing attention *during* label utterance following early cues prior to the

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

naming event. Such an effect would suggest that the occurrence of gesture before naming in naturalistic communication (e.g. Donnellan et al., 2022) may be optimal for language learning due to the (endogenous) attentional shift that it precipitates. However, these predictions from observational studies about the importance of gesture timing to word learning have not yet been tested in controlled studies. Observational studies are unable to systematically control the distribution of cues, their timing or other potential sources of information that may interact with gesture and speech.

In particular, we do not yet know whether a pointing cue to an intended referent occurring immediately before (versus after) speech may be critical for learning, nor whether the time window of sensitivity might be less than the 2s observed in Trueswell et al. (2016) and is perhaps closer to the 360ms asynchrony investigated by Habets et al. (2011) in their study of iconic gestures. Furthermore, although multiple sources of information may aid accurate referent selection, disambiguation of meaning does not necessarily reflect long-term learning. Accurate referent selection under referential ambiguity may reflect 'fast' in-moment problem-solving by the learner, whereas retention of novel words may occur as a 'slow' and gradual process, during which multiple exposures strengthen or weaken word–referent pairs over time (McMurray et al., 2012).

In these respects, investigating the timing of pointing cues is critical for refining models of word learning. If word learning is influenced by *when*, in addition to *whether*, the learner is directed to the intended referent, then this would suggest that strength of associations when acquiring word–referent mappings is influenced by the quality (and not only the quantity) of integration of visual and speech information (Bhat et al., 2022). Such findings would signify the need to refine standard associative learning models where temporal contiguity has not been considered (McMurray et al, 2012; Yu & Smith, 2012) and would provide evidence that the temporal relation found between gesture and speech production also has an effect on language learning. An alternative perspective is that the relative timing is more of an accident of production constraints (e.g. Chu & Hagoort, 2014) and has no impact on word learning.

The current study

In this study, across two studies we examine how adult learners identify word–referent pairings by using environmental cues to reduce referential ambiguity and how this might affect their subsequent retention of novel words. Our research addresses three novel questions: (1) What are the effects on learning accuracy of pointing cues that occur before, versus after, a referent is labelled? (2) Do any facilitative effects of pointing cues on referent selection accuracy also apply to longer-term retention of words? (3) What temporal dynamics of looking behaviour reflect learning from pointing cues presented before, versus after, labelling the referent?

Studies 1 and 2 investigated the temporal process of how pointing cues are integrated with auditory and visual information to support accurate cross-situational word learning. We manipulated the timing of a pointing cue (Study 1) and employed an eye-tracker to uncover how the dynamics of visual attention are affected by pointing cue timing (Study 2). In each study, we tested how our manipulations

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

affected both immediate recall and retention (after a delay) of novel word–referent mappings. Given that Nirme et al. (2020) found some evidence that gesture–speech timing affected judgements of naturalness of the communicative situation, we also measured the extent to which participants were aware of the variation in timing between gesture and speech. We used a static photograph of a finger and hand as a pointing cue. This stimulus was chosen to limit additional visual information, such as oromotor movements associated with speech or eye gaze. Previous studies of pointing gestures and speech in human–machine interaction have sometimes used a virtual avatar (e.g. Kranstedt et al., 2006; Nirme et al. 2020) or recorded human gestures (e.g. Cavicchio & Busà, 2023; Habets et al., 2011). However, naturalistic gestures extend over a few hundred milliseconds (Donnellan et al., 2022) and determining when a deictic gesture begins to provide referential information is imprecise. In this study, we aimed to investigate the close temporal relation of speech and pointing relative to word learning with control over the precise timing of the gesture cue. Furthermore, previous research has identified that operationalisation of gesture as a pointing hand is effective as a cue to learning in cross-situational word learning (Monaghan et al., 2017). We reflect on potentially using more naturalistic gestures in future investigations in the General Discussion. All pre-registrations, data, experimental stimuli and tasks, and code for all analyses in this study are available on the Open Science Framework (OSF): https://osf.io/2m9pe/?view_only=9d64688d03d84704aa5f2e8f8eb34dc9.[1]

Study 1: When are pointing cues in word learning most useful?

Study 1 investigated whether cue timing effects apply to adults' use of pointing cues in cross-situational word learning. As endogenous cues appear to induce slower attention shifts than exogenous cues (Shepherd & Müller, 1989), pointing cues that occur sometime before, rather than after, a label may be critical to encoding robust label–target associations and minimizing spurious label–foil associations. We manipulated the timing of pointing cues relative to label utterance across two conditions: pointing appeared *before* or *after* the verbal label. In the HSP, Trueswell et al. (2016) found that shifting an obscured word 2 seconds earlier than the word's original position was sufficient to reduce the accuracy score of those guessing the missing word from ~ 60% to ~ 43%. Furthermore, if the obscuring 'beep' was moved too early, guessers did not relate the visual event to the missing word, as they were perceived as too temporally discontinuous. However, shifting attention between potential referents during word learning can happen very quickly (e.g. within 225ms, Halberda, 2006), and Habets et al. (2011) already found a semantic integration change from 360ms to 180ms asynchronies for iconic gestures and word naming. We therefore assessed whether sensitivity to cue timing can be observed in a smaller temporal window than tested by Trueswell et al. (2016) in the HSP by presenting pointing cues just 1 second before and after a novel label, at a point in between the parameters of Trueswell et al.'s (2016), Nirme et al. (2020) and Habets et al.'s (2011) studies.

We hypothesised that participants would respond more accurately on both immediate and retention trials when tested on words trained in the early pointing

[1] Please note that two additional experiments were pre-registered with those reported in this manuscript; the results of which are reported on OSF for full transparency.

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

condition compared to the late condition. Early pointing cues may support cross-situational word learning by highlighting the target prior to (or at) label utterance, reducing spurious associations between the label and non-target foils. Late pointing cues may be less useful for word–referent mappings as any attentional shift that occurs due to the pointing cue will be after the crucial information (the label) has been uttered, reducing the chance to reconcile the auditory label and the visual referent together and robustly encode the association.

Method

Participants were twenty monolingual English-speaking adults without any sensory deficits (age M = 20.9 years, SD = 5.16, range = 18.0 – 39.0; 5 male, 15 female), as specified in the pre-registration. They were recruited via leaflets and the *** University research participation system, which allows all members of the University community to partake in research. Informed, written consent was obtained from all individuals prior to participation. Participants were either paid £3.50 or received course credit for taking part. The number of participants was specified in the pre-registration and based on previous studies that test cross-situational word learning using a similar paradigm (e.g. Monaghan et al., 2015; Monaghan & Mattock, 2012).

Materials All stimuli used can be found on OSF. Thirty-two novel objects and 32 novel two-syllable words were taken from the NOUN database (Horst & Hout, 2016). Sound files for each word were made using the Serena system voice (Macintosh computer, OS 10.13). Each object and word were paired randomly for each participant to produce 32 word–object mappings. Pictures and audio were presented on a Macintosh computer (OS 10.13, 21.5-inch monitor, 1920 × 1080 resolution) using PsychoPy3 (Pierce & MacAskill, 2018). Participants used closed cup headphones.

Procedure Testing took place in a quiet room. Both studies included two training and test conditions and were run using a similar procedure. Participants first completed a warm-up with two familiar objects and words presented as they would be during training. The order of conditions was counterbalanced across all participants. During the first condition, participants were administered the first training block with one set of 16 word–referent pairs, followed by an immediate testing block, then a 5-minute distractor task (colouring in a geometric picture), before completing a retention testing block. They then repeated this process with another set of 16 word–referent pairs for the second condition.

Each correct word–referent pairing appeared four times per training condition, with 16 word–referent pairings to be learnt per condition. Screen position of the objects was pseudo-randomised so that the target appeared an equal number of times on the left and on the right. The order of trials within training blocks was pseudo-randomised with the constraint that referents appeared no more than twice in a row. Target objects also acted as foils for their non-associated words and were pseudo-randomised with the constraint of appearing an equal number of times across all trials. To ensure that participants could disambiguate words and referents based on cross-situational information, co-occurrences of the same targets and foils were minimised across trials.

Training blocks Participants completed two cue conditions, an ‘early’ and a ‘late’ pointing cue condition, in counterbalanced order. These cues were blocked, which enabled us to probe participants’ awareness of cue timing differences at debrief, without the need for leading questions about the asynchrony. At all times,

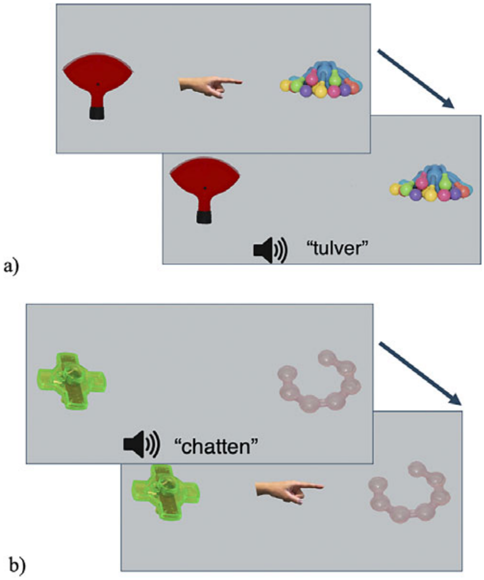

Figure 1. Studies 1 and 2: Training trials, a) early pointing cue, b) late pointing cue condition.

participants saw two novel objects on screen with a pointing cue – a picture of a hand pointing to the target appeared simultaneously with the referent. The target word in both conditions was played 500 milliseconds after referent presentation. In the early condition, participants saw the pointing cue 1 second *before* word utterance (Figure 1a). In the late condition, the pointing cue appeared 1 second *after* word utterance. In both conditions, the two referents appeared for the duration of the trial (3 seconds), label utterance occurred at the same time at the 2 second mark after the referents had first appeared, and the cue lasted for 1 second (Figure 1b). The timing of the pointing cue with the novel label was adjusted to ensure an equal amount of time before and after label utterance in both conditions.

Testing blocks In order to test learning accuracy for the word–referent pairs, participants were administered two testing blocks: *immediate*, which occurred immediately after training, and *retention*, which occurred after a 5-minute distractor task (colouring in a complex picture). Each word was tested on one immediate trial and on one retention trial. During test trials, all 16 referent objects were presented simultaneously on screen, and the learner was asked to click on the correct referent for each target word, requested in a random order ('*which is the [target word]?*'; chance level = 0.0625; Figure 2). The on-screen positions of the referents differed for immediate and retention trials. Participants were asked at debrief after the study had finished if they had noticed any difference between the two training blocks and their response was recorded.

Statistical analysis

As pre-registered, accuracy of correct word–referent pairs was scored as 1 (correct) or 0 (incorrect) and entered into general linear mixed effects models (GLMEs), using *glmer* from the *lme4* package (v.1.1-20, Bates et al., 2015) in R Studio [v1.1.463; R v.3.6.3]. Separate analyses were conducted for immediate testing blocks, retention

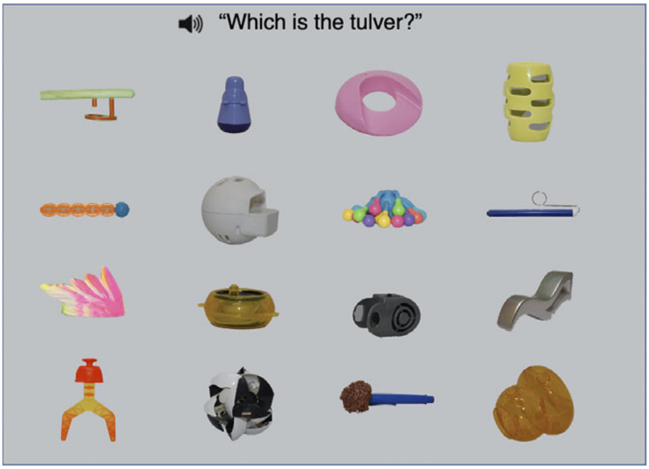
“Which is the tulver?”

Figure 2. Studies 1 and 2 testing trial example: participants see all 16 referents for given condition and are asked to click on the corresponding object for novel words.

testing blocks and all testing blocks combined (i.e. immediate and retention testing blocks). This enabled direct comparison between trial types, reflecting the discrete processes that may underlie immediate referent selection and retention of novel words after a delay. All model fitting sequences began with a baseline model that contained only random effects. Subsequent models were then built progressively by adding individual fixed effects and comparing each model to the previously best-fitting model using log-likelihood comparisons (Barr et al., 2013), selecting the more complex model if it was a significantly better fit. A frequentist approach was utilised, where comparisons p <.05 were classed as statistically significant.

For all models, we used sum-to-zero coding. For models predicting immediate testing accuracy, a fixed effect of pointing cue condition (‘1’ = early, ‘−1’ = late) was included. For models predicting retention accuracy, a fixed effect of pointing cue condition (‘1’ = early, ‘−1’ = late) was tested, followed by a fixed effect of accuracy for each word on immediate testing trials (‘1’ = correct, ‘−1’ = incorrect) and then for presence of interactions between condition and immediate accuracy. For models predicting overall accuracy across trial types, we first tested the fixed effects of pointing cue condition (‘1’ = early, ‘−1’ = late), then the effect of trial type (‘1’ = immediate, ‘−1’ = retention) and then for the presence of interactions between condition and trial type. For all models, random effects of participant, target word and target object and test order (early or late condition first) were included, and random slopes of condition were fitted for each random intercept unless this prevented the model from converging.

Results and discussion

The final best-fitting models and results for all three analyses are presented in Table 1 and Figure 3. Participants performed statistically above chance in both conditions on immediate and retention trials.

Table 1. Study 1: Best-fitting general linear model results predicting trial accuracy by pointing cue condition

Immediate trial accuracy				
Fixed effect	estimate	SE	*z*-value	*p*-value
(*intercept*)	1.08	0.42	2.58	.010
Pointing cue condition	0.30	0.14	2.20	.028
Retention trial accuracy				
(*intercept*)	0.38	0.29	1.32	.187
Pointing cue condition	0.37	0.20	1.87	.062
Immediate accuracy	1.47	0.14	10.35	<.001
Overall accuracy				
(*intercept*)	1.05	0.43	2.43	.015
Pointing cue condition	0.43	0.16	2.73	.006
Trial type	0.15	0.07	2.15	.032

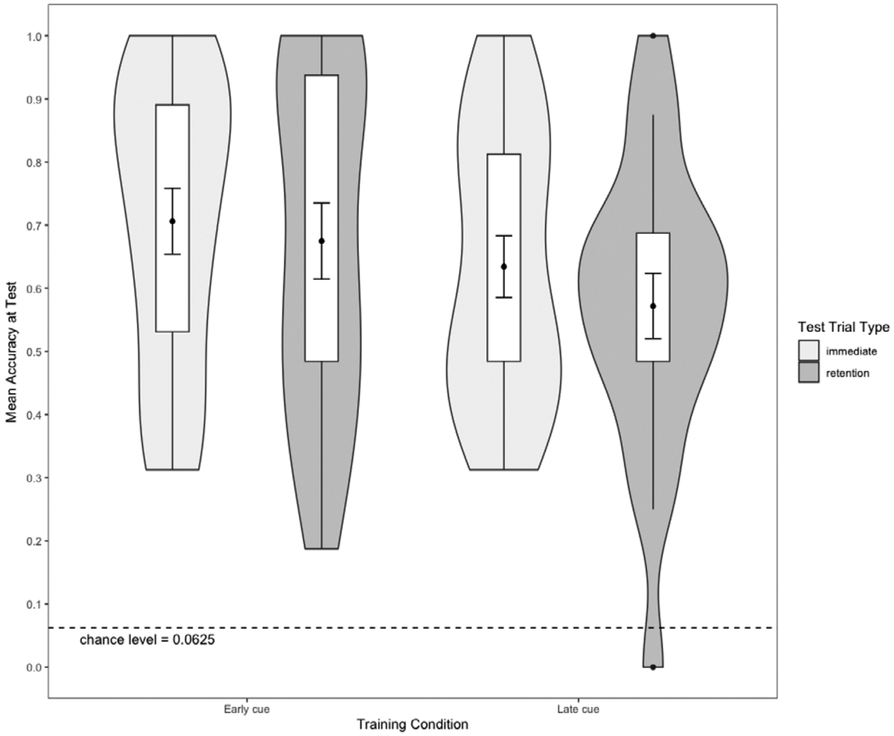

Figure 3. Study 1: The effect of pointing cue timing on behavioural response – mean accuracy across testing trials with standard error bars of all participants, grouped by pointing cue condition and trial type.

The final random-effects structure included by-participant and by-target word random intercepts with slopes for condition and random intercepts of target object and test order, across all models. Slopes of condition for target object and test order did not converge despite using allFit() procedures; these had the lowest variance so were removed (Barr et al., 2013). The best-fitting model for immediate testing trials demonstrated a fixed effect of pointing cue condition ($\chi^2(1) = 4.35$, p =.037).

Participants were significantly more likely to respond accurately in the early pointing condition compared to the late pointing condition (p =.028). The best-fitting model for retention trials included fixed effects of immediate accuracy and condition ($\chi^2(1)$ = 146.1, p <.001), although there was no significant effect of condition once immediate accuracy was also included (p =.062). Participants were significantly more likely to respond correctly on retention trials if they responded correctly on immediate trials for the same word (p <.001)

Overall, participants had higher accuracy (immediate and retention test trials) in the early condition (M = 0.69) compared to the late condition (M = 0.60). For overall accuracy, the best-fitting model contained fixed effects of both pointing cue condition and trial type ($\chi^2(1)$ = 4.50, p =.034), indicating that participants were more likely to respond correctly when tested on words learnt in the early pointing cue condition compared to the late pointing cue condition (p =.006) and were more likely to respond correctly in retention than immediate test trials overall (p =.032).

At debrief, only four of the 20 participants reported noticing a difference between conditions related to the pointing cue. This was unexpected, as the conditions were split into two distinct training blocks and the timing differences between words and pointing spanned a 1-second interval, which we expected to be easily detectable.

The results of Study 1 indicate that temporal ordering of cues with word utterance is important when initially establishing word–referent pairs, consistent with the cued attention literature (Hauer and Macleod 2006; Yoshida and Burling, 2012). Our results not only confirm the importance of cue timing to referent selection (Trueswell et al., 2016) but also indicate that the effect of temporal co-occurrence is more fine-grained than −2 to + 2 seconds. Pointing cues during training that occur just 1 second before label utterance significantly improved accuracy at test when compared to those that occurred 1 second after word utterance. Whether gestures occurring even closer to naming, as in the 500ms used in Nirme et al. (2020) or the 360ms used in Habets et al. (2011), may boost learning further is an open question that we revisit in the General Discussion. Further, the effect of temporal synchrony of pointing and spoken label during referent selection also influenced retention accuracy in our cross-situational paradigm.

Interestingly, only four of the 20 participants reported noticing that the pointing cue appeared at different time points within trials across the two conditions. This suggests that the temporal synchrony of pointing cue and spoken information was not explicitly available to the majority of participants, indicating that strategic use of information was likely not driving performance and that differences in test accuracy between conditions were not due to conscious manipulation of attention by learners. These results, however, do not yet indicate how learners' attention to objects is affected by the timing of a pointing cue and what pattern of visual attention relates to learning – we therefore examine this using an eye-tracker in Study 2.

Study 2: How do early pointing cues support more accurate word learning than late pointing cues?

We hypothesised that the advantage of early pointing cues over late cues was due to *where* attention was allocated during, rather than following, label utterance. Early pointing may benefit learning by endogenously cuing orientation of visual attention to the target referent before the word is named (Hauer and Macleod, 2006), thus strengthening the link between word and referent. That is, participants may have learned more effectively in the early condition because they were already looking at the target object when they heard the referring label. We therefore repeated the procedure of Study 1 to replicate the behavioural effects, but also monitored participants' gaze during training trials using an eye-tracker, allowing us to pinpoint where their attention was directed during label utterance. We made two additional predictions relating to the temporal dynamics of multiple cue integration during word learning: (1) if the early pointing cue promotes attention to the target over the foil, participants would have increased overall relative looking time to the target compared to the foil during training trials in the early condition (relative to the late pointing cue condition), and (2) if the early pointing cue advantage for learning is due to where attention is located when the word is spoken, then greater accuracy would be predicted by fixations to the target during and immediately after the spoken label, but not prior to the spoken label.

Method

Participants were twenty monolingual English-speaking adults without any sensory deficits who had not partaken in Study 1 (*M* age = 19.9, *SD* = 4.15, range = 18.0 – 37.0; 5 male, 15 female), as specified in the pre-registration. They were recruited and reimbursed as per the procedures outlined in Study 1.

Materials The materials remained the same as in Study 1, with the following exceptions: a Tobii Pro X3-120 eye-tracker was used (sampling rate 120Hz) in conjunction with a Windows computer (17-inch monitor, screen resolution 1600 × 900) to track binocular participant gaze throughout training trials. Participants were seated approximately 60cm away from the eye-tracker.

Procedure Participants' eye positions were calibrated using the Tobii Eye-Tracker Manager five-point calibration system before the experiment. The rest of the procedure followed that of Study 1.

An average of binocular data from the left and right eye was taken to give a single (x, y) coordinate for each gaze point. Where data from one eye were missing, data from the other eye were taken. If data from both eyes were missing, linear interpolation within participant and within trial was used to smooth the data.

The data were split into time bins of 250 milliseconds, and three distinct areas of interest (AOIs) were identified: cue, foil and target object. Fixations within these AOIs were detected using the *saccades* package (von der Malsburg, 2015) in R [v1.1.463], allowing for isolation of fixations whilst disregarding artefacts such as blinks. All processing code is available on OSF.

Statistical analysis

We first constructed GLME analyses in the same way as for Study 1 with behavioural response data at test only (Analysis 1). We then examined the effect of pointing cue timing on the learning process during training, first descriptively and then by employing growth curve analysis (GCA) to analyse target fixation proportion across conditions (Analysis 2). GCA allows for modelling of differences between participants whilst allowing for within-participant differences across time (Mirman et al., 2008). We used the best-fitting orthogonal polynomials for the time form function, testing up to cubic polynomials. GCAs were fitted according to Mirman (2014) using *lme4* in R Studio. A baseline model was constructed that predicted mean fixation proportion to target with fixed effects of all time terms, random slopes of all time terms per participant and random slopes of time terms for each participant per condition. These models failed to converge despite applying techniques to retain maximal random-effects structure (Barr et al., 2013; Mirman, 2014), resulting in a baseline model of all time terms with random effects of all time terms per participant. Subsequent models were then built up by adding a fixed effect of pointing cue timing condition (early or late) to the intercept only and then adding a fixed effect of pointing cue timing condition to all time terms. Each model was compared to a baseline model, or previous best-fitting model, using log-likelihood comparisons. For all models, the early pointing cue training condition was used as the reference level. We then conducted post hoc t-tests to compare mean target fixation proportions between time bins.

Analysis 3 identified when looking behaviour during training trials had the biggest effect on accuracy at test. Target fixation data were split into three distinct training phases, each comprising four time bins (Figure 6):

a) Phase 1: before the verbal label in both conditions and after cue occurrence in the early pointing cue condition (−1000 – 0 milliseconds)
b) Phase 2: after the verbal label in both conditions (0 – 1000 milliseconds)
c) Phase 3: after the occurrence of the pointing cue in the late condition (1000 – 2000 milliseconds)

GLMEs were constructed with fixed effects of eye-tracking behaviour per phase and built in the same format as for all other analyses. Only the fixed effects differed; instead of a fixed effect of condition, average fixation proportion to target for each of the training phases (per word and per participant; coded as Phase 1, Phase 2 and Phase 3) was used. An added fixed effect of condition was not included due to a high variance inflation factor between condition and target fixation proportion (>3; Zuur et al., 2010). Interactions between time periods were not tested due to high VIF values within interaction models.

To further understand our results, we also conducted an additional post hoc analysis, Analysis 4, that was not pre-registered. This was split into two models: Analysis 4a identified the effect of word–referent exposure on average fixation proportion during the most crucial phase of training as determined by Analysis 3 using a linear mixed-effects model. Analysis 4b identified the effect of the interaction between fixation proportion during the most crucial phase of training, as determined by Analysis 3, and word–referent exposure on test accuracy using a general linear mixed-effects model. Models were constructed using the same processes as described previously, with the following exceptions. Analysis 4a tested fixed effects of word–referent exposure (number of occurrences as a continuous variable, 1 to 4) and condition ('1' = early, '−1' = late). Analysis 4b tested fixed effects of an interaction between word–referent exposure (number of occurrences as a continuous variable, 1 to 4) and average target fixation proportion during training, immediate accuracy ('1' = correct, '−1' = incorrect) for retention trial analysis and trial type ('1' = immediate, '−1' = retention) for overall accuracy. In addition, for Analysis 4b, a random effect of test trial number was included.

Results and discussion

Analysis 1: The effect of pointing cue timing on behavioural response The final random-effects structure included by-participant random intercepts with slopes

Table 2. Study 2, Analysis 1: The effect of pointing cue timing on behavioural response – best-fitting general linear model results predicting trial accuracy by pointing cue condition

Immediate trial accuracy				
Fixed effect	estimate	*SE*	*z*-value	*p*-value
(intercept)	1.03	0.37	2.75	.006
Pointing cue condition	0.26	0.13	2.00	.045
Retention trial accuracy				
(intercept)	0.14	0.30	0.46	.646
Pointing cue condition	0.47	0.17	2.75	.006
Immediate accuracy	1.29	0.13	9.73	<.001
Overall accuracy				
(intercept)	0.89	0.40	2.20	.028
Pointing cue condition	0.38	0.12	3.09	.002
Trial type	0.21	0.07	2.99	.003
Pointing cue: trial type	−0.15	0.07	−2.22	.026

for condition and by-target word random intercepts across all models. Random intercepts by-target object and by-test order did not converge, and random slopes of condition for all other random intercepts also did not converge despite using allFit(); these had the lowest variance so were removed. The results, presented in Table 2 and Figure 4, replicated those of Study 1. Participants again performed above chance in all conditions. Participants were more accurate at test on words learnt when the pointing cue occurred 1 second before label utterance (rather than 1 second after) across immediate trials (model fit: $\chi^2(1) = 4.28$, p =.038). Study 2 also demonstrated an additional effect of condition on retention trials where Study 1 did not: a model that included fixed effects of condition and immediate accuracy provided the best fit for retention test trial data (model fit: $\chi^2(1) = 111.18$, p <.001). Participants were more likely to respond accurately on retention trials for words learned in the early pointing cue condition (p =.006) and, as per Study 1, more likely to respond accurately for words that were correctly disambiguated in immediate test trials (p <.001).

The best-fitting model predicting overall accuracy included fixed effects of pointing cue condition, trial type and an interaction between pointing cue condition and trial type (model fit: $\chi^2(1) = 4.85$, p =.028). This model showed that participants were more likely to respond accurately in immediate trials than retention trials (p =.003), more likely to respond accurately in the early cue condition than the late cue condition (p =.002), and the interaction demonstrated that learners were more likely to respond accurately in retention trials for words learnt in the early compared to late pointing cue condition (p =.026). Only three of the 20 participants reported noticing a difference between pointing cue conditions at debrief – again suggesting that the difference in performance appeared to be independent of any conscious manipulation of attention.

Analysis 2: Target fixation proportion during training using GCA Figure 5 shows how mean fixation proportion to target, foil and cue alters across trial time by condition (using *geom_smooth* in the *ggplot2* package, local polynomial regression fitting, Wickham, 2016) in R [v1.1.463]. In the early pointing condition, participants looked predominantly at the target with a peak around word utterance, but began to look at the foil towards the end of the trial. In the late pointing condition, fixations at the beginning of the trial were split roughly equally between target and foil, but participants began to discriminate between target and foil after word utterance, with fixation to target rising after the pointing cue.

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

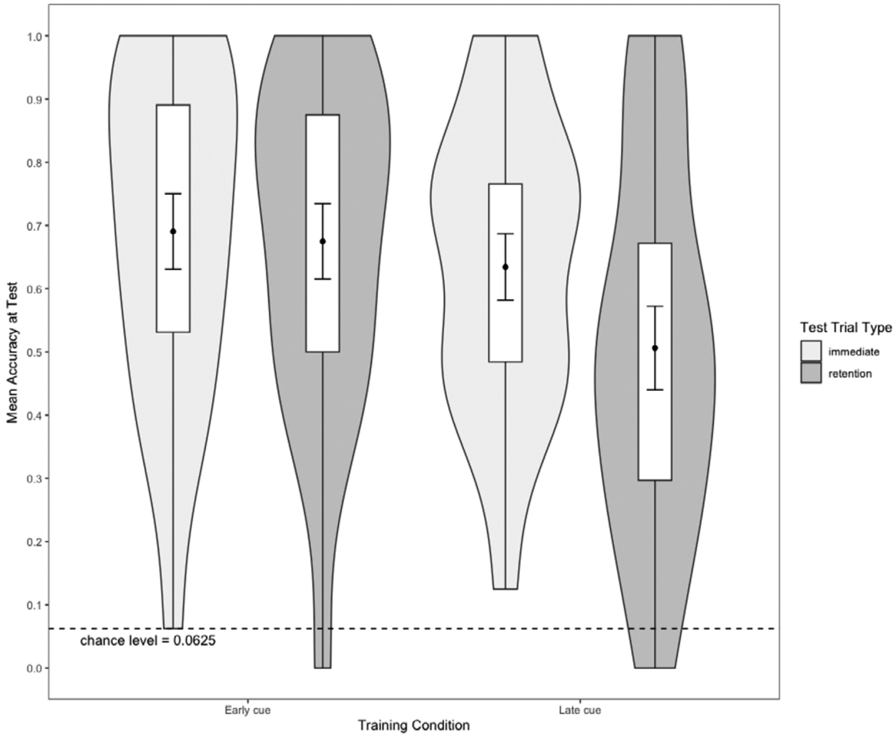

Figure 4. Study 2, Analysis 1: The effect of pointing cue timing on behavioural response – average accuracy across testing trials with standard error bars of all participants, grouped by pointing cue condition and trial type.

The GCA model and data fits are shown in Figure 6,[2] with Table 3 showing fixed-effect parameter estimates and standard error (*p*-values estimated using normal approximation for *t*-values). The overall time course of mean target fixations was best captured with a third-order (cubic) orthogonal polynomial (model fit: $\chi^2(1) = 20.22$, p <.001). The effect of condition improved model fit on the intercept and all time terms (all p <.001). The GCA analysis indicated that target fixation proportion was significantly different between the two conditions, with participants exhibiting a mirrored effect (Figure 6): participants in the early pointing cue condition looked longer at the target at the beginning of trials and decreased their fixation over the duration of trials, whilst participants in the late condition looked less at the target at the beginning of trials and increased their fixation over the duration of trials. To further test where differences between the early and late condition were significant, a series of post hoc independent samples two-tailed *t*-tests for each time bin were carried out. These reflected the same pattern as the GCAs; the *t*-tests demonstrated a significant difference at almost all time bins (8 out of 11 time bin differences were p <.001; Table 4).

[2] Due to technical issues, some data at the beginning of the trial were lost. The drop in fixation proportion to target at time bin 8 (2000 ms) in the late condition was likely due to cue appearance, but this was not captured by a quartic orthogonal polynomial.

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

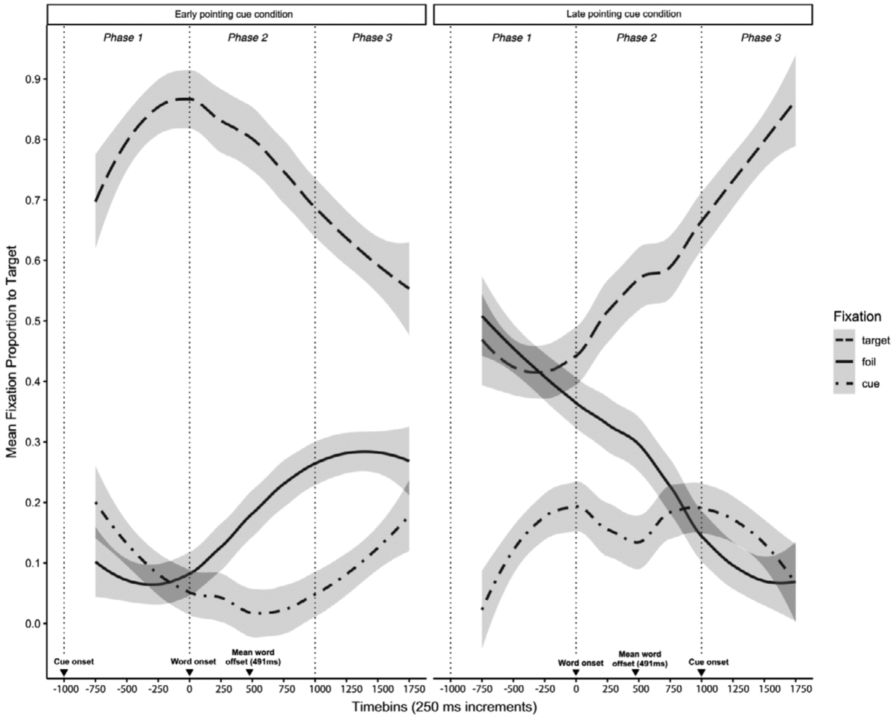

Figure 5. Study 2, Analysis 2: Target fixation proportion during training using GCA – mean fixation proportion (aggregated across all participants and trials) during training in 250ms time bins by pointing cue condition. Grey shaded areas indicate 95% confidence intervals. Phase 1 = after pointing cue in early condition and before word occurrence in both conditions; Phase 2 = after word onset; and Phase 3 = after pointing cue in late condition. Note that as this figure shows aggregated mean fixation proportion across participants and trials per condition, looks to cue in the late pointing condition prior to word occurrence likely stem from participants expecting the cue to appear from previous within-condition trials.

In line with our hypothesis, participants were more likely to fixate on the target before and during word utterance in the early compared to the late condition. However, the increase in target fixation prior to cue onset over trials in the late pointing cue condition demonstrates that, over multiple exposures to word–referent pairs, participants could identify the correct target prior to the cue's appearance. The cue in the late pointing condition thus appeared to act as a confirmation of a referent, whereas in the early pointing condition, the cue appeared to act as a predictor of the referent prior to label occurrence. We then assessed how these patterns during training might have affected participants' performance at test.

Analysis 3: When does target fixation during training predict word learning accuracy? The final random-effects structure included by-participant random intercepts with slopes for condition and by-target word random intercepts, across all models. Random intercepts by target object and by test order did not converge, and random slopes of condition for all other random intercepts also did not converge despite using allFit(); these had the lowest variance so were removed. Results of the models are reported in Table 5. There was a significant effect of fixation proportion to target in Phase 2 (after verbal label in both conditions) on immediate trial accuracy

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

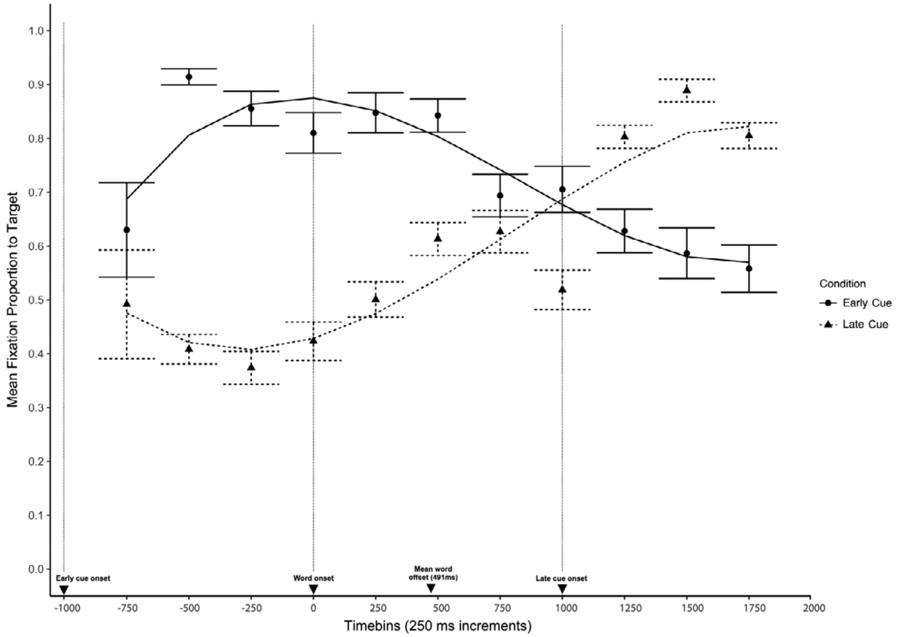

Figure 6. Study 2, Analysis 2: Target fixation proportion during training using GCA – GCA showing mean fixation proportion to target in 250ms time bins, by pointing cue condition. Data points indicate mean and standard error bars for target fixation proportion, aggregated across all participants and trials. Lines indicate model fit.

Table 3. Study 2, Analysis 2: Target fixation proportion during training using GCA – results of GCA of mean target fixation proportion – estimates of time terms between pointing cue condition and model comparison of best-fitting model

	Early cue condition				Late cue condition			
Term	estimate	SE	*t*-value	*p*-value	estimate	SE	*t*-value	*p*-value
(intercept)	0.73	0.02	43.74	<.001	−0.15	0.02	−9.16	<.001
Linear	−0.26	0.06	−4.37	<.001	0.75	0.05	13.95	<.001
Quadratic	−0.20	0.05	−4.19	<.001	0.34	0.05	6.39	<.001
Cubic	0.14	0.05	2.99	<.001	−0.24	0.05	−4.56	<.001

	Model comparisons	
Term	χ^2(df)	*p*-value
(intercept)	45.49(1)	<.001
Linear	137.74(1)	<.001
Quadratic	36.38(1)	<.001
Cubic (full)	20.22(1)	<.001

(model fit: $\chi^2(1) = 6.10$, p =.014). There was also a significant effect of fixation proportion in Phase 2 and immediate test trial accuracy when testing retention trial accuracy (model fit: $\chi^2(1) = 109.12$, p <.001). Finally, there was a significant effect of fixation proportion to target in Phase 2 (after verbal label in both conditions) and trial

Table 4. Study 2, Analysis 2: Target fixation proportion during training using GCA – post hoc t-tests comparing mean target fixation proportion at 250 ms time bins by pointing cue condition

	Early cue		Late cue		Comparison			
Time bin, *ms*	*M*	*SE*	*M*	*SE*	*t*-value*(df)*	95% CI	*p*-value	Cohen's *d*
–750	0.75	0.07	0.82	0.07	−0.72 (25.57)	−0.28, 0.13	.478	−0.27
–500	0.91	0.01	0.41	0.02	16.16 (29.32)	0.44, 0.57	<.001	5.11
–250	0.86	0.03	0.37	0.03	10.89 (37.90)	0.39, 0.57	<.001	3.44
0 (word onset)	0.81	0.04	0.42	0.04	7.43 (37.89)	0.29, 0.49	<.001	2.35
250	0.85	0.04	0.50	0.03	7.01 (37.39)	0.25, 0.45	<.001	2.22
500	0.84	0.03	0.61	0.03	5.28 (38.00)	0.14, 0.32	<.001	1.67
750	0.69	0.04	0.63	0.04	1.20 (38.00)	−0.05, 0.18	.238	0.38
1000	0.71	0.04	0.52	0.04	3.31 (37.16)	0.07, 0.30	.002	1.05
1250	0.63	0.04	0.80	0.02	−3.81 (28.90)	−0.27, −0.08	<.001	−1.20
1500	0.58	0.04	0.89	0.02	−5.87 (26.25)	−0.41, −0.20	<.001	−1.85
1750	0.56	0.04	0.80	0.02	−4.93 (29.4)	−0.35, −0.14	<.001	−1.56

Table 5. Study 2, Analysis 3: When does target fixation during training predict word learning accuracy? Best-fitting general linear model results predicting trial accuracy with fixed effects of target fixation proportion during training

Immediate accuracy				
Fixed effect	estimate	*SE*	*z*-value	*p*-value
(intercept)	0.09	0.42	0.23	.822
Target fixation proportion (Phase 2)	1.13	0.45	2.54	.011
Retention accuracy				
(intercept)	−0.40	0.53	−0.75	.452
Target fixation proportion (Phase 2)	1.13	0.55	2.05	.041
Immediate accuracy	1.28	0.13	9.64	<.001
Overall accuracy				
(intercept)	0.93	0.50	0.19	0.85
Target fixation proportion (Phase 2)	1.07	0.36	2.97	.003
Trial type	0.22	0.07	−3.21	.001

Note: Only fixation proportion during training Phase 2 (after the label utterance) was a significant predictor of accuracy.

type when testing overall accuracy (model fit: $\chi^2(1) = 10.19$, p =.001). GLMEs fitted for Phases 1 (before verbal label in both conditions, after cue in early condition) and 3 (after cue in late condition) did not identify significant effects of average fixation proportion to target on accuracy in any of the test trials, indicating that looking behaviour during training before the word occurred and after the cue occurred in the late pointing cue condition did not influence performance at test. An additional analysis testing total fixation proportion during the trial across all time periods (see OSF) did not yield any significant predictive effects on accuracy. Thus, fixation to target during Phase 2 (after verbal label in both conditions) immediately after word utterance was the crucial time period for accurate learning.

Analysis 4a: Does word–referent exposure influence fixation to target? Next, we analysed fixation proportion to target during Phase 2 (after verbal label in both conditions), taking into account the number of times participants had been exposed to the word–referent pair. Each word–referent pairing had four exposures during training, and the expectation of cross-situational word learning is that participants successfully learn word–referent pairs after multiple exposures.

Table 6. Study 2, Analysis 4a: Does word–referent exposure influence fixation to target? Best-fitting linear model results predicting target fixation proportion by pointing cue condition and word–referent exposure

Fixed effect	estimate	SE	*t*-value	*p*-value
(intercept)	0.69	0.02	27.78	<.001
Word–referent exposure	−0.002	0.006	−0.44	0.66
Pointing cue condition	0.23	0.02	8.11	<.001
Word–referent exposure: condition	0.04	0.006	−7.96	<.001

The final model had by-participant random intercepts of by-target word random intercept with slopes of condition and a random intercept of target object. A slope of condition did not converge for target object despite using allFit() and was removed. For average fixation proportion during Phase 2 (after verbal label), the best fitting model included significant fixed effects of word–referent exposure, condition, and an interaction between the two ($\chi^2(1) = 62.41$, p <.001; Table 6). This indicated that mean target fixation proportion increased with exposure for the late pointing condition but decreased for the early condition. Figure 7 illustrates how participants in the early pointing cue condition looked less at the target during label utterance as word–referent exposure increased, whereas participants in the late pointing cue condition exhibited the opposite pattern, looking more at the target during label

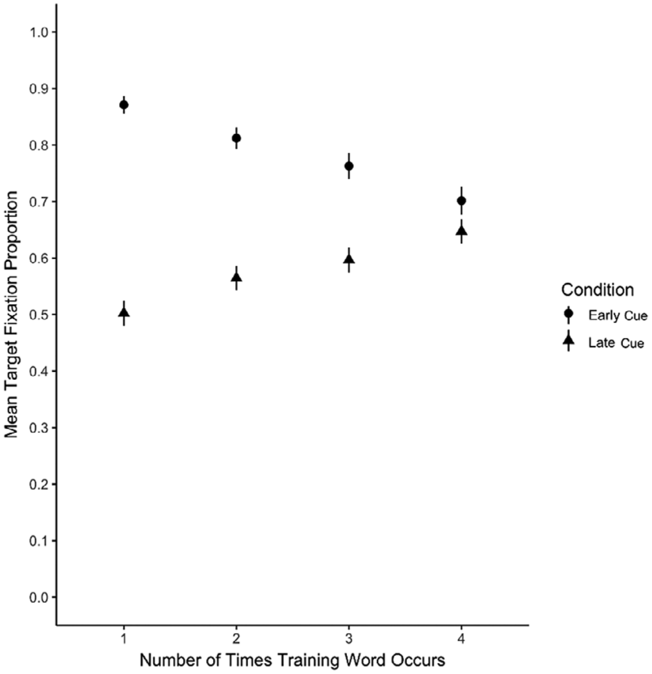

Figure 7. Study 2, Analysis 4a: Does word–referent exposure influence fixation to target? Mean target fixation proportion (aggregated across all participants, all words, and all trials) and standard error bars during label utterance (Phase 2 [after verbal label in both conditions]; Figure 5) by word–referent exposure and pointing cue condition.

utterance after multiple exposures. These profiles likely reflect different learning strategies over time between the two conditions.

Analysis 4b: Does the interaction between word–referent exposure and target fixation proportion during training affect accuracy at test? All models had random intercepts of participant, target word, target object and test trial number; test order did not converge despite using allFit() and had the lowest variance so was removed. For immediate accuracy, the interaction between word–referent exposure and average target fixation proportion during Phase 2 of training (model fit: $\chi^2(1) = 3.94$, p =.047; Table 7) indicated that participants were more accurate if they fixated longer on the target with increasing word–referent exposures (p =.045).

For retention data, the model contained a fixed effect of immediate accuracy at test and the interaction between word–referent exposure and average target fixation proportion during Phase 2 of training (model fit: $\chi^2(1) = 362.21$, p <.001; Table 7). This indicated that participants were more likely to respond accurately in retention test trials if they had responded correctly on the corresponding immediate test trial (p <.001), and they were more likely to respond accurately overall if they fixated longer on the target with increasing word–referent exposures during training (p =.015).

For target fixation data predicting overall accuracy, fixed effects of trial type and average target fixation proportion during first word–referent exposures were found (model fit: $\chi^2(1) = 54.90$, p <.001, Table 7). Participants were more likely to respond accurately in immediate trials than retention trials (p <.001) and were more likely to respond accurately overall if they fixated longer on the target with increasing word–referent exposures during training (p <.001).

Together with the GCA (Analysis 2), Analyses 4a and 4b indicate that participants learned words more accurately when the pointing cue occurred 1 second before the word, rather than 1 second after, primarily because they exhibited higher target fixation during the period surrounding label utterance. Furthermore, from the first exposures to word–referent pairs, participants already demonstrated higher target fixation proportion during label utterance in the early pointing cue condition (Figure 7), which predicted higher accuracy at test.

Table 7. Study 2, Analysis 4b: Does average target fixation proportion by word–referent exposure during training affect accuracy? General linear model results showing interaction between average target fixation proportion during Phase 2 and word–referent exposure on accuracy at test

Immediate trial accuracy				
Fixed effect	estimate	*SE*	*z*-value	*p*-value
(intercept)	1.16	0.52	2.25	.024
Word–referent exposure * target fixation proportion in Phase 2	0.09	0.04	2.01	.045
Retention trial accuracy				
(intercept)	0.08	0.46	0.18	.861
Immediate accuracy	1.34	0.08	16.79	<.001
Word–referent exposure * target fixation proportion in Phase 2	0.12	0.05	2.44	.015
Overall accuracy				
nintercept)	0.84	0.49	1.73	.083
Test trial type	0.25	0.04	6.61	<.001
Word–referent exposure * target fixation proportion in Phase 2	0.10	0.03	3.37	<.001

General discussion

The contribution of cross-situational statistics to word learning is well documented, but the mechanisms through which environmental cues facilitate cross-situational word learning are not well understood. In this study, we showed how studies of pointing cue use in word learning can align with the long-standing tradition of studies exploring visual attentional cueing. We highlighted how the effectiveness of pointing cues in language learning is determined by the timing of endogenous cue reorientation, potentially tailored to exploit the coordination of attention at the moment of labelling to optimise word learning.

Study 1 demonstrated that early pointing cues under referential ambiguity yield superior learning to late pointing cues, indicating that *when* cues occur in relation to label utterance has a direct influence on word–object mapping accuracy. Study 2 replicated these results and confirmed that this superior learning was due to the early cue directing visual attention to the target referent during label utterance. Both studies demonstrated that immediate referent selection accuracy was a predictor of later retention accuracy and that this effect was a stronger predictor of retention than any manipulation of pointing cue condition – indicating that the dynamics of referent selection are vital to subsequent retention (McMurray et al. 2012; Yu & Smith, 2012). These results are consistent with studies that examine the time course of how, and when, endogenous cues orient attention to objects (Berger et al., 2005; Yoshida & Burling, 2012). However, these effects have not previously been merged with word learning, and our study investigating cross-situational word learning with different temporal arrangements of pointing cues provides an example of how endogenous cueing during similar word learning tasks may interplay with speech to support learning.

Studies that examine pointing cues under naturalistic settings have also indicated different effects of temporal order for cued attention during word learning. In naturalistic settings, gestures appear more frequently before, rather than after, speech (Bergmann et al., 2011; Donnellan et al., 2022). Frank et al. (2013) found that pointing gestures were used to introduce new topics and tended to be used at the beginning of discourses about objects during semi-naturalistic mother–infant interactions. Regarding language acquisition, children also looked at an object less as it was talked about more, mirroring the pattern of target fixation behaviour in the early pointing condition (Figure 7). Furthermore, novel words are learnt by infants most accurately when they are centred in view and largest in size during label utterance (Pereira et al., 2014), and children's attention to referents is highest during, and just after, label utterance in naturalistic mother–infant interaction videos (Trueswell et al., 2016).

In adult communication, adjusting the timing of gesture and naming has been shown to affect processing. Nirme et al. (2020) found that moving the gesture later affected judgements of naturalness for communicative situations when the gesture overlapped with a pause, Habets et al. (2011) found enhanced semantic integration when iconic gestures occurred before rather than simultaneously with naming, and Cavicchio and Busà (2023) found that iconic gestures resulted in quicker identification of an action for non-native speakers. Thus, it has been established that gesture-naming timing is critical for effective processing of potential word learning situations, and in our study, we showed how manipulating this timing can affect both referent selection and retention of novel words. It appears that gesturing before speaking is beneficial for learning, and so the temporal ordering may not merely be a

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

consequence of production constraints, but may instead be meeting the contingent need of the learner in acquiring new words (Holler & Levinson, 2019). Overall, the benefit of endogenous cues to cross-situational word learning appears to be mediated by quality rather than quantity: *when* a learner fixates upon a target referent may matter more than *how much* they fixate on a target referent. As Study 2 demonstrated, simply looking at a target prior to label utterance is not sufficient to improve learning. Our analyses showed that target fixation prior to word occurrence during training (Study 2, Phase 1, before verbal label in both conditions, after cue in early condition) did not predict accuracy at test, despite participants in the early pointing cue condition having more time to fixate on the target before label utterance. Rather, the predictive value of early pointing cues leads to a learner fixating upon the correct referent when label utterance occurs from the very first exposures to novel words, and this may confer an advantage in overall resilience of forming word–referent mappings. This difference is apparent even when varying the relative timing of the pointing cue to label utterance by only 1 second, as participants performed significantly less accurately in the late pointing cue condition across both studies. Consistent with these findings, MacDonald et al. (2017) found that adult learners still tracked a single hypothesis and spent less time on alternative word–referent pairs when a gaze cue to a target object was present (as opposed to absent) even after being given the same amount of time to visually inspect the objects during cross-situational training in both conditions. The authors suggested this was because gaze increased opportunity to maintain attention on the target referent.

When examining adult cross-situational word learning, Yu et al. (2012) found that strong and weak learners exhibited a pattern of looking behaviour that only began to differ around the middle stages of their training, likely due to gradual aggregation of statistical co-occurrences over time. This is consistent with our results in Study 2, where participants in the late pointing cue condition increasingly fixated on the target over trials with increased word–referent exposure (Figure 7). However, during the early pointing cue condition, participants began trials by fixating upon the target because they were cued towards it. In Yu et al. (2012), strong learners had increased attention to the referent towards the end of trials, rather than the beginning. With an early pointing cue, learners in Studies 1 and 2 may have been provided with a shortcut that enabled them to direct their attention towards the target from the very first exposure, resulting in more accurate performance at test. This is in line with the eye-tracking data showing that fixations to target in the first exposures to word–referent pairs, rather than the last exposures, were predictive of word learning accuracy.

Increased looking and attention to the referent when an unfamiliar label is uttered may benefit learning by increasing the initial strength of association between label and target, which then builds up gradually over multiple situations. The results of Study 2 that demonstrated high target fixation during first exposures to words during the early cue condition support this interpretation. Reducing attention to foil objects may also decrease the likelihood of forming spurious word–object associations, supporting learning of precise word–referent mappings intended by the speaker (e.g. Yu & Ballard, 2007; McMurray et al., 2012). Associative models of word learning (MacWhinney, 2005; McMurray et al., 2012; Yu & Smith, 2012) contend that a learner builds up weights on associations between labels and foils, as well as targets. We show that directing attention to the target with a pointing cue *prior* to the word being spoken may prevent the learner from making false associations between a foil and the label, limiting the formation of competing associations. However, cues that

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

occur *after* the word is spoken do not appear to prevent some competing false label–foil associations from being formed, resulting in reduced accuracy at test relative to the early pointing cue condition. Applying a cue to indicate the target referent after the label has been spoken does not provide the same quality of information as when attention is already drawn to the target referent prior to the label being spoken. Therefore, the presence of cues is not the only factor that promotes optimal word learning – the contiguity of those cues in relation to labelling must also be effective.

Another benefit for learning conferred by pointing cues preceding labelling concerns prediction. Ramscar et al. (2010) manipulated the ordering of objects and labels during word learning in adults and found that learning was more accurate when objects were presented prior to labels, rather than when labels preceded objects. This may be due to differences in the informativeness of labels and objects as conditioning cues; when objects occur prior to labels, learners must process several object features as distinctive cues that compete for relevance when predicting the label. However, when learners are exposed to the label first, this provides a far more constrained source of information to predict objects from. Consistent with this, learners in our study appeared to use the early pointing cue as a predictor of the referent, whereas in the late pointing condition, the cue may have simply confirmed the participant's assumption, resulting in a weaker prediction for the learner.

An alternative explanation for why early pointing cues facilitate more accurate word learning is that participants are more familiar with this ordering of gesture and speech, assuming that the majority of gestures precede naming referents in naturalistic communication. Under this view, early gesturing is an accidental property of the communicative environment, and learners become attuned to this. Though this is a possibility, applying equally to the interpretation of previous studies adjusting gesture and speech ordering (e.g. Cavicchio and Busà, 2023; Habets et al., 2011; Nirme et al. 2020; Trueswell et al., 2016), we believe this is less likely than our favoured interpretation that a positive effect of ordering exists due to cognitive mechanisms integrating speech and visual information. This is because the eye-tracking data demonstrate how the learner explores the scene, providing evidence not only of a learning boost but also how patterns of looking to the target at the point of naming is beneficial for learning, and precipitated by the gesture. Being used to the relative order of speech and gesture would not necessarily result in these subtle patterns of looking. As our results are consistent with the general attentional cueing literature (Berger et al., 2005; Brignani et al., 2009; Hauer and Macleod 2006; Shepherd & Müller, 1989), rather than being specific to situations involving gestures and naming production, we contend that it is preferable to explain results with broader cognitive theories than more specific theories. Also, as only seven of 40 participants noticed a difference in cue timing between conditions, violation of ordering familiarity did not influence conscious processing for the majority of our sample. In Nirme et al. (2020), judgements of naturalness were only impacted when iconic gestures overlapped with a pause, and manipulations of 500ms before or after naming did not otherwise influence judgements. For our study, moving the gesture to 1s before or after naming similarly resulted in no difference in participants' perceptions. This suggests that learners have little meta-awareness of the context and process surrounding word learning itself and likely have little explicit control over how cues, labels and referents are sequenced in communicative situations. Nonetheless, in naturalistic settings, speakers use gesture and speech in ways that are beneficial to learning, despite being unaware of their temporal contiguity.

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

Limitations

This study has a number of limitations. Firstly, our use of a finger and hand as a pointing cue may raise the question of whether an arrow might yield the same results. However, the advantage of using a finger pointing cue is simply that they play a more prominent role in naturalistic language learning than arrows. Whether visual attention grabbers such as lights and arrows outweigh social cues, such as head turn and eye gaze, has been addressed elsewhere (e.g. see Axelsson et al., 2012; Hartley et al., 2020; Wu & Kirkham, 2010). A similar limitation concerns the naturalism of a static photograph of a human hand with an index point as a pointing cue; although easily recognisable by human learners, this was not as naturalistic as having an actor pointing at different objects. As described previously (see 'Current Study' section), this static cue was chosen to afford more control over precise timing on informative value during word–referent mapping, as compared to dynamic video stimuli, where informative value occurs over time (Donnellan et al., 2022). Although we believe it is unlikely that video stimuli would produce vastly different results, we do recommend that future studies examine the role of more naturalistic social and non-social cues under the same conditions of referential ambiguity, or even weigh different types of social cues against one another.

Secondly, despite pointing cues being reliable indicators of referents, they do not occur in many naturalistic learning situations, such as during language acquisition. In their semi-naturalistic mother–infant video corpus, Frank et al. (2013) report that pointing cues had a recall value of 10%, whereas maternal eye gaze had a recall value of 36%. In Trueswell et al. (2016), highly informative vignettes that contained maternal gestures were rare, and in Iverson et al. (1999), mothers only used pointing cues during word learning 15% of the time. Pointing cues are likely only one of several cues that can support cross-situational word learning.

Thirdly, we did not intermix early and late cues within the same block, and instead attempted to minimise variation by blocking the task by cue timing. This blocking may have introduced additional bias into the results if learners were distinctly aware of the difference; however, as noted, the majority of participants failed to notice a difference between the conditions.

A final limitation – and opportunity for further exploration – was that we investigated only two time intervals between gesture and naming: a 1s asynchrony, that was situated between the 2s of Trueswell et al. (2016) and the 500ms and 360ms intervals of Nirme et al. (2020) and Habets et al. (2011), respectively. Testing multiple asynchronies will allow us to determine the precise optimal difference between gesture and naming to support learning, potentially closer to the 370ms interval between gesture and naming found in naturalistic discourse (Donnellan et al., 2022). Evidence for a quantitative effect on ordering of gesture and naming will enable us to specify more fully the fine-grained learning mechanisms that apply in learning novel words.

Conclusion

These studies offer multiple insights into how pointing cues can facilitate disambiguation of meaning when a learner is faced with referential ambiguity. The value of pointing cues appears to be in compensating for referential ambiguity by providing accurate information about referents. Cues are particularly useful when highlighting

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

referents prior to labels; when a perfectly disambiguating pointing cue occurs before a novel word is spoken, this provides a superior benefit to the learner than when a pointing occurs after a novel word. These temporal effects are consistent with how pointing cues interoperate with speech in naturalistic studies and show how attention literature regarding endogenous cues is also applicable to cross-situational word learning. The studies presented here provide a controlled setting that demonstrates how and when pointing can support cross-situational statistical learning, and translate well-investigated attention and memory phenomena into effects of cueing during word learning.

Data availability statement. All data and code supporting this study are available on OSF at https://osf.io/2m9pe/?view_only=9d64688d03d84704aa5f2e8f8eb34dc9.

Acknowledgments. This research was funded by the Leverhulme Trust (RWC, Leverhulme Trust Doctoral Scholar [DS-2014-014]) and the International Centre for Language and Communicative Development (LuCiD) at Lancaster University funded by the Economic and Social Research Council UK (PM and CH [ES/L008955/1]).

Competing interest. There are no competing interests to declare. All pre-registrations, data, code and additional outputs can be found on the OSF via this link: https://osf.io/2m9pe/?view_only=9d64688d03d84704aa5f2e8f8eb34dc9.

References

Axelsson, E. L., Churchley, K., & Horst, J. S. (2012). The right thing at the right time: Why ostensive naming facilitates word learning. *Frontiers in Psychology*, 3, 1–8. https://doi.org/10.3389/fpsyg.2012.00088

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Berger, A., Henik, A., & Rafal, R. (2005). Competition between endogenous and exogenous orienting of visual attention. *Journal of Experimental Psychology: General*, 134(2), 207–221. https://doi.org/10.1037/0096-3445.134.2.207

Bergmann, K., Aksu, V., & Kopp, S. (2011). *The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011).* https://pub.uni-bielefeld.de/record/2392953.

Beun, R.-J., & Cremers, A. H. M. (1998). Object reference in a shared domain of conversation. *Pragmatics and Cognition*, 6(1–2), 121–152.

Bhat, A. A., Spencer, J. P., & Samuelson, L. K. (2022). Word-Object Learning via Visual Exploration in Space (WOLVES): A neural process model of cross-situational word learning. *Psychological Review*, 129(4), 640–695 https://doi.org/10.1037/rev0000313

Brignani, D., Guzzon, D., Marzi, C.A., Miniussi, C. (2009) Attentional orienting induced by arrows and eye-gaze compared with an endogenous cue. *Neuropsychologia*, 47*(2)*, 37–381. https://doi.org/10.1016/j.neuropsychologia.2008.09.011

Carey, S., & Bartlett, E. (1978) Acquiring a single new word. *Papers and Reports on Child Language Development*, 15, 17–29.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, 110(28), 11278–11283. https://doi.org/10.1073/pnas.1309518110

Cavicchio, F., & Busà, M. G. (2023). The role of representational gestures and speech synchronicity in auditory input by L2 and L1 speakers. *Journal of Psycholinguistic Research*, 52, 1721–1735.

Chu, M., & Hagoort, P. (2014). Synchronization of speech and gesture: Evidence for interaction in action. *Journal of Experimental Psychology: General*, 143(4), 1726–1741. https://doi.org/10.1037/a0036281

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

Donnellan, E., Özder, L. E., Man, H., Grzyb, B., Gu, Y., & Vigliocco, G. (2022). Timing relationships between representational gestures and speech: A corpus based investigation. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, 2052–2058).

Dunn, K. J., Frost, R. L. A., & Monaghan, P. (2024). Infants' attention during cross-situational word learning: Environmental variability promotes novelty preference. *Journal of Experimental Child Psychology*, 241, 105859.

Fitneva, S. A., & Christiansen, M. H. (2011). Looking in the wrong direction correlates with more accurate word learning. *Cognitive Science*, 35(2), 367–380. https://doi.org/10.1111/j.1551-6709.2010.01156.x

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, 9(1), 1–24. https://doi.org/10.1080/15475441.2012.707101

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, 73(2), 135–176. https://doi.org/10.1016/S0010-0277(99)00036-0

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Harvard University Press.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, 28(1), 99–108. https://doi.org/10.1037/0012-1649.28.1.99

Habets, B., Kita, S., Shao, Z., Özyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience*, 23(8), 1845–1854. https://doi.org/10.1162/jocn.2010.21462

Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive Psychology*, 53(4), 310–344. https://doi.org/10.1016/j.cogpsych.2006.04.003

Hartley, C., Bird, L.-A., & Monaghan, P. (2020). Comparing cross-situational word learning, retention, and generalisation in children with autism and typical development. *Cognition*, 187, 104265. https://doi.org/10.1016/j.cognition.2019.03.001

Hauer, B. J. A., & Macleod, C. M. (2006) Endogenous versus exogenous attentional cuing effects on memory. *Acta Psychologica*, 122*(3)*, 305–320, https://doi.org/10.1016/j.actpsy.2005.12.008

Holler, J., & Levinson, S. (2019) Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8), 639–652.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R., Brand, R. J., & Brown, E. (2000) Breaking the language barrier: An Emergentist Coalition Model for the origins of word learning. *Monographs of the Society for Research in Child Development*, 65(3), I-135.

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, 48(4), 1393–1409. https://doi.org/10.3758/s13428-015-0647-3

Iverson, J. M., Capirci, O., Longobardi, E., & Caselli, M. C. (1999) Gesturing in mother-child interactions. *Cognitive Development*, 14(1), 57–75. https://doi.org/10.1016/S0885-2014(99)80018-5

Jonides, J. (1981) Towards a model of the mind's eyes' movement. *Canadian Journal of Psychology*, 34(2), 103–112, https://doi.org/10.1037/h0081031

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, 24, 145–167.

Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., & Wachsmuth, I. (2006). Deictic object reference intask-oriented dialogue. *Situated Communication*, 166, 135–189. https://doi.org/10.1515/9783110197747.155

Levelt, W. J., Richardson, G., & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language*, 24(2), 133–164.

MacDonald, K., Yurovsky, D., & Frank, M. C. (2017). Social cues modulate the representations underlying cross-situational learning. *Cognitive Psychology*, 94, 67–84. https://doi.org/10.1016/j.cogpsych.2017.02.003

MacWhinney, B. (2005). Extending the competition model. *International Journal of Bilingualism*, 9(1), 69–84. https://doi.org/10.1177/13670069050090010501

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology*, 20(2), 121–157. https://doi.org/10.1016/0010-0285(88)90017-5

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, 119(4), 831–877. https://doi.org/10.1037/a0029872

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, 92, 350–371.

McNeill, D. (2000). *Language and gesture*. Cambridge University Press.

Mirman, D. (2014). *Growth curve analysis and visualization using R*. CRC Press, Taylor & Francis Group, LLC.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59(4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. *Topics in Cognitive Science*, 9(1), 21–34. https://doi.org/10.1111/tops.12239

Monaghan, P., Brand, J., Frost, R. L. A., & Taylor, G. (2017). Multiple variable cues in the environment promote accurate and robust word learning. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 817–822.

Monaghan, P., & Mattock, K. (2012). Integrating constraints for learning word-referent mappings. *Cognition*, 123(1), 133–143. https://doi.org/10.1016/j.cognition.2011.12.010

Monaghan, P., Mattock, K., Davies, R. A. I., & Smith, A. C. (2015). Gavagai is as gavagai does: Learning nouns and verbs from cross-situational statistics. *Cognitive Science*, 39(5), 1099–1112. https://doi.org/10.1111/cogs.12186

Nirme, J., Haake, M., Gulz, A., & Gullberg, M. (2020). Motion capture-based animated characters for the study of speech–gesture integration. *Behavior Research Methods*, 52, 1339–1354.

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic Bulletin & Review*, 21(1), 178–185. https://doi.org/10.3758/s13423-013-0466-4

Pierce, J. W., & MacAskill, M. R. (2018). *Building experiments in PsychoPy*. Sage.

Posner, M. I. (1981). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, 32(1), 3–25. https://doi.org/10.1080/00335558008248231

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6), 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

Roembke, T., & McMurray, B. (2016). Observational word learning: Beyond propose-but-verify and associative bean counting. *Journal of Memory and Language*, 87, 105–127. https://doi.org/10.1016/j.jml.2015.09.005

Shepherd, M. & Müller, H. J. (1989) Movement versus focusing of visual attention. *Perception & Psychophysics*, 46(2), 146–154. https://doi.org/10.3758/BF03204974

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1–2), 39–91.

Smith, K., Smith, A. D. M., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, 35(3), 480–498. https://doi.org/10.1111/j.1551-6709.2010.01158.x

Trueswell, J. C., Lin, Y., Armstrong, B., Cartmill, E. A., Goldin-Meadow, S., & Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent–child interactions. *Cognition*, 148, 117–135. https://doi.org/10.1016/j.cognition.2015.11.002

Von der Malsburg, T. (2015). saccades: Detection of fixations in eye-tracking data. https://cran.r-project.org/package=saccades

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org.

Wu, R., & Kirkham, N. Z. (2010). No two cues are alike: Depth of learning during infancy is dependent on what orients attention. *Journal of Experimental Child Psychology*, 107(2), 118–136. https://doi.org/10.1016/j.jecp.2010.04.014

Yoshida, H., & Burling, J. (2012) Highlighting: A mechanism relevant for word learning. *Frontiers in Psychology*, 3(262), 1–12. https://doi.org/10.3389/fpsyg.2012.00262

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13), 2149–2165. https://doi.org/10.1016/j.neucom.2006.01.034

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word–referent learning: Prior questions. *Psychological Review*, 119, 21–39. https://doi.org/10.1037/a0026182

Yu, C., Zhong, Y., & Fricker, D. (2012). Selective attention in cross-situational statistical learning: Evidence from eye tracking. *Frontiers in Psychology*, 3, 148. https://doi.org/10.3389/fpsyg.2012.00148

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, 37(5), 891–921. https://doi.org/10.1111/cogs.12035

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

Cite this article: Cheung, R. W., Hartley, C., & Monaghan, P. (2024). Better early than late: the temporal dynamics of pointing cues during cross-situational word learning, *Language and Cognition* 16: 1960–1986. https://doi.org/10.1017/langcog.2024.39

https://doi.org/10.1017/langcog.2024.39 Published online by Cambridge University Press