What Predicts Adult Word Learning in Naturalistic Interactions? A Corpus Study

Francesco Cabiddu[1] (francesco.cabiddu@ucl.ac.uk)
Christopher Edwards[1*] (christopher.edwards.20@ucl.ac.uk)
Harriet Hill-Payne[1] (harriet.hill-payne.22@ucl.ac.uk)
Ed Donnellan[1, 2] (ed.donnellan@warwick.ac.uk)
Yan Gu[1, 3] (yan.gu@essex.ac.uk)
Gabriella Vigliocco[1] (g.vigliocco@ucl.ac.uk)

[1]Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, UK
[2]Department of Psychology, University of Warwick, University Road, Coventry, CV4 7AL, UK
[3]Department of Psychology, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK

Abstract

Alongside the linguistic input, young children leverage multimodal cues (e.g., prosody, gestures) to learn novel words in face-to-face interactions. It is unclear whether multimodal cues play a similar role in adults. Here, we used ECOLANG, a corpus of semi-naturalistic dyadic conversations where English-speaking adults incidentally learned about unknown objects and their names by interacting with a partner who knew those objects. We examined whether multimodal cues (prosodic, indexical, and iconic) predicted learners' ability to learn the objects' names, above and beyond individual differences and linguistic predictors. We found that the number of repetitions of the label predicted word learning. Additionally, learners with lower working memory abilities benefited from speakers producing representational gestures while labelling the unknown objects. We discuss implications for theories of word learning and approaches of situated cognition.

Keywords: adult word learning; multimodal word learning; naturalistic learning; gestures; eye-gaze; prosody.

Introduction

Word learning often occurs in face-to-face interactions, where we learn new words from more knowledgeable people (caregivers, teachers, peers). Social interaction offers rich multimodal inputs, enabling learners to integrate information across channels (e.g., linguistic, prosodic, and visual/gestural). Multimodality is key in theories viewing language as a situated phenomenon (Murgiano et al., 2021; Reggin et al., 2023). These theories propose that cognitive representations are shaped by sensorimotor experiences in the actual physical and social setting in which they occur. Thus, multimodality is seen as central for learning, with implications also for educational approaches (e.g., Macrine & Fugate, 2022; Mathias & Kriegstein, 2023). The role of multimodal cues in word learning has mostly been studied in children. Caregivers' cues, including linguistic (word frequency and length; Cabiddu et al., 2023; Jones et al., 2023), prosodic (word pitch; Shi et al., 2022), indexical (pointing to word referents; Booth et al., 2008), and iconic (gestures depicting word meanings; Vogt & Kauschke, 2017) impact child word learning. Rather than focusing on specific cues, some recent studies have begun to explore how different cues are used together by caregivers in naturalistic settings and which cues predict child learning (Donnellan et al., 2023; Motamedi et al., 2024). This research moves beyond studies that consider cues in isolation and often rely on controlled laboratory experiments thus providing insight into real-world learning.

Such an approach has not been applied to adults. This is notable given that incidental learning in conversation with teachers, colleagues, and others is a common way for adults to learn new words in everyday life. Learning of new words in adulthood is typically assessed experimentally, hence we do know little about which factors play a role in naturalistic interactions. Similarly to the child literature, the experimental studies available indicate that learner's individual differences, linguistic properties, and multimodal cues impact adult word learning.

Regarding **individual differences**, adults with better working memory have been found to be better at learning novel words from artificial languages (Bulgarelli & Weiss, 2021; Martin & Ellis, 2012) and in object-referent mapping tasks (Neveu & Kaushanskaya, 2023). Additionally, vocabulary skills have been shown to influence learning through familiarity with semantically (James et al., 2023) or phonologically similar words in the mental lexicon (Papagno et al., 1991; Papagno & Vallar, 1992; Storkel et al., 2006). Most studies, however, do not involve social interaction. In social interaction, working memory and vocabulary of the learner might play a lesser role (Brandt et al., 2022) because social partners have been found to help learners direct their attention to correct referents (Verga & Kotz, 2017) and facilitate learners' lexical recall by channelling their attention to specific information (Elekes & Sebanz, 2020).

Regarding **linguistic predictors**, it has been shown that adults can leverage the frequency of co-occurrence between word labels and referents in word learning via associative mechanisms (Hendrickson & Perfors, 2019; Vouloumanos, 2008; Yu & Smith, 2007). Word length also has an independent effect, as long words are harder to learn due to increased processing load (Brennan & Cullinan, 1976; Krishnan et al., 2017; Papagno & Vallar, 1992).

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Moreover, the sentence context in which novel words are presented also matters. First, sentence-level concreteness has been shown to aid in sentence processing and recall (Meltzer et al., 2016; Pham & Archibald, 2023; Romani et al., 2008), and might facilitate word learning due to grounding the encoding of a novel word in perception and action (Vigliocco et al., 2009). Second, words presented in more varied lexical contexts are typically recalled, named, and learned faster by adults (Adelman et al., 2006; Johns et al., 2016; Lohnas et al., 2011). Analyses of lexical-semantic networks, where words are connected based on adult semantic relatedness, show that new words are more likely to be integrated into the lexicon when they occur in diverse contexts, thereby establishing more semantic links with other words in the learning input (Hills et al., 2009, 2010). Finally, sentence length has been linked to child vocabulary growth (Anderson et al., 2021; Braginsky et al., 2019). Words embedded in shorter sentences are likely more easily learned due to the facilitation in decoding the target word's meaning. However, given that adults are proficient language users, they might benefit more from longer sentences, which provide richer semantic information related to target words.

Finally, regarding **multimodal predictors**, presenting novel words while performing representational gestures (i.e., gestures depicting characteristics of the object being described) improves adult word learning in free recall and recognition (Kelly et al., 2009; Macedonia et al., 2010; Sánchez-Borges & Álvarez, 2023). There is also evidence that indexical cues, such as the speaker's manipulation of an object while talking about it, gazing at, or pointing to a target object positively influence word-referent mapping in laboratory studies (MacDonald et al., 2017; Kobayashi et al., 2023; Yasuda & Kobayashi, 2022; Yu et al., 2005). These indexical cues offer a non-verbal means of referencing the object, directing the learner's attention and singling out a referent present in the visual field (Bohn & Frank, 2019; McNeill, 1992; Kuhn et al., 2009).

In addition, prosodic cues such as higher pitch and longer word duration (both characteristics of infant-directed language) have been shown to facilitate adult word learning in recognition and memory experiments (Filippi et al., 2014; Golinkoff & Alioto, 1995; Ma et al., 2020; Sommers & Barcroft, 2007). These cues may draw attention to the target word, as they deviate from the speaker's average acoustic profile (Golinkoff & Alioto, 1995).

Thus, previous studies show that different types of factors can affect word learning in adults, however they do not allow us to establish their relevance when considered together in naturalistic settings because learning is assessed in laboratory studies that manipulate a single cue (e.g., prosody) or only a small number of predictors (e.g., word's frequency and referential gaze). Here, we use a corpus that mimics a learning dynamic commonly found in everyday life: a more knowledgeable person names and describes objects that are unknown to their conversational partner, but teaching and learning is not the explicit objective of the interaction. We examine the concurrent role of different multimodal cues while statistically controlling for the contribution of established individual and linguistic predictors.

The Current Study

We used the ECOLANG (Adult) corpus (Gu et al., submitted), comprising 33 adult dyadic conversations between a "teacher" (more knowledgeable person asked to freely talk about objects that are unknown to their conversational partner) and a "learner" (less knowledgeable person). Dyads talked about objects known and unknown to the learners. The teachers talked about the objects without being explicitly instructed to teach the learners about them. Learners were not instructed to learn about the objects, making the learning incidental. After the interaction, learners were asked to recall the unknown objects' names. The corpus is annotated for a range of multimodal cues, enabling analysis of their impact on word learning alongside individual and linguistic predictors.

Our first question concerns whether multimodality plays a significant role above and beyond individual differences of the learner and linguistic dimensions. On the one hand, multimodal cues could play a lesser role. Adults have sophisticated cognitive resources and, as proficient language users, might rely more strongly on linguistic input only. On the other hand, laboratory studies suggest that multimodality might play a role in adult language comprehension and learning (Dargue et al., 2019; Ma et al., 2020). If multimodality plays a significant role in naturalistic adult word learning, adults should benefit from multimodal cues when learning novel words in conversation, with these cues explaining additional variance beyond that accounted for by individual and linguistic factors.

Our second question is whether the impact of multimodal cues is modulated by the learner's individual differences and the linguistic properties of the message. Multimodal cues may not independently affect adult word learning, but rather interact with individual differences (e.g., supporting learners with low working memory) and/or linguistic factors (e.g., facilitating the processing of longer words). There is evidence that gestures improve adult speech processing in more complex discourse or in noisy environments (Drijvers & Özy, 2017; Gluhareva & Prieto, 2017; Holle et al., 2010), and that gestures and prosody reduce cognitive load in word processing (Osorio et al., 2023; Zhang et al., 2021). Therefore, we assess whether reliance on multimodality depends on processing demands and capacity. If this is the case, we should find that multimodal cues moderate the effect of individual and linguistic predictors on adult word learning.

Method

The ECOLANG corpus

ECOLANG (Gu et al., submitted) comprises video-recorded conversations between 33 English-speaking adults talking to a familiar adult. Demographics are detailed at https://osf.io/23s9w/?view_only=37f05d7f28ca467e956ea189703660b0. Adults chat about 12 familiar (e.g., giraffe) and

12 unfamiliar objects (e.g., cassowary) randomly chosen from a pool of 18 familiar and 19 unfamiliar objects (see example in Figure 1). Objects were categorized as "unfamiliar" based on a separate norming study. Before the interaction session, teachers received training on unfamiliar objects. The training provided information about the objects' appearance, origin, use, and other features including name orthography and pronunciation. Object lists and training materials are available at https://osf.io/fmehc/?view_only=5187acaaeb90406f9fd52efa51dfad3c. After the interaction, we asked the learner whether they already knew those objects and their names. The analyses reported here exclude those cases in which the objects were already known. Conversations lasted 2-3 minutes per object.

Learners' vocabulary and working memory were assessed using the tests listed in Table 1. Teachers' words and utterances were auto-transcribed, manually corrected, and aligned with speech files in Praat (Boersma & Weenink, 2023). Multimodal cues overlapping with utterances were annotated in ELAN (Sloetjes & Wittenburg, 2008) following the ECOLANG coding manual, with each utterance and cue coded for its object reference. The corpus includes annotations for representational gestures, points to objects, object manipulations, and eye-gaze to objects. We additionally computed linguistic measures, and prosodic measures in Praat. All cues alongside their definition are shown in Table 1. Analyses focused only on the unfamiliar objects.

Test Phase and Outcome Measure

After the conversation, learners were presented with pictures of the unfamiliar objects and asked to name them. Attempts were transcribed using the International Phonetic Alphabet. The Levenshtein distance between the learners' label attempts and the teachers' label pronunciation was computed. Levenshtein distance is the minimum number of single-character edits needed to change one string into another (Levenshtein, 1966). For each dyad and item, learners' attempts to the teacher's productions were compared, using the lowest score as the final measure of label learning (i.e., the listener's best attempt). Scores were scaled from 0 to 1, with exact label replication scoring 1 and no attempt scoring 0. For scaling, we used the formula 1-X/(Y+1), where X is the raw Levenshtein distance and Y is the highest Levenshtein distance recorded in the sample. This formula gives 0.1 as the minimum value. Scores of 0 could only be obtained by learners who made no attempts at recalling a target label.

Statistical Analyses

We analysed 1153 teachers' utterances that mentioned a target unfamiliar object. Linguistic predictors and multimodal cues were computed for each teacher and object category (Table 1), except for lexical diversity which was computed at the teacher level to obtain a sufficient number of word tokens (>100). We conducted statistical analyses in R (version 4.3.2; R Core Team, 2023), standardising all predictors ($M = 0$, $SD = 1$).

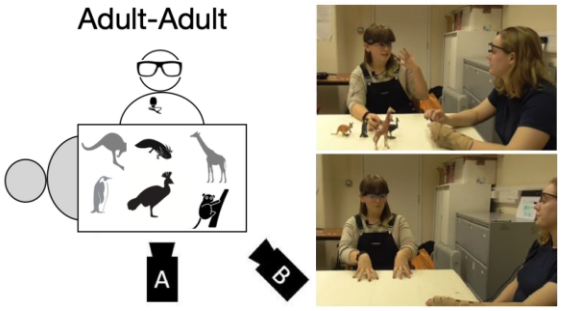

Figure 1: the ECOLANG setup. Participants were seated at a 90-degree angle to each other with objects placed on the table. Two cameras captured the teacher (A) and the interaction space (B). The teacher wore a lapel microphone and a head-mounted eye-tracker (Tobii Pro Glasses 2). The interlocutors engaged in discussions about sets of familiar (grey) and unfamiliar (black) objects (6 objects in each set), distributed across four categories: food, musical instruments, animals, and tools. Each object was discussed both in its presence (top photo) and absence (bottom photo) within the same conversation (order counterbalanced across participants). Adapted from Gu et al. (submitted).

The outcome measure, bounded between 0 and 1, led to significant violations of model assumptions in a linear mixed-effects model. Thus, we used an ordered beta regression model (Kubinec, 2023) with the glmmTMB package (version 1.1.8-9; Brooks et al., 2017), accommodating continuous values between 0 and 1. We focused on: (a) the role of multimodal predictors beyond individual differences and linguistic predictors, and (b) the moderating effect of multimodal predictors on linguistic predictors and individual differences. Thus, we performed model comparisons in blocks (Cernat, 2023), starting with a base model of all linguistic and individual differences. A second model, including all multimodal effects, was compared to it using a likelihood ratio test to assess improvements in model fit. Next, we added bivariate interactions between each multimodal and linguistic or individual difference predictor stepwise, retaining only significant interactions for the final model. We used random intercepts of participant and object category throughout the initial model comparison steps to ensure model convergence at each step. Once the final model with significant interactions was established, we sought the maximal random structure by adding slopes for each predictor on participant or object category, ensuring the most complex feasible random structure allowing convergence (Bolker et al., 2011).

Table 1: List of predictors. We examined teachers' individual differences, linguistic and multimodal cues that align temporally with utterances containing a novel label (McNeill, 1992). Each teacher contributed to the computation of each predictor with an average of 34.9 (*SD* = 11.8) total utterances, and 8.73 (*SD* = 4.26) utterances for each object category (food, music instruments, animals, and tools).

Predictor	Description
Vocabulary Size	Ghent University Vocabulary Test (Brysbaert et al., 2016)
Working Memory	Dual N-Back Test (Jaeggi et al., 2008)
Label Repetition	Number of teacher's utterances mentioning a target label
Label Length	Number of phonemes in a target label
Sentence Concreteness	The average concreteness of words in an utterance, using Brysbaert et al.'s (2014) norms on lemmatised utterances
Mean Length of Utterance (MLU)	The average number of words in an utterance (Ezeizabarrena & Garcia Fernandez, 2018)
Lexical Diversity	Moving-average type-token ratio with a 100-word token window (Covington & McFall, 2010)
Label Pitch	Mean pitch of a target label in semitone: 12* log2(target Hertz/50) (Shi et al., 2022)
Label Speaking Rate	Mean speaking rate of a label: log(N syllables / duration in seconds) (Shi et al., 2022)
Representational Gestures	Proportion of labelling utterances overlapping with a representational gesture, which depicted properties of a label, such as shape or function (ECOLANG; Gu et al., submitted). We additionally omitted enumerative representational gestures (e.g., holding up two fingers to represent the quantity or number two)
Pointing Gestures	Proportion of labelling utterances overlapping with a gesture using deixis (e.g., index finger pointing) to identify specific referents (ECOLANG; Gu et al., submitted). Note, some points in the corpus refer to absent objects, but we only considered points to physically present objects
Object Manipulations	Proportion of labelling utterances overlapping with a communicative action performed while touching an object (e.g., holding a toy to focus a learner's attention; ECOLANG; Gu et al., submitted)
Eye-Gaze to Objects	Proportion of labelling utterances overlapping with a gaze fixation on the target object, lasting over 3 video frames (ECOLANG; Gu et al., submitted). We additionally merged object gaze annotations separated by less than 150ms, which likely indicated the same fixation event (Wass et al., 2013)

We additionally pruned interaction terms that did not generalise well to unseen data. This was done via 5-fold cross-validation, considering both fixed and random effects. We also carried out power simulations to verify that the final model had sufficient power to detect small (practically significant) effect sizes. The simulations also indicated a high type I error (~ 0.1). Thus, we simulated power by applying a false-discovery rate correction (Benjamini & Hochberg, 1995), effectively reducing the type I error to the 0.05 level. The correction was then applied to the p-values of the final model. In the Results section, we report both the non-adjusted and adjusted p-values for the final model. The simulations also indicated that a random effect structure including the random effect intercepts for participants and object category was the only one supported by the data (i.e., avoiding convergence issues when fitting the model on resampled data). This structure was used in the final model.

The final model's assumptions were verified using the DHARMa package (Hartig & Lohse, 2022) for generalised linear mixed models' residual diagnostics. No multicollinearity was detected, with Variance Inflation Factor scores < 2. Models' output and scripts to reproduce data manipulation and analyses are available at https://doi.org/10.17605/OSF.IO/HBGWM.

Results

We found that adding multimodal simple effect terms to a base model which only includes individual differences and linguistic predictors did not significantly improve the model fit (X^2 = 8.48, p = .205; *AIC/BIC base model* = 145.92, 183.30; *AIC/BIC simple multimodal terms* = 149.44, 204.07). Among the predictors in the base model, we found two significant effects (Figure 2). First, label learning improved with increased teacher's label repetition (*Odds Ratio* = 1.39, *95% CI* = 1.13 - 1.71, *p / p adjusted* = .002, .014), supporting the positive effect of label repetition in adult word learning (Hendrickson & Perfors, 2019; Vouloumanos, 2008; Yu & Smith, 2007). Second, mean length of utterance negatively impacted label learning (*Odds Ratio* = .76, *95% CI* = .60 - .98, *p / p adjusted* = .034, 109), consistent with research indicating that shorter sentences facilitate word acquisition due to easier decoding of target words embedded in less

complex grammar (Braginsky et al., 2019). However, this effect became non-significant after adjusting the p-values. No other simple effect was found to be significant.

Further, the fit improved when the model additionally included multimodal interaction terms (X^2 = 4.26, p = .039; *AIC/BIC interaction terms* = 124.68, 196.56). This supports the idea that multimodal cues play a role in naturalistic adult word learning beyond individual and linguistic predictors, specifically moderating their role.

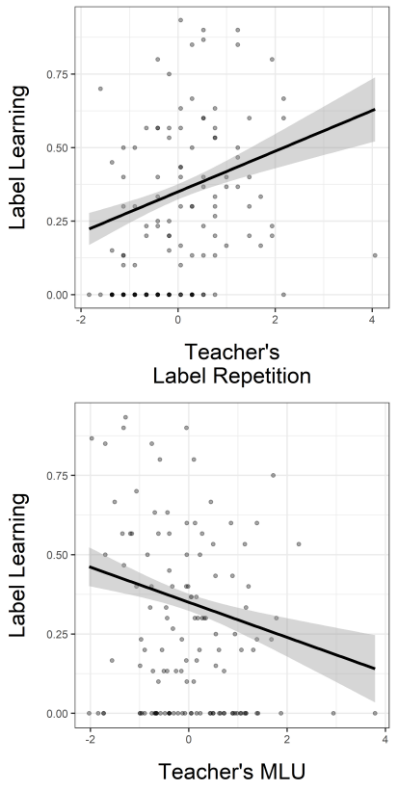

Figure 2: Learners' label learning relative to the teacher's label repetition and mean length of utterance (MLU). The plots display the observed data points along with the regression lines based on model predictions and their 95% confidence bands.

Our analysis revealed two significant moderating effects of multimodal cues (Figure 3). First, the learner's working memory moderated the effect of teacher's representational gestures on learning (*Odds Ratio* = .76, *95% CI* = .61 - .94, *p / p adjusted* = .012, .049). Learners with lower working memory benefited from teachers' representational gestures accompanying label production, possibly because these gestures help map the semantic characteristics of referents to the target labels. This result aligns with evidence that gestures aid adults' comprehension of speech in complex discourse or noisy environments (Drijvers & Özy, 2017; Gluhareva & Prieto, 2017; Holle et al., 2010), likely by reducing the cognitive load during speech processing.

Finally, we found a significant moderation of teachers' pitch on label repetition (*Odds Ratio* = 1.26, *95% CI* = 1.06 - 1.51, *p / p adjusted* = .011, .049), where a lower pitch from the teachers aided label learning, specifically when they infrequently produced the object labels. This effect was unexpected, given previous evidence showing a facilitatory effects of high pitch on word learning (e.g., Filippi et al., 2014; Golinkoff & Alioto, 1995; Ma et al., 2020; Sommers & Barcroft, 2007).

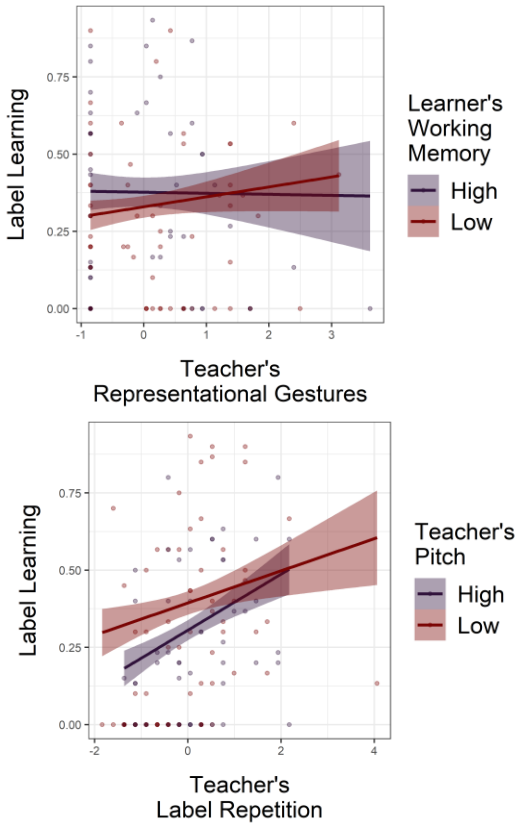

Figure 3: The moderating effects of multimodal cues on label learning. Median split is applied to moderators in the legend for graphical purposes only, as in the statistical model moderators are continuous variables.

Discussion

We explored the role of multimodal cues in adult word learning during social interactions. While previous research has demonstrated that multimodality supports children's learning in naturalistic settings (e.g., Donnellan et al., 2023), the limited literature on adults have focused on laboratory settings in which only individual cues are manipulated while the others are kept constant. Our findings show that adults are sensitive to and use multimodal cues in learning new words. This aligns with recent work indicating the benefits of multimodal cues in speech comprehension beyond childhood (Dargue et al., 2019). Our results highlight the importance of considering the specific situated communicative context of learning (e.g., Murgiano et al., 2021; Reggin et al., 2023).

This research challenges views that deem multimodal information marginal in language processing and learning

(e.g., Kuperberg & Jaeger, 2016). Ignoring multimodality may lead to underestimating the information available to learners in naturalistic settings and its interactions with linguistic cues and the individual characteristics of the learner. This underscores the importance of considering multimodality in linguistic studies.

Furthermore, the interaction effects highlight the need to acknowledge the complexity of word learning contexts. Focusing only on multimodal cues without accounting for individual differences and linguistic predictors could overlook crucial effects, potentially leading studies to underestimate the significant role of multimodality in adult speech processing.

We found that iconic cues, but not indexical cues, predicted listeners' label learning, suggesting that adults might particularly benefit from the semantic conveyance of representational gestures in naturalistic word learning. It needs to be further investigated whether such semantic facilitation plays a more significant role than simply drawing the learner's attention to object referents (McNeill et al., 1994). Iconic gestures enhance the semantic richness and semantic strength of word representations (Straube et al., 2009), and this likely impacts the encoding and retrieval of the word phonological representation. Our finding suggests that not only linguistic predictors (i.e., label repetition) but also representational gestures matter for label learning. This highlights the need to consider how learning word forms in the real world not only entails phonological processing but it is necessarily intertwined with the semantic characteristics of words' referents.

We observed an effect of the sentence-level predictor of mean length of utterance, but this did not hold in the conservative analysis that corrected for family-wise error rates. This may be because of the limited sample of utterances used to compute these measures (recall that all measures were based on utterances containing the target labels). It is possible that an alternative analysis, which calculates the measure over a larger sample of utterances produced by teachers in the conversation, might yield different results. This might also be true for the null effects of the other sentence-level variables (sentence concreteness and lexical diversity) which might be more accurately measured over large utterance samples. The ECOLANG corpus also provides utterances from each teacher that do not contain a target label, but are still part of a conversation whose main topic of discussion is the target object. An alternative analysis might additionally include these utterances in the calculation of the sentence-level predictors. However, we believe such an analysis would not be appropriate given that our outcome measure is label learning, which is likely influenced by the sentence context immediately surrounding the word label.

Surprisingly, we found that teachers' use of higher pitch predicted *poorer* learning scores when the teacher used the label infrequently. This finding contradicts previous studies showing that higher pitch support word learning (Filippi et al., 2014; Golinkoff & Alioto, 1995; Ma et al., 2020; Sommers & Barcroft, 2007). One possibility is that adults might have an overall preference for low-pitched voices (e.g., Klofstad, 2016; Tsantani et al., 2016), which could, in turn, impact word learning. Alternatively, a second possibility would involve examining the behavioural coordination between the teacher and the learner. Namely, it is possible that teachers raise their pitch to enhance the saliency of the target word only when the learner is not fully engaged in the conversation, aiming to grab their attention more efficiently towards the target label. These less engaged learners might then perform more poorly at test. The negative effect of engagement might be especially important when there are fewer occasions to capitalise on phonological/semantic information (i.e., when the label is produced infrequently). Instead, frequent label repetition might compensate for the lack of attention by the learner. At present we are annotating the learner's behaviour and we will carry out analyses including learner and contingent behaviours as soon as annotations are completed.

Although we found a moderation effect involving working memory, the lack of *simple* effects from working memory and vocabulary might be related to our task demands. Social interaction may enhance learning by enabling learners to more efficiently focus their attention on the referent, aligning with the teacher's attentional cues (Elekes & Sebanz, 2020; Verga & Kotz, 2017). This aligns with studies showing negligible effects of working memory and vocabulary in naturalistic social interaction tasks (e.g., Brandt et al., 2022). To observe the influence of these variables, task difficulty may need to be further increased. This is evident in our parallel analysis of semantic learning, using the same ECOLANG corpus of adult conversations (Edwards et al., 2024), where we see an effect of working memory and vocabulary, as the nature of the task requires deeper processing levels (providing a description of the objects after the conversation), and participants must remember more extensive physical and functional information about the objects.

Conclusion

Our research provides evidence that multimodal cues play a significant role in adult word learning during social interactions. This study highlights the importance of studying learning in naturalistic contexts. Our results support models of embodied and situated cognition, and also challenge traditional views that marginalise multimodality in language processing. Our findings underscore the need to incorporate multimodal cues in models of language learning, highlighting the interplay between individual differences, linguistic, and multimodal variables in situated learning contexts. This work paves the way for future research aimed at capturing the complexity of word learning interactions to fully understand adult language processing and acquisition.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological*

Science, *17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Anderson, N. J., Graham, S. A., Prime, H., Jenkins, J. M., & Madigan, S. (2021). Linking Quality and Quantity of Parental Linguistic Input to Child Language Skills: A Meta-Analysis. *Child Development*, *92*(2), 484–501. https://doi.org/10.1111/cdev.13508

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, 57*, 289--300. http://www.jstor.org/stable/2346101.

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer.* (6.3.05) [Computer software]. http://www.praat.org

Bohn, M., & Frank, M. C. (2019). The Pervasive Role of Pragmatics in Early Language. *Annual Review of Developmental Psychology*, *1*(1), 223–249. https://doi.org/10.1146/annurev-devpsych-121318-085037

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2011). *GLMMs in action: Gene-by-environment interaction in total fruit production of wild populations of Arabidopsis thaliana*. https://glmm.wdfiles.com/local--files/examples/Banta_2011_part1.pdf

Booth, A. E., McGregor, K. K., & Rohlfing, K. J. (2008). Socio-Pragmatics and Attention: Contributions to Gesturally Guided Word Learning in Toddlers. *Language Learning and Development*, *4*(3), 179–202. https://doi.org/10.1080/15475440802143091

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). *Consistency and Variability in Children's Word Learning Across Languages. Open Mind : Discoveries in Cognitive Science, 3*, 52–67. https://doi.org/10.1162/opmi_a_00026

Brandt, A. C., Schriefers, H., & Lemhöfer, K. (2022). A laboratory study of naturalistic second language learning: Acquiring grammatical gender from simple dialogue. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(5), 658–679. https://doi.org/10.1037/xlm0001068

Brennan, D. G., & Cullinan, W. L. (1976). The Effects of Word Length and Visual Complexity on Verbal Reaction Times. *Journal of Speech and Hearing Research*, *19*(1), 141–155. https://doi.org/10.1044/jshr.1901.141

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01116

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bulgarelli, F., & Weiss, D. J. (2021). Desirable Difficulties in Language Learning? How Talker Variability Impacts Artificial Grammar Learning. *Language Learning*, *71*(4), 1085–1121. https://doi.org/10.1111/lang.12464

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2023). CLASSIC Utterance Boundary: A Chunking-Based Model of Early Naturalistic Word Segmentation. *Language Learning, 73*(3), 942–975. https://doi.org/10.1111/lang.12559

Cernat, A. (2023). *Longitudinal Data Analysis Using R*. Leanpub. https://leanpub.com/long-data-r

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Dargue, N., Sweller, N., & Jones, M. P. (2019). When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. *Psychological Bulletin*, *145*(8), 765–784. https://doi.org/10.1037/bul0000202

Donnellan, E., Jordan-Barros, A., Theofilogiannakou, N., Brekelmans, G., Murgiano, M., Motamedi, Y., Grzyb, B., Gu, Y., & Vigliocco, G. (2023). The impact of caregivers' multimodal behaviours on children's word learning: A corpus-based investigation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45). https://escholarship.org/uc/item/6km748xv

Drijvers, L., & Özy, ürek A. (2017). Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Edwards, C., Cabiddu, F., Hill-Payne, H., D'Estalenx, Q., Gu, Y., Donnellan, E., & Vigliocco, G. (2024). The impact of speaker's multimodal behaviours on adults' learning of semantic information: A corpus-based investigation. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Elekes, F., & Sebanz, N. (2020). Effects of a partner's task on memory for content and source. *Cognition*, *198*, 104221. https://doi.org/10.1016/j.cognition.2020.104221

Ezeizabarrena, M.-J., & Garcia Fernandez, I. (2018). Length of Utterance, in Morphemes or in Words?: MLU3-w, a Reliable Measure of Language Development in Early Basque. *Frontiers in Psychology*, *8*. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.02265

Filippi, P., Gingras, B., & Fitch, W. T. (2014). Pitch enhancement facilitates word learning across visual contexts. *Frontiers in Psychology*, *5*.

https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01468
Gluhareva, D., & Prieto, P. (2017). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research*, *21*(5), 609–631. https://doi.org/10.1177/1362168816651463
Golinkoff, R. M., & Alioto, A. (1995). Infant-directed speech facilitates lexical learning in adults hearing Chinese: Implications for language acquisition. *Journal of Child Language*, *22*(3), 703–726. https://doi.org/10.1017/s0305000900010011
Gu, Y., Donnellan, E., Grzyb, B., Brekelmans, G., Murgiano, M., Brieke, R., Perniss, P., & Vigliocco, G. (Submitted). *The ECOLANG Multimodal Corpus of adult-child and adult-adult conversation. University College London*.
Hartig, F., & Lohse, L. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models* (0.4.6) [Computer software]. https://cran.r-project.org/web/packages/DHARMa/index.html
Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. *Cognition*, *189*, 11–22. https://doi.org/10.1016/j.cognition.2019.03.005
Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The Associative Structure of Language: Contextual Diversity in Early Word Learning. *Journal of Memory and Language*, *63*(3), 259–273. https://doi.org/10.1016/j.jml.2010.06.002
Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal Analysis of Early Semantic Networks Preferential Attachment or Preferential Acquisition? *Psychological Science*, *20*(6), 729–739. https://doi.org/10.1111/j.1467-9280.2009.02365.x
Holle, H., Obleser, J., Rueschemeyer, S.-A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage*, *49*(1), 875–884. https://doi.org/10.1016/j.neuroimage.2009.08.058
Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833. https://doi.org/10.1073/pnas.0801268105
James, E., Gaskell, M. G., Murphy, G., Tulip, J., & Henderson, L. M. (2023). Word learning in the context of semantic prior knowledge: Evidence of interference from feature-based neighbours in children and adults. *Language, Cognition and Neuroscience*, *38*(2), 157–174. https://doi.org/10.1080/23273798.2022.2102198
Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review*, *23*(4), 1214–1220. https://doi.org/10.3758/s13423-015-0980-7
Jones, G., Cabiddu, F., Barrett, D. J. K., Castro, A., & Lee, B. (2023). How the characteristics of words in child-directed speech differ from adult-directed speech to influence children's productive vocabularies. *First Language*, *43*(3), 253–282. https://doi.org/10.1177/01427237221150070
Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, *24*(2), 313–334. https://doi.org/10.1080/01690960802365567
Klofstad, C. A. (2016). Candidate Voice Pitch Influences Election Outcomes. *Political Psychology*, *37*(5), 725–738. https://doi.org/10.1111/pops.12280
Kobayashi, H., Yasuda, T., & Liszkowski, U. (2023). Marked pointing facilitates learning part names: A test of lexical constraint versus social pragmatic accounts of word learning. *Journal of Child Language*, *50*(2), 296–310. https://doi.org/10.1017/S0305000921000891
Krishnan, S., Watkins, K. E., & Bishop, D. V. M. (2017). The effect of recall, reproduction, and restudy on word learning: A pre-registered study. *BMC Psychology*, *5*(1), 28. https://doi.org/10.1186/s40359-017-0198-8
Kubinec, R. (2023). Ordered Beta Regression: A Parsimonious, Well-Fitting Model for Continuous Data with Lower and Upper Bounds. *Political Analysis*, *31*(4), 519–536. https://doi.org/10.1017/pan.2022.20
Kuhn, G., Tatler, B. W., & Cole, G. G. (2009). You look where I look! Effect of gaze cues on overt and covert attention in misdirection. *Visual Cognition*, *17*(6–7), 925–944. https://doi.org/10.1080/13506280902826775
Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299
Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.
Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2011). Contextual Variability in Free Recall. *Journal of Memory and Language*, *64*(3), 249–255. https://doi.org/10.1016/j.jml.2010.11.003
Ma, W., Fiveash, A., Margulis, E. H., Behrend, D., & Thompson, W. F. (2020). Song and infant-directed speech facilitate word learning. *Quarterly Journal of Experimental Psychology*, *73*(7), 1036–1054. https://doi.org/10.1177/1747021819888982
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
Macedonia, M., Müller, K., & Friederici, A. D. (2010). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping*, *32*(6), 982–998. https://doi.org/10.1002/hbm.21084
Macrine, S. L., & Fugate, J. M. (2022). *Movement Matters: How Embodied Cognition Informs Teaching and Learning*. The MIT Press. https://doi.org/10.7551/mitpress/13593.001.0001

Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition*, *34*(3), 379–413. https://doi.org/10.1017/S0272263112000125

Mathias, B., & Kriegstein, K. von. (2023). Enriched learning: Behavior, brain, and computation. *Trends in Cognitive Sciences*, *27*(1), 81–97. https://doi.org/10.1016/j.tics.2022.10.007

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought* (pp. xi, 416). University of Chicago Press.

McNeill, D., Cassell, J., & McCullough, K.-E. (1994). Communicative effects of speech-mismatched gestures. *Research on Language and Social Interaction*, *27*(3), 223–237. https://doi.org/10.1207/s15327973rlsi2703_4

Meltzer, J. A., Rose, N. S., Deschamps, T., Leigh, R. C., Panamsky, L., Silberberg, A., Madani, N., & Links, K. A. (2016). Semantic and phonological contributions to short-term repetition and long-term cued sentence recall. *Memory & Cognition*, *44*(2), 307–329. https://doi.org/10.3758/s13421-015-0554-y

Motamedi, Y., Murgiano, M., Grzyb, B., Gu, Y., Kewenig, V., Brieke, R., Donnellan, E., Marshall, C., Wonnacott, E., Perniss, P., & Vigliocco, G. (2024). Language development beyond the here-and-now: Iconicity and displacement in child-directed communication. *Child Development*. https://doi.org/10.1111/cdev.14099

Murgiano, M., Motamedi, Y., & Vigliocco, G. (2021). Situating Language in the Real-World: The Role of Multimodal Iconicity and Indexicality. *Journal of Cognition*, *4*(1), 38. https://doi.org/10.5334/joc.113

Neveu, A., & Kaushanskaya, M. (2023). Paired-associate versus cross-situational: How do verbal working memory and word familiarity affect word learning? *Memory & Cognition*, *51*(7), 1670–1682. https://doi.org/10.3758/s13421-023-01421-7

Osorio, S., Straube, B., Meyer, L., & He, Y. (2023). The role of co-speech gestures in retrieval and prediction during naturalistic multimodal narrative processing. *Language, Cognition and Neuroscience*, *0*(0), 1–16. https://doi.org/10.1080/23273798.2023.2295499

Papagno, C., & Vallar, G. (1992). Phonological Short-term Memory and the Learning of Novel Words: The Effect of Phonological Similarity and Item Length. *The Quarterly Journal of Experimental Psychology Section A*, *44*(1), 47–67. https://doi.org/10.1080/14640749208401283

Papagno, C., Valentine, T., & Baddeley, A. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language*, *30*(3), 331–347. https://doi.org/10.1016/0749-596X(91)90040-Q

Pham, T., & Archibald, L. M. D. (2023). The role of working memory loads on immediate and long-term sentence recall. *Memory (Hove, England)*, *31*(1), 61–76. https://doi.org/10.1080/09658211.2022.2122999

Puimège, E., & Peters, E. (2019). Learning L2 vocabulary from audiovisual input: An exploratory study into incidental learning of single words and formulaic sequences. *The Language Learning Journal*, *47*(4), 424–438. https://doi.org/10.1080/09571736.2019.1638630

R Core Team. (2023). *R: A language and environment for statistical computing* [Manual]. R Foundation for Statistical Computing. https://www.R-project.org/

Reggin, L. D., Gómez Franco, L. E., Horchak, O. V., Labrecque, D., Lana, N., Rio, L., & Vigliocco, G. (2023). Consensus Paper: Situated and Embodied Language Acquisition. *Journal of Cognition*, *6*(1), 63. https://doi.org/10.5334/joc.308

Romani, C., Mcalpine, S., & Martin, R. C. (2008). Concreteness Effects in Different Tasks: Implications for Models of Short-Term Memory. *Quarterly Journal of Experimental Psychology*, *61*(2), 292–323. https://doi.org/10.1080/17470210601147747

Sánchez-Borges, I., & Álvarez, C. J. (2023). Comparing mnemonic effects of iconic gestures and pictures on word memory. *Quarterly Journal of Experimental Psychology*, *76*(2), 294–304. https://doi.org/10.1177/17470218221082654

Shi, J., Gu, Y., & Vigliocco, G. (2022). Prosodic modulations in child-directed language and their impact on word learning. *Developmental Science*, *26*(4), e13357. https://doi.org/10.1111/desc.13357

Sloetjes, H., & Wittenburg, P. (2008). Annotation by Category: ELAN and ISO DCR. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf

Sommers, M. S., & Barcroft, J. (2007). An integrated account of the effects of acoustic variability in first language and second language: Evidence from amplitude, fundamental frequency, and speaking rate variability. *Applied Psycholinguistics*, *28*(2), 231–249. https://doi.org/10.1017/S0142716407070129

Storkel, H. L., Armbr, üster J., & Hogan, T. P. (2006). Differentiating Phonotactic Probability and Neighborhood Density in Adult Word Learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175–1192. https://doi.org/10.1044/1092-4388(2006/085)

Straube, B., Green, A., Weis, S., Chatterjee, A., & Kircher, T. (2009). Memory effects of speech and gesture binding: Cortical and hippocampal activation in relation to subsequent memory performance. *Journal of Cognitive Neuroscience*, *21*(4), 821–836. https://doi.org/10.1162/jocn.2009.21053

Tsantani, M. S., Belin, P., Paterson, H. M., & McAleer, P. (2016). Low Vocal Pitch Preference Drives First Impressions Irrespective of Context in Male Voices but Not in Female Voices. *Perception*, *45*(8), 946–963. https://doi.org/10.1177/0301006616643675

Verga, L., & Kotz, S. A. (2017). Help me if I can't: Social interaction effects in adult contextual word learning.

Cognition, *168*, 76–90. https://doi.org/10.1016/j.cognition.2017.06.018

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*(2), 219–247. https://doi.org/10.1515/LANGCOG.2009.011

Vogt, S., & Kauschke, C. (2017). Observing iconic gestures enhances word learning in typically developing children and children with specific language impairment. *Journal of Child Language*, *44*(6), 1458–1484. https://doi.org/10.1017/S0305000916000647

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*(2), 729–742. https://doi.org/10.1016/j.cognition.2007.08.007

Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, *45*(1), 229–250. https://doi.org/10.3758/s13428-012-0245-6

Yasuda, T., & Kobayashi, H. (2022). Ostensive gaze shifting changes referential intention in word meanings: An examination of children's learning of part names. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(2), 272–283. https://doi.org/10.1037/xlm0000859

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., Ballard, D. H., & Aslin, R. N. (2005). The Role of Embodied Intention in Early Lexical Acquisition. *Cognitive Science*, *29*(6), 961–1005. https://doi.org/10.1207/s15516709cog0000_40

Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J. I., & Vigliocco, G. (2021). More than words: Word predictability, prosody, gesture and mouth movements in natural language comprehension. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1955), 20210500. https://doi.org/10.1098/rspb.2021.0500

Acknowledgments

The research reported in this article was supported by an ERC Advanced Grant (743035) to GV, and an NWO Rubicon Grant (019.182SG.023) to YG.