The impact of speakers' multimodal behaviours on adults' learning of semantic information: A corpus-based investigation

Christopher Edwards[1*] (christopher.edwards.20@ucl.ac.uk)
Francesco Cabiddu[1*] (francesco.cabiddu@ucl.ac.uk)
Harriet Hill-Payne[1] (harriet.hill-payne.22@ucl.ac.uk)
Quitterie D'Estalenx[2] (qmad201@exeter.ac.uk)
Ed Donnellan[1, 3] (ed.donnellan@warwick.ac.uk)
Yan Gu[1, 4] (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 Computer Science, University of Exeter, North Park Road, EX4 4QF, UK
[3]Department of Psychology, University of Warwick, University Road, Coventry, CV4 7AL, UK
[4]Department of Psychology, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK

Abstract

Adults often learn new semantic information in face-to-face communication with other adults (e.g., teachers, colleagues). More knowledgeable individuals provide an ensemble of multimodal behaviours that can shape the information that their interlocutors learn. Using the naturalistic ECOLANG corpus of dyadic conversations, we ask whether multimodal behaviours (pitch, speaking rate, representational gestures, points, object manipulations, and gaze) support adults' semantic learning of unknown objects above and beyond verbal properties of utterances (number of utterances, lexical diversity, mean length of utterances, concreteness) and learners' individual differences (vocabulary, working memory). We found that individual differences, pointing and object manipulations affected learning, with verbal and multimodal factors also interacting to predict adult semantic learning. Our results highlight the relevance of accounts of multimodal learning in adulthood and the importance of considering naturalistic interaction in its complexity to understand the factors that influence adult learning.

Keywords: concepts; semantic learning; multimodal communication; gesture; multimodal corpus

Introduction

You are visiting the Metropolitan Museum in New York with a friend, a Roman history enthusiast, when your attention is drawn to a curious curved implement. Your friend informs you that that is a "strigil", explaining that it was a Roman cleaning tool and demonstrating how it was scraped across the skin. A few months later you are in the British Museum in London; you notice a strigil, and point it out to another friend, describing its function and recreating its scraping motion. You have learned a concept.

We learn new concepts throughout the lifespan, often during interactions with other, more knowledgeable people: infants learn from their caregivers; adults learn from teachers, colleagues and friends. This process of conceptual learning – the encoding of semantic information that we can later retrieve for relevant tasks – does not only occur in formal settings (e.g., schools, universities), but often takes place incidentally, during face-to-face communication with others (Cronin-Golomb & Bauer, 2023; De Felice et al., 2023). But which factors might support learning of new semantic information in these scenarios? Some are individual – for example, our working memory. Yet some other factors are situational – learning also depends on how information is made available to us in our environment. Our concept of a strigil will differ depending on whether our friend explains its function, or merely comments on its shape.

Face-to-face communication provides learners with a rich composite of semantic information (words and gestures that imagistically relate to properties of referents) and other multimodal cues that direct attention to a visually available referent (e.g., gaze, points) or to important information in the utterance (prosodic modulation). Thus, the non-verbal behaviours of a teacher can support learning in at least two different ways. First, they can enhance meaning-making by providing additional information. Second, they may strengthen signal robustness by introducing redundancies and leaving cross-modal memory traces – as argued by Paivio's Dual Coding Theory (1991) and Mayer's related theory of multimedia learning (2009) – or by providing information with sensory grounding, activating sensorimotor networks that have been argued to underpin conceptual knowledge, as proposed by the embodied cognition account (Barsalou, 2008; Hostetter & Alibali, 2019).

Variability in multimodal communicative behaviours is therefore likely to influence how and how well we learn. Here, we investigate several multimodal behaviours that speakers spontaneously produce in face-to-face communication, (lexical diversity, length of utterance, utterance, concreteness, speaking rate, speech pitch, representational gestures, pointing, object manipulation and object gaze), examining their role in adults' immediate semantic learning.

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.*

The role of verbal and multimodal behaviours in learning semantic information

Despite a lack of work directly investigating the impact of a teacher's behaviours on adults' learning of new semantic knowledge during face-to-face communication, there is an extensive body of literature on their impact on word learning in childhood and memory recall in adults.

Verbal behaviours Verbal properties of input can vary considerably, with this shaping the nature of the semantic information learners receive. Lexical diversity, or the variety of words used during input (measured by type-token ratio, or the ratio of unique words to total words across input), indexes the semantic richness of the information provided: more diverse lexical input is likely to allow for the development of richer semantic representations, and has been linked with vocabulary learning in development (Cychosz et al., 2021; Rowe & Snow, 2020), where it has been shown to predict 3-to-4-year-olds' ability to match novel words to their meanings (Dong et al., 2021). Finally, sentence length has been linked to child vocabulary growth (Anderson et al., 2021; Braginsky et al., 2019). The length of utterances has also been related children's learning, where caregivers' use of longer sentences has been shown to predict vocabulary and conceptual development (Bornstein et al., 1998; Brown, 1973; Hoff & Naigles, 2002). Evidence concerning adults' semantic learning is scarce, but longer utterances may condense and integrate more information, facilitating learners' development of richer semantic representations. Finally, utterance concreteness – the extent to which words refer to perceptible concepts – has been claimed to facilitate learning through the relatively greater activation of the brain's sensorimotor networks by concrete words compared to abstract ones (Paivio, 2013; Vigliocco et al., 2009), with research showing benefits to sentence processing and recall (Meltzer et al., 2016; Pham & Archibald, 2023; Romani et al., 2008).

Multimodal behaviours In addition to these verbal properties, a number of multimodal cues have also been shown to support learning and processing. Prosodic modulations of speech pitch that direct attention to new information have been tied to immediate learning. Throughout the lifespan, individuals show sensitivity to prosodic pitch, with higher pitch seemingly prompting both adults and children to mark accompanying units of information in speech as salient (Cristia, 2013; Wagner, 2020). While it is well-established that infants are particularly dependent on higher pitched speech (Cox et al., 2022; Cristia, 2013), and that immediate learning of linguistic input (e.g., labels) is predicted by use of higher pitch (Graf Estes & Hurley, 2013; Ma et al., 2011; Shi et al., 2023), evidence suggests these effects extend to adults; for adults, the use of higher pitch to accentuate new information improves language processing (Bock & Mazzella, 1983; Heim & Alter, 2006) and recall (Filippi et al., 2014; Fraundorf et al., 2010; Lee & Fraundorf, 2017; Sanford et al., 2006). Speech characterized by high prosodic pitch might therefore reflect greater prosodic scaffolding of information, aiding learners' processing of semantic information (Helfrich & Weidenbecher, 2011). Furthermore, slower speaking rates are associated with benefits to recall and learning. In child-directed language (CDL), caregivers reliably present novel information at slower speeds than given information to the child's benefit (Shi et al., 2023). In adults, evidence suggests there is a negative relationship between speaking rate and verbal recall, with increases in speaking rate in both auditory-only and audiovisual input impeding recall (and especially older adults) (Sommers et al., 2020).

Visual co-speech cues also play a role in the transmission of information. Indexical cues like points and gazes towards objects, as well as manual manipulations of objects, guide attention towards referents, allowing learners to map linguistic information to sensory experience, not only marking such information as salient, but also reinforcing concepts by encoding them from multiple sources. This is important in early development as a means of providing conceptual grounding for words (Bohn & Frank, 2019; Rader & Zukow-Goldring, 2010). In adulthood, these benefits seem to depend on the cue; evidence from classroom-based paradigms suggests that while pointing continues to be associated with improved learning of information (Pi et al., 2017; Rueckert et al., 2017), gaze does not seem to be as useful (Pi et al., 2019; van Wermeskerken & van Gog, 2017).

Alongside these indexical cues, speakers also produce representational gestures depicting some of the referents' semantic features (e.g., holding the hands vertically in front of the chest to indicate playing a pipe) (McNeill, 1992). Representational gestures can aid learning as they provide additional semantic information. Research has demonstrated benefits to observing representational gestures alongside speech for verbal recall and comprehension tasks in adults (Cohen & Otterbein, 1992; Dargue et al., 2019; Dargue & Sweller, 2020; Feyereisen, 2006).

Thus, previous studies suggest not only that verbal properties of speech might influence semantic learning, but that a multitude of multimodal behaviours used by a knowledgeable speaker may also lead to their interlocutor developing better and richer representations of new concepts.

The current study

This study aims to assess whether a more knowledgeable speaker's use of verbal and multimodal behaviours when talking about objects unknown to their interlocutor has an impact on the interlocutor's learning. To this end, our research is motivated by two key questions. First, do the multimodal behaviours used by a speaker predict their interlocutor's semantic learning over and above the verbal properties of their speech? And second, *which* of these multimodal behaviours predict this learning better?

We approach these questions by using data from the ECOLANG corpus: a new multimodal dataset of naturalistic interactions between two familiar adults (Gu et al., submitted). In the corpus, participants discuss a series of familiar and novel objects. One participant (referred to hereafter as the 'Teacher') is more knowledgeable than the

other, since they have been previously informed about all the objects being discussed; they are asked to lead the conversation with their partner (the 'Learner'), though they are not explicitly told to 'teach' them. Following the interaction, the Learner is asked to describe the novel objects.

Our work thus addresses several gaps in the current literature on language, communication and learning. First, while the multimodal features of CDL and their relative influence on children's learning have been extensively researched and theorized (Donnellan et al., 2023; Rowe & Snow, 2020; Vigliocco et al., 2019), how such multimodal behaviours might impact adults' learning is far less understood. Indeed, since adults, with their large vocabularies (Brysbaert et al., 2016; Segbers & Schroeder, 2017) and better developed theory of mind (Moran, 2013; Valle et al., 2015), can deploy a range of cognitive resources unavailable to children, the dynamics of children's learning during face-to-face communication may differ. Here, we investigate how multimodal *adult*-directed language influences learning.

Second, research on multimodal communication and learning has been dominated by work on word learning, and in particular the mapping problem – learners' successful association of novel labels to referents (Quine, 1960; Wojcik et al., 2022). However, this approach is relatively insensitive to how the semantic representations underpinning word meanings may vary, and consequently how multimodal communication behaviours may shape learners' development of these representations. By assessing *semantic* learning, we aim to identify learning benefits of multimodal communication that may have been marginalized by research hitherto.

Finally, in using data from the ECOLANG corpus, we are able to assess adults' verbal and multimodal behaviours as they are produced spontaneously in face-to-face communication, during a semi-naturalistic task, thereby allowing us closer access to how learners use such behaviours in the real world. While there are many studies on CDL that have used naturalistic interactions, we are not aware of studies with adult learners. Looking at naturalistic data is important because we are able to capture the real-world distribution of verbal and multimodal behaviours and examine their influence *together* – reflecting the nature of face-to-face communication as a multimodal gestalt (Holler & Levinson, 2019). This distinguishes our research from other work on multimodal communication and learning, which examined individual cues, or pairs of cues, in isolation.

Given the wide-ranging evidence supporting learning and recall benefits for multimodal behaviours, we predicted that multimodal behaviours would show an overall benefit to semantic learning. In particular, we expected to find that multimodal cues exerted a positive effect on learning, above and beyond that of verbal variables. Since no previous study has assessed the role of verbal *and* multimodal cues together and there is very little research that has investigated learning of semantic information, we were more tentative regarding the role of specific verbal and multimodal cues.

Methods

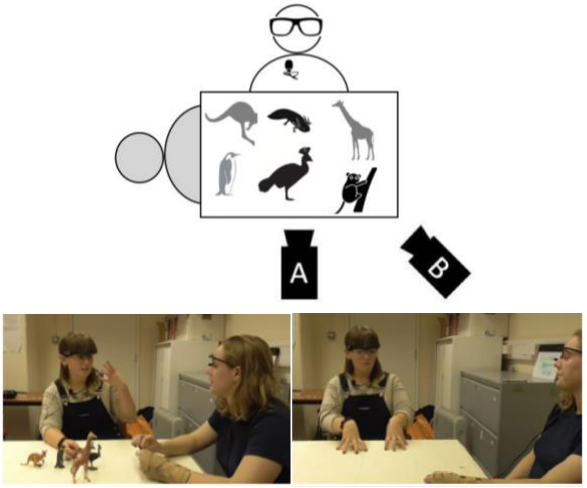

Figure 1: Top panel demonstrates recording setup. Teacher (white) and Learner (grey) sit at a table with objects from a particular category (e.g., animals) on the table. Two cameras record the interaction: A focuses on the Teacher, and B on the interaction space. Participants wear Tobii eye-tracking glasses and a lapel microphone. Grey objects (kangaroo, penguin, giraffe) are previously known to both Teacher and Learner. Black objects (axolotl, cassowary, tarsier) are novel and only known about by the Teacher. Bottom panel provides examples of recording, taken by Camera A, when objects are present (left) and absent (right). Adapted from Gu et al., submitted.

The ECOLANG corpus

This corpus comprises data from 33 dyads of familiar adults. Demographics are detailed at https://osf.io/23s9w/?view_only=37f05d7f28ca467e956ea189703660b0. The interaction session took place in a laboratory room in the Experimental Psychology Department at UCL. Participants talked about 12 familiar ('known') and 12 unfamiliar ('unknown') objects grouped across four categories: fruits (e.g., *cherimoya*), musical instruments (e.g., *xun*), animals (e.g., *axolotl*), and tools (e.g., *strigil*) (see Figure 1). Objects were selected from a pool of 39 stimuli, of which 19 were classified as previously unknown to participants, based on a norming survey conducted amongst native English speakers (Gu et al., submitted). Prior to the interaction session, the Teacher was sent training videos about the novel objects, demonstrating their appearance, and describing their origin, use and other features. They were asked not to discuss the training content with the Learner. During the interaction, the Teacher freely described the (unknown and known) objects in sets of six (each category) in two conditions: with the objects present on the table in front of them, and without. The sequence of categories, and conditions within each category (i.e., object absent-first or present-first), were counterbalanced across participants. The dyad discussed each set of objects for four to five minutes in both present and absent sessions. This process was repeated for the four categories; full recording sessions lasted between

32 and 40 minutes. The Learner's working memory (Dual N-Back Test; Jaeggi et al., 2008), and vocabulary (Ghent University Vocabulary Test; Brysbaert et al., 2016) were also assessed before interaction.

Testing At the end of the interaction session, Learners were tested on their learning of semantic information about the objects; they were shown an image prompt of each novel object they discussed and asked to provide a definition, giving as many details as possible. Learners' learning of the name of the object was also tested. Results pertaining to label learning are provided in Cabiddu et al. (2024).

Annotation of the Teachers' behaviours As described in Gu et al (submitted), speech was manually transcribed using Praat (Boersma & Weenink, 2019) and ELAN (Sloetjes & Wittenburg, 2008). Speech was initially transcribed on an utterance level, defined as a unit that expresses a single event (Berman et al., 1994). Teachers' utterance pitches were automatically extracted using Praat, with utterances returned as having undefined pitch excluded, along with utterances that were outliers in mean, minimum or maximum pitch.

Gestures, object manipulations and gaze fixations were annotated using ELAN according to the ECOLANG coding manual (Gu et al., submitted). Representational gestures were defined as gestures that represent properties of referents, such as the shape or function of an object (e.g., hands placed vertically in front of the mouth to represent playing a pipe). Points were defined as gestures that single out a particular referent through deixis. Object manipulations were any movement or action performed while touching an object that were deemed to be communicatively meaningful (e.g., holding an object to direct the Learner's attention to it). Object gaze was measured as gaze fixations on the objects that lasted for 3 or more consecutive frames. Raw recordings from the eye tracking glasses were processed in order to establish gaze position; they were then manually annotated by an expert coder on ELAN to mark the specific object fixated on by the participant. Further information on annotation will be available in Gu et al. (submitted).

Measures

Predictors For the 33 dyads in the sample, we extracted N = 9,527 utterances in which Teachers talked about unknown objects. The analyses reported here focus on these utterances only. Across this sample, we considered three verbal measures and six multimodal measures. Unless specified, these measures were computed by object category to ensure samples of utterances were large enough to robustly compute predictors. The measures are:

Mean Length of Utterance (MLU; Brown, 1973) was measured in words rather than morphemes, following others (Dickinson & Porche, 2011; Dong et al., 2021; Potratz et al., 2022).

Lexical diversity was measured as moving average type-token ratio (MATTR; Covington & McFall, 2010). MATTR scores were taken at the participant level, since participants did not always talk for long enough about individual object categories to produce usable samples.

Concreteness was calculated as the mean concreteness of each utterance, using concreteness ratings from Brysbaert et al. (2014)

Pitch was calculated as the mean F0 of utterances and transformed to semitones following Shi et al. (2023): 12* log2(target Hertz/50).

Speaking rate was calculated as log(N syllables / duration in seconds), again following Shi et al. (2023).

We measured the Teacher's use of other multimodal cues by rate, whereby cues which temporally overlapped with any utterance concerning the object were counted and divided by the total time spent talking about an object category. Points, object manipulations and representational gestures were computed per minute, with gaze computed per second. Given the fact that participants spent half their time discussing objects while they were absent, the rates for points, object manipulations and object gaze (cues which directly engage with the objects) were calculated using only the utterances in the present condition, while the rate for representational gestures, pitch and speaking rate measures were calculated across utterances in both present and absent conditions.

Learning outcomes Learning was assessed using a semantic learning score, calculated by measuring the cosine similarity between all of the Teacher's utterances about an object during interaction, and the Learner's utterances about the object during testing. Teacher and Learner utterances were concatenated and then filtered to remove non-content words; the cosine similarity of the word embeddings of the resulting texts was then calculated. Embeddings were generated using the GloVe model (Pennington et al., 2014). For each object per participant pair, a cosine similarity score was calculated between each word of the Teacher output and each word of the Learner output, and the maximum score for each Teacher word taken (i.e., the score between each Teacher word and the most similar Learner word). These scores were averaged across the object output. Scores ranged from 0 to 1, where 1 indicates that texts were identical. Object scores were then grouped by category, as with the predictors.

Analysis

The outcome variable was the Learner's semantic learning score. The predictors of interest were MLU, MATTR, mean concreteness of utterance, pointing, representational gestures, object manipulation, and gaze to object. Total utterances were included in the analysis to control for quantity of input. The Learner's working memory and vocabulary scores were also included to control for individual cognitive differences, given evidence supporting the importance of working memory function and prior vocabulary knowledge to learning in development (Cowan, 2014; Gray et al., 2022; James et al., 2023). All predictors were standardized ($M = 0$, $SD = 1$).

We used R for statistical analyses (version 4.3.2; R Core Team, 2023). As the outcome measure is bounded between 0 and 1, we fitted an ordered beta regression model (Kubinec,

2023), using the glmmTMB package (version 1.1.8-9; Brooks et al., 2017), allowing to fit a single linear model to both bounded continuous responses.

As in Cabiddu et al. (2024), we carried out model comparisons in blocks (e.g., Cernat, 2023). We fitted a base model including all verbal behaviours, measures of participant individual differences (vocabulary and working memory) and other controls (total number of utterances). We then fitted a second model that additionally contained all multimodal simple effects, and carried out model comparison via likelihood ratio test to test whether multimodal cues significantly improved the model fit. Next, we added bivariate interactions between each multimodal cue and each verbal or individual difference predictor in a stepforward fashion, only keeping significant interactions in the final model to contain model complexity. For interactions that improved the model fit in the model comparison, we also assessed their predictive accuracy via 5-fold cross-validation, considering fixed and random effects; we removed interaction terms that did not generalize to unseen data during cross-validation. To verify that the model had enough power to detect small effect sizes, we ran power simulations. These simulations also indicated a high type I error rate (~ 0.1). We therefore simulated power by applying a false-discovery rate correction to the model's p-values (Benjamini & Hochberg, 1995), effectively reducing the type I error to a 0.05 level, with this correction applied to the final model's p-values. Finally, we used the power simulations to determine our final random effect structure; the simulations indicated that including only the random effect intercepts for participants converged even when fitting the model on resampled data. This structure was used in the final model.

Model assumptions were checked using the DHARMa package (Hartig & Lohse, 2022). Variance Inflation Factors (VIF) were calculated to check for multicollinearity between predictors; VIF were < 2 for all predictor variables. Scripts to reproduce data manipulation and analysis can be found at https://doi.org/10.17605/OSF.IO/E8JHN.

Results

Adding multimodal predictors and significant interactions to the base model containing only verbal predictors and individual differences significantly improved the model fit (X^2 = 20.05, p = .005).

Among our predictors, we found a significant main effect of pointing, with higher rates of pointing (*Odds Ratio* = 1.11, *95% CI* = 1.03-1.2, p = .023) predicting better learning outcomes (Figure 2). We also found significant main effects for our participants' cognitive characteristics. Higher working memory scores predicted better learning outcomes (*Odds Ratio* = 1.14, *95% CI* = 1.04-1.25, p = .013), while there was a quadratic effect of the Learner's vocabulary knowledge: higher scores in the vocabulary test predicting better performance, but only for learners with smaller vocabularies (*Odds Ratio* = .88, *95% CI* = .81-.96, p = .013).

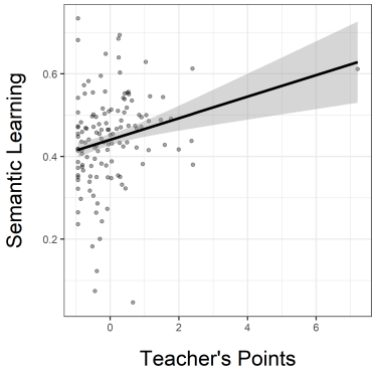

Figure 2: Main effect of rate of points. The plot displays the observed data points along with the regression lines based on model predictions and their 95% confidence intervals.

We also found a main effect for lower rates of object manipulation (*Odds Ratio* = .9, *95% CI* = .84-.98, p = .029). However, this was driven by the moderating effect of total utterances produced: (*Odds Ratio* = .91, 95% CI = .85-.97, p = .013): object manipulations negatively predicted learning only when relatively more utterances about the objects were produced (see Figure 3, top).

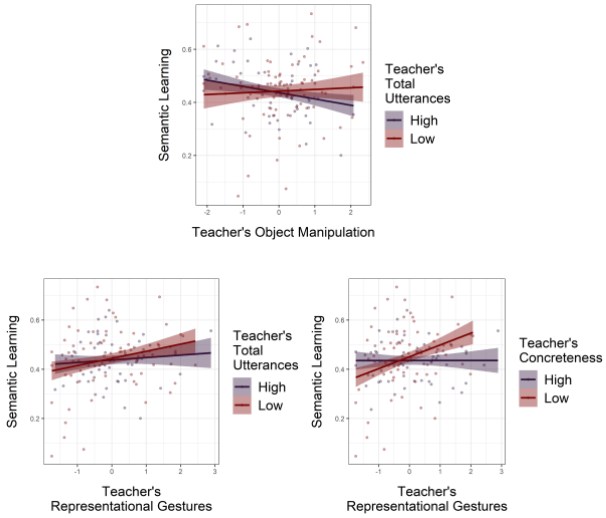

Figure 3: Moderating effects of multimodal cues on Learner semantic learning. The plots display observed data points alongside the regression lines based on model predictions and their 95% confidence bands. Median split is applied to moderators for graphical purposes only; in the statistical model moderators are continuous variables.

We also found two interaction effects for representational gestures (see Figure 3, bottom). First, representational gestures interacted with total number of utterances (*Odds Ratio* = .89, 95% CI = .83-.96, p = .013), with a higher rate of representational gestures predicting better learning outcomes when more utterances about the objects were produced. Second, representational gestures also interacted with concreteness (*Odds Ratio* = .86, 95% CI = .81-.92, p =

<0.001). When utterances were less concrete, a higher rate of representational gestures predicted better learning outcomes.

Discussion

We examined several verbal and multimodal behaviours performed by a speaker during face-to-face communication and assessed their influence over their partner's immediate learning of new semantic information from the interaction. We found that participants' individual characteristics played some role, alongside variability in several of the Teacher's multimodal behaviours predicted learning above these.

First and foremost, we found that our model including multimodal behaviours and interactions fit the data significantly better than the base model, which only contained participants' cognitive characteristics and verbal measures. This shows for the first time that multimodal behaviours influence whether and how interlocutors learn semantic information during naturalistic interaction.

We found several main effects. First, as expected, we observed that participants' cognitive characteristics predicted learning. We observed better learning for Learners with stronger working memory performance, and a quadratic effect of vocabulary, where participants with smaller vocabularies learned less, but where no relationship was seen for participants with larger vocabularies.

Among our predictors, The Teacher's points to the object also positively predicted the Learner's semantic learning. This is consistent with previous research on adult learning, particularly in classroom settings, where several studies have suggested learning benefits for pointing over other cues like gaze (Pi et al., 2019; Rueckert et al., 2017; van Wermeskerken & van Gog, 2017).

By contrast, the Teacher's object manipulations *negatively* predicted learning, against evidence from developmental studies on learning (Jant et al., 2014). However, this effect was driven by the moderating effect of the total number of utterances produced by the Teacher: when Teachers spoke relatively more about objects, increased object manipulations predicted poorer learning outcomes. One possible explanation for this is that during naturalistic conversation between adults, speakers may make object manipulations that are less tightly linked to speech content (compared to caregivers); this may have distracted Learners when processing more verbal semantic information.

There was no main effect of representational gesture found, despite meta-analysis finding general memory and learning benefits for observing representational gestures (Dargue et al., 2019; Hostetter, 2011). However, we did observe two interactions between the Teacher's use of representational gestures and other variables: the total number of the Teacher's utterances, and the mean concreteness of their utterances. These interactions may explain the absence of a main effect. First, representational gestures significantly predicted learning when Teachers spoke relatively *less* about objects. This is in fact consistent with studies showing benefits for observation of representational gesture at the word and sentence level (Feyereisen, 2006; So et al., 2012), as well as studies that have found absent or moderated learning effects of representational gestures in comprehending longer discourse (Dahl & Ludvigsen, 2014; Dargue & Sweller, 2020). This suggests that when speech is relatively less, gestures may play a proportionally more important role in developing semantic representations, but when speech is greater, gestures are superseded by richer verbal content. Second, representational gestures predicted learning when utterances were relatively *abstract*. This may point to an embodiment function of representational gestures, whereby learners benefit from concrete support of relatively abstract semantic information, in line with accounts that emphasize the sensorimotor grounding of abstract conceptual knowledge (Barsalou & Wiemer-Hastings, 2005; Zdrazilova et al., 2018).

There were some other noteworthy null effects that emerged in our data, namely for MLU, lexical diversity, prosodic pitch and speaking rate. There may be a several reasons for this. First, these factors have all been most strongly tied to learning in children. It may be the case that adults' relatively greater cognitive resources and linguistic skills mean that these cues become relatively less important. Second, the findings here relate to semantic learning, while these measures have been most strongly connected with word learning. It may be that case that particular cues help learners encode word forms (e.g., higher pitch), or strengthen connections between labels and referents (e.g., lexical diversity), rather than contributing to the development of semantic representations. Finally, no previous study has investigated these behaviours together, in a naturalistic setting. It is plausible that cues previously found to benefit learning did so under experimental conditions in which more useful cues were absent – when cues are restricted learners may look for multimodal support where they can find it, but when given multiple cues, come to depend on some more than others.

Conclusion

Here we assessed for the first time the learning of semantic information during naturalistic face-to-face communication. We show that speakers' use of multiple multimodal behaviours play an important role in how their interlocutors learn new semantic information, demonstrating that these cues influence immediate learning beyond verbal properties of input, and identifying specific behaviours that might support this learning. This work underlines the importance of approaching language as a multimodal gestalt, reinforces the value of investigating language, learning and communication in ecologically valid scenarios, and bears clear significance for educational theorists and practitioners.

Acknowledgement

The work reported here was supported by an ERC Advanced Grant (743035) to GV, and an NWO Rubicon Grant (019.182SG.023) to YG.

References

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating Abstract Concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking* (pp. 129–163). Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9780511499968.007

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Berman, R. A., Slobin, D. I., Aksu-Koç, A. A., Bamberg, M., Dasinger, L., Marchman, V., Neeman, Y., Rodkin, P. C., Sebastián, E., & et al. (1994). *Relating events in narrative: A crosslinguistic developmental study* (pp. xiv, 748). Lawrence Erlbaum Associates, Inc.

Bock, J. K., & Mazzella, J. R. (1983). Intonational marking of given and new information: Some consequences for comprehension. *Memory & Cognition*, *11*(1), 64–76. https://doi.org/10.3758/BF03197663

Bohn, M., & Frank, M. C. (2019). *The Pervasive Role of Pragmatics in Early Language*.

Bornstein, M. H., Haynes, M. O., & Painter, K. M. (1998). Sources of child vocabulary competence: A multivariate model. *Journal of Child Language*, *25*(2), 367–393. https://doi.org/10.1017/S0305000998003456

Brown, R. (1973). Development of the first language in the human species. *American Psychologist*, *28*(2), 97–106. https://doi.org/10.1037/h0034209

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

Cabiddu, F., Edwards, C., Hill-Payne, H., Donnellan, E., Gu, Y., & Vigliocco, G. (2024). What Predicts Adult Word Learning in Naturalistic Interactions? A Corpus Study. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Cohen, R. L., & Otterbein, N. (1992). The mnemonic effect of speech gestures: Pantomimic and non-pantomimic gestures compared. *European Journal of Cognitive Psychology*, *4*(2), 113–139. https://doi.org/10.1080/09541449208406246

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

Cowan, N. (2014). Working Memory Underpins Cognitive Development, Learning, and Education. *Educational Psychology Review*, *26*(2), 197–223. https://doi.org/10.1007/s10648-013-9246-y

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., & Fusaroli, R. (2022). A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour*, *7*(1), 114–133. https://doi.org/10.1038/s41562-022-01452-1

Cristia, A. (2013). Input to Language: The Phonetics and Perception of Infant-Directed Speech. *Language and Linguistics Compass*, *7*(3), 157–170. https://doi.org/10.1111/lnc3.12015

Cronin-Golomb, L. M., & Bauer, P. J. (2023). Self-motivated and directed learning across the lifespan. *Acta Psychologica*, *232*, 103816. https://doi.org/10.1016/j.actpsy.2022.103816

Cychosz, M., Edwards, J. R., Bernstein Ratner, N., Torrington Eaton, C., & Newman, R. S. (2021). Acoustic-Lexical Characteristics of Child-Directed Speech Between 7 and 24 Months and Their Impact on Toddlers' Phonological Processing. *Frontiers in Psychology*, *12*, 712647. https://doi.org/10.3389/fpsyg.2021.712647

Dahl, T. I., & Ludvigsen, S. (2014). How I See What You're Saying: The Role of Gestures in Native and Foreign Language Listening Comprehension. *The Modern Language Journal*, *98*(3), 813–833.

Dargue, N., & Sweller, N. (2020). Two hands and a tale: When gestures benefit adult narrative comprehension. *Learning and Instruction*, *68*, 101331. https://doi.org/10.1016/j.learninstruc.2020.101331

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

De Felice, S., Hamilton, A. F. D. C., Ponari, M., & Vigliocco, G. (2023). Learning *from* others is good, *with* others is better: The role of social interaction in human acquisition of new knowledge. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1870), 20210357. https://doi.org/10.1098/rstb.2021.0357

Dickinson, D. K., & Porche, M. V. (2011). Relation Between Language Experiences in Preschool Classrooms and Children's Kindergarten and Fourth-Grade Language and Reading Abilities: Preschool Language Experiences and Later Language and Reading. *Child Development*, *82*(3), 870–886. https://doi.org/10.1111/j.1467-8624.2011.01576.x

Dong, S., Gu, Y., & Vigliocco, G. (2021). The impact of child-directed language on children's lexical development. *Proceedings of the Annual Meeting of*

the Cognitive Science Society, *43*(43). https://escholarship.org/uc/item/38x9h9h4

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

Feyereisen, P. (2006). Further investigation on the mnemonic effect of gestures: Their meaning matters. *European Journal of Cognitive Psychology*, *18*(2), 185–205. https://doi.org/10.1080/09541440540000158

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

Fraundorf, S. H., Watson, D. G., & Benjamin, A. S. (2010). Recognition memory reveals just how CONTRASTIVE contrastive accenting really is. *Journal of Memory and Language*, *63*(3), 367–386. https://doi.org/10.1016/j.jml.2010.06.004

Graf Estes, K., & Hurley, K. (2013). Infant-Directed Prosody Helps Infants Map Sounds to Meanings. *Infancy*, *18*(5), 797–824. https://doi.org/10.1111/infa.12006

Gray, S. I., Levy, R., Alt, M., Hogan, T. P., & Cowan, N. (2022). Working Memory Predicts New Word Learning Over and Above Existing Vocabulary and Nonverbal IQ. *Journal of Speech, Language, and Hearing Research*, *65*(3), 1044–1070. https://doi.org/10.1044/2021_JSLHR-21-00397

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.*

Heim, S., & Alter, K. (2006). Prosodic pitch accents in language comprehension and production: ERP data and acoustic analyses. *Acta Neurobiologiae Experimentalis*, *66*(1), 55–68. https://doi.org/10.55782/ane-2006-1587

Helfrich, H., & Weidenbecher, P. (2011). Impact of Voice Pitch on Text Memory. *Swiss Journal of Psychology*, *70*(2), 85–93. https://doi.org/10.1024/1421-0185/a000042

Hoff, E., & Naigles, L. (2002). How Children Use Input to Acquire a Lexicon. *Child Development*, *73*(2), 418–433.

Holler, J., & Levinson, S. C. (2019). Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences*, *23*(8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, *137*(2), 297–315. https://doi.org/10.1037/a0022128

Hostetter, A. B., & Alibali, M. W. (2019). Gesture as simulated action: Revisiting the framework. *Psychonomic Bulletin & Review*, *26*(3), 721–752. https://doi.org/10.3758/s13423-018-1548-0

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 of the United States of America*, *105*(19), 6829–6833.

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

Jant, E. A., Haden, C. A., Uttal, D. H., & Babcock, E. (2014). Conversation and Object Manipulation Influence Children's Learning in a Museum. *Child Development*, *85*(5), 2029–2045. https://doi.org/10.1111/cdev.12252

Lee, E.-K., & Fraundorf, S. (2017). Effects of contrastive accents in memory for L2 discourse. *Bilingualism: Language and Cognition*, *20*(5), 1063–1079. https://doi.org/10.1017/S1366728916000638

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word Learning in Infant- and Adult-Directed Speech. *Language Learning and Development*, *7*(3), 185–201. https://doi.org/10.1080/15475441.2011.579839

Mayer, R. E. (2009). *Multimedia learning, 2nd ed.* (pp. xiii, 304). Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

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

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

Moran, J. M. (2013). Lifespan development: The effects of typical aging on theory of mind. *Behavioural Brain Research*, *237*, 32–40. https://doi.org/10.1016/j.bbr.2012.09.020

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology / Revue Canadienne de Psychologie*, *45*(3), 255–287. https://doi.org/10.1037/h0084295

Paivio, A. (2013). Dual coding theory, word abstractness, and emotion: A critical review of Kousta et al. (2011). *Journal of Experimental Psychology: General*, *142*(1), 282–287. https://doi.org/10.1037/a0027004

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543. https://doi.org/10.3115/v1/D14-1162

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

Pi, Z., Hong, J., & Yang, J. (2017). Effects of the instructor's pointing gestures on learning performance in video lectures. *British Journal of Educational Technology*, *48*(4), 1020–1029. https://doi.org/10.1111/bjet.12471

Pi, Z., Zhang, Y., Zhu, F., Xu, K., Yang, J., & Hu, W. (2019). Instructors' pointing gestures improve learning regardless of their use of directed gaze in video lectures. *Computers & Education*, *128*, 345–352. https://doi.org/10.1016/j.compedu.2018.10.006

Potratz, J. R., Gildersleeve-Neumann, C., & Redford, M. A. (2022). Measurement Properties of Mean Length of Utterance in School-Age Children. *Language, Speech, & Hearing Services in Schools*, *53*(4), 1088–1101. https://doi.org/10.1044/2022_LSHSS-21-00115

Quine, W. V. O. (1960). *Word and object: An inquiry into the linguistic mechanisms of objective reference* (pp. xv, 294). John Wiley.

Rader, N. D. V., & Zukow-Goldring, P. (2010). How the hands control attention during early word learning. *Gesture*, *10*(2–3), 202–221. https://doi.org/10.1075/gest.10.2-3.05rad

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

Rowe, M. L., & Snow, C. E. (2020). Analyzing input quality along three dimensions: Interactive, linguistic, and conceptual. *Journal of Child Language*, *47*(1), 5–21. https://doi.org/10.1017/S0305000919000655

Rueckert, L., Breckinridge, C. R., Avila, A., & Trejo, T. (2017). Gesture enhances learning of a complex statistical concept. *Cognitive Research: Principles and Implications*, *2*(1). https://doi.org/10.1186/s41235-016-0036-1

Sanford, A. J. S., Sanford, A. J., Molle, J., & Emmott, C. (2006). Shallow Processing and Attention Capture in Written and Spoken Discourse. *Discourse Processes*, *42*(2), 109–130. https://doi.org/10.1207/s15326950dp4202_2

Segbers, J., & Schroeder, S. (2017). How many words do children know? A corpus-based estimation of children's total vocabulary size. *Language Testing*, *34*(3), 297–320. https://doi.org/10.1177/0265532216641152

Shi, J., Gu, Y., & Vigliocco, G. (2023). Prosodic modulations in child-directed language and their impact on word learning. *Developmental Science*, *26*(4), e13357. https://doi.org/10.1111/desc.13357

So, W. C., Sim Chen-Hui, C., & Low Wei-Shan, J. (2012). Mnemonic effect of iconic gesture and beat gesture in adults and children: Is meaning in gesture important for memory recall? *Language and Cognitive Processes*, *27*(5), 665–681. https://doi.org/10.1080/01690965.2011.573220

Sommers, M. S., Spehar, B., Tye-Murray, N., Myerson, J., & Hale, S. (2020). Age Differences in the Effects of Speaking Rate on Auditory, Visual, and Auditory-Visual Speech Perception. *Ear & Hearing*, *41*(3), 549–560. https://doi.org/10.1097/AUD.0000000000000776

Valle, A., Massaro, D., Castelli, I., & Marchetti, A. (2015). Theory of Mind Development in Adolescence and Early Adulthood: The Growing Complexity of Recursive Thinking Ability. *Europe's Journal of Psychology*, *11*(1), Article 1. https://doi.org/10.5964/ejop.v11i1.829

van Wermeskerken, M., & van Gog, T. (2017). Seeing the instructor's face and gaze in demonstration video examples affects attention allocation but not learning. *Computers & Education*, *113*, 98–107. https://doi.org/10.1016/j.compedu.2017.05.013

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

Vigliocco, G., Motamedi, Y., Murgiano, M., Wonnacott, E., Marshall, C., Milán-Maillo, I., & Perniss, P. (2019). *Onomatopoeia, gestures, actions and words: How do caregivers use multimodal cues in their communication to children?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/v263k

Wagner, M. (2020). Prosodic Focus. In *The Wiley Blackwell Companion to Semantics* (pp. 1–75). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118788516.sem133

Wojcik, E. H., Zettersten, M., & Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, *13*(4), e1596. https://doi.org/10.1002/wcs.1596

Zdrazilova, L., Sidhu, D. M., & Pexman, P. M. (2018). Communicating abstract meaning: Concepts revealed in words and gestures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170138. https://doi.org/10.1098/rstb.2017.0138