J Cult Cogn Sci (2019) 3:103–112
https://doi.org/10.1007/s41809-019-00047-z

REVIEW

Language and perception: Introduction to the Special Issue "Speakers and Listeners in the Visual World"

Mila Vulchanova · Valentin Vulchanov · Isabella Fritz · Evelyn A. Milburn

Received: 12 September 2019 / Revised: 28 September 2019 / Accepted: 1 October 2019 / Published online: 14 October 2019

Abstract Language and perception are two central cognitive systems. Until relatively recently, however, the interaction between them has been examined only partially and not from an over-arching theoretical perspective. Yet it has become clear that linguistic and perceptual interactions are essential to understanding both typical and atypical human behaviour. In this editorial, we examine the link between language and perception across three domains. First, we present a brief review of work investigating the importance of perceptual features, particularly shape bias, when learning names for novel objects—a critical skill acquired during language development. Second, we describe the Visual World Paradigm, an experimental method uniquely suited to investigate the language-perception relationship. Studies using the Visual World Paradigm demonstrate that the relationship between linguistic and perceptual information during processing is both intricate and bi-directional: linguistic cues guide interpretation of visual scenes, while perceptual information shapes interpretation of linguistic input. Finally, we turn to a discussion of co-speech gesture focusing on iconic gestures which depict aspects of the visual world (e.g., motion, shape). The relationship between language and these semantically-meaningful gestures is likewise complex and bi-directional. However, more research is needed to illuminate the exact circumstances under which iconic gestures shape language production and comprehension. In conclusion, although strong evidence exists supporting a critical relationship between linguistic and perceptual systems, the exact levels at which these two systems interact, the time-course of the interaction, and what is driving the interaction, remain largely open questions in need of future research.

Keywords Language · Perception · Language development · Language processing · Gesture

M. Vulchanova (✉) · V. Vulchanov · I. Fritz · E. A. Milburn
Language Acquisition and Language Processing Lab, Norwegian University of Science and Technology, Edvard Bulls veg 1, 7034 Trondheim, Norway
e-mail: mila.vulchanova@ntnu.no

Introduction

Language and perception are two central cognitive systems. Until relatively recently, however, the interaction between them has been examined only partially and not from an over-arching theoretical perspective (e.g. Miller and Johnson-Laird 1976). Yet it has become clear that language and perception interactions are essential to understand both typical and atypical human behaviour. Recent work in 'embodied cognition' and 'cognitive linguistics' has shown that language processing involves the construction of situation models and early activation of perceptual representations (see Barsalou 2009 for review).

Beyond these empirical demonstrations, though, there is a notable absence of an explanatory framework in which language-perception interactions can be understood (see for example Chatterjee 2010).

There is a rich bi-directional interface between language and perception. Visual perceptual experience informs language and the conceptual system and can shape language processing. At the level of sound, the visual cues of speech can enhance speech perception or even distort it, as demonstrated in the well-known McGurk effect (McGurk and MacDonald 1976; MacDonald and McGurk 1978). Visual information has been shown to activate (prime) language-related information early in development (Mani and Plunkett 2010). It is also the case that atypically developing children display a problem in matching object-images to corresponding linguistic labels (von Koss Torkildsen et al. 2007). However, the mechanism underlying this interaction (and its failure in some populations) has not been identified. Further open questions concern the extent to which visual perception contributes to word meaning (in long-term memory), and whether the comprehension of certain categories of words depend on the visual system. Here evidence is mixed and existing accounts are conflicting (Bedny et al. 2008; Bedny and Caramazza 2011; Glenberg and Gallese 2012; Pulvermüller 2012).

Language in turn also influences perception at several levels. Language mediates eye-movements to images present immediately in the visual context, as demonstrated in studies employing the visual-world paradigm (VWP) (Cooper 1974; Tanenhaus et al. 1995; Spivey et al. 2002; Allopenna et al. 1998; Altmann and Kamide 1999), and language also mediates motion processing of visual stimuli (Coventry et al. 2013). However, while existing studies document the effect of language context, and provide evidence that speakers rely heavily on linguistic cues in deciding what to anticipate as the speech signal unfolds, they fail to show what is the exact nature of the prediction, and the level at which linguistic and visual information integrate (Magnuson 2019).

The most straightforward explanation of why language and perception are inextricably related is the fact that we can talk about what we perceive. Furthermore, perceptual terms which are grounded in spatial relations in the world, including motion, are often used across languages to form analogies for the expression of more abstract terms, such as e.g. time. More importantly, in Pylyshyn's apt formulation, the main reason for the inherent relationship between language and perception is that "the perceptual system is the primary means through which language acquires a semantics" (Pylyshyn 1978). In an early vision of artificial intelligence systems, he observes that a system with a knowledge data-base and a language processor might succeed in carrying out a coherent dialogue, but without a perceptual component, it "would not know what it was talking about". This observation highlights an important aspect of the two systems, and their interrelationship, e.g., the fact that there are both intra-systemic relations and inter-systemic relations. Thus, in language, there are relations between words based on their linguistic features, such as phonological similarity, whereby a preceding word may prime a following word purely based on some phonological overlap between the two. Priming effects are also observed on the basis of semantic association. The fact that such effects are also observed outside of any context, whether linguistic or visual, supports the idea that they are due to purely intra-linguistic relations as a result of how these words are stored in long-term memory (the "mental lexicon", cf Jackendoff and Jackendoff 2002; Altmann 2001), and how these words are associated in the lexical network, e.g., reflecting neighbourhood effects, frequency effects. In a similar way, and independently of language, the perceptual system processes and stores "percepts" or representations of objects and events. Situated and embodied cognition theories claim that the predictive power of human intelligence resides in simulation which implements "the concepts that underlie knowledge" (Barsalou 2009). Currently perceived situations then "activate situated conceptualizations that produce predictions via simulations on relevant modalities" based on an inference mechanism. Even though Barsalou's suggestion assumes multi-modal simulation associated with frequently experienced situations, the integration of intra-linguistic and extra-linguistic (world–world) information is not addressed in detail. Exactly how and at what level of processing do linguistic and perceptual information integrate, at what temporal scale, and which of the two dominates in this process is still a matter of debate. In addition, and more importantly, we still lack comprehensive theoretical models of this interaction (cf. the critical discussion in Magnuson 2019).

In the current paper, we review evidence from three specific domains where the bi-directional relation between language and perception is of particular relevance: the importance of object features and affordances in the acquisition of object labels; language-mediated eye movements in the presence of visual context; and gesture as part of an integrated language-vision communication system.

The developmental perspective: learning object labels

The acquisition of object names is often described as a specific challenge for the infant. Following on Quine's (1960) *gavagai* thought experiment, what a specific word refers to is difficult to determine. In the context of a rabbit scurrying through a field, this expression may refer to the colour of the rabbit, it's fur, the event of scurrying, its' paws or legs, or express an evaluative attitude, e.g., "What a marvellous creature!". Still, children emerge successful from the process of learning words and the categories they refer to. It has been suggested that, in this process, visual object recognition plays an important role.

In seminal work over the past couple of decades Linda Smith has studied the underlying cognitive mechanisms that not only accompany, but also determine, word learning in infants and toddlers. One well-attested developmental phenomenon is the so-called *shape bias* in the acquisition of common nouns, which often refer to object categories. The shape bias is usually documented in experiments where infants are exposed to novel objects and their novel names, whereby the objects in the stimulus design can be either grouped on the basis of colour, shape or texture. These experiments confirm the hypothesis that early on in development, infants primarily attend to, and rely on, shape similarity over other object features, such as colour or texture, in attributing the novel label (Landau et al. 1988). They also generalise a newly acquired name to novel instances by shape. Attention to object shape has also been found to be a reliable predictor of noun vocabulary growth (Smith et al. 2002; Poulin-Dubois 1999; Gershkoff-Stowe and Smith 2004).

Independently, it has also been shown that shape plays an important role in early object recognition, and that a robust shape bias and object recognition develop between 18 and 24 months. Crucially, the ability to recognise common objects from sparse shape representations, develops at this stage (Smith 2009). Like the shape bias, this ability is more strongly linked to early vocabulary size than age (Smith 2003; Pereira and Smith 2009). The sparse representation idea originates from Biederman's (1995) proposal that the internal representations of objects that humans form are based on sparse geometric models ("caricatures") of the 3-dimensional structure of object shape.

Capitalising on previous research documenting the shape bias and the trajectory of sparse object recognition, Yee et al. (2012) studied the development and the relationship between these two abilities in the same cohort of children. In two experiments, Yee et al. (2012) focused on the period between 18 and 24 months where the development of both skills has been attested to peak and stabilise. Both experiments used a Shape Bias task, a Shape Caricature recognition task and an Object Recognition task. The tasks were preceded by practice trials with 3-dimensional common objects (a flower, a spoon and a duck) whereby infants were familiarised with the experimental tasks by being provided with an illustration of the procedure illustrated on the basis of identifying and retrieving an object upon hearing its name. The experimental tasks were three-alternative forced choice tasks where the child was asked to select an object by its name. In addition, data were collected from the parents using the MacArthur-Bates Communicative Development Inventory (CDI) checklist.

The first experiment exploited a cross-sectional design (n = 55, M = 27, age range 18–24 months). For the purposes of analysis, participating children were divided into three noun vocabulary groups: low, with 55 or less nouns, medium, with 56–125 nouns and high, for children with more than 125 nouns in their vocabulary according to the CDI results. Pairwise linear correlations among the variables in that study revealed that children's performance in the Shape Bias and the Shape Caricature Recognition task were correlated and this performance was correlated to noun, and total vocabulary size. Furthermore, vocabulary was a better predictor of scores in both the Shape Bias and Shape Caricature Recognition task than age. These results come to suggest that shape caricature development may pave the way for the development of shape bias in novel noun acquisition. Thus, children can only demonstrate a stable shape bias after they can

already recognise objects from sparse shape representations. These predictions were borne out by the longitudinal data from the second experiment where 10 infants were tested once every 3 weeks starting when participants were 18 months old and until they turned 24 months. The analysis of variance of the data revealed a main effect of vocabulary size whereby performance on each task increased as a function of vocabulary size, and a main effect of task. Post hoc analyses revealed that children performed better on the Object Recognition task than the Shape Caricature task, and better on the Shape Caricature task than the Shape Bias task. Additional analyses carried out to establish the temporal ordering of success at shape recognition and shape bias based on a 0.62 correct criterion revealed a pattern of development where the skills required to perform adequately on the Shape Caricature task did not depend completely on the set of skills necessary to perform on the Shape Bias task. This leads to the conclusion that success at the Shape Recognition task actually precedes success at the Shape Bias task. This study thus supports the idea that shape bias, which is an important prerequisite for successful word acquisition, and visual object recognition from sparse representations are developmentally related, and, in addition, that robust object recognition is a prerequisite for, and supports the development of, the shape bias in typically developing children. Further support for these results comes from the evidence from late talkers, where both abilities have been shown to be absent (Jones and Smith 2005).

The importance of shape as central among perceptual properties of objects is further confirmed by findings of impaired categorisation skills and impaired shape bias in children with autism potentially leading to atypical categorisation and problems with word learning and semantics (Hartley and Allen 2014a; Field et al. 2016; Abdelaziz et al. 2018). There is also evidence that children with autism are not as successful in categorising objects on the basis of black and white contour sketches in comparison to more realistic colour images of the object (Hartley and Allen 2014b). Furthermore, symbolic understanding of pictures in children with autism in the study by Hartley and Allen (2014b) was facilitated by iconicity, and particularly colour, but not language. This comes to suggest that impaired object categorisation and atypical reliance on object features in categorisation may be linked to the attested absence of the shape bias in word learning in autism, thus supporting the evidence from the Yee et al. (2012) study. No research, to the best of our knowledge, has directly compared the emergence of object recognition from sparse representations and the development of the shape bias in that population.

Interestingly, these findings are not consistent with the well-documented exceptional ability in high functioning and highly-verbal individuals with autism on abstract pattern recognition, as reflected in superior performance on tasks such as the Block Design from the Wechsler scales or matrices (Raven 1998; Dawson et al. 2007; Vulchanova et al. 2012). These results may suggest a dissociation between abstract pattern recognition on the one hand, and recognising and categorising real objects from their characteristic features, on the other. This points in the direction of atypical association between the symbols (words, intra-linguistic relations; abstract images) and their referents out in the world (the real objects). This hypothesis is in need of further investigation in well-designed and controlled testing environments.

Language comprehension and visual context: an integrated perspective

An important tool for investigating the bidirectional interface between visual perception and language processing has been the Visual World Paradigm (VWP). In 1974, Cooper showed participants visual displays while playing short passages and noticed that participants were likely to look at objects in the display that were also referred to in the text. Participants' eye movements were also closely time-locked to the text that they heard. Although not popularised in wider psycholinguistic researcher until much later (Tanenhaus et al. 1995), the basic framework of Cooper's study is still used in VWP studies today.

The primary advantage of the VWP is the strong, systematic, relationship between the auditory linguistic stimulus and eye movements around the visual display. Participants' eye movements are recorded and analysed, and where, when, and how fast participants look can inform research questions about how they are processing the language they hear. Another advantage of the VWP is its flexibility: it can be used to address questions at levels of language processing from the phonological to the discursive. The VWP has been used to study, among others, phonological effects on

word recognition (Allopenna et al. 1998), facilitative effects of selectional restrictions (Altmann and Kamide 1999), the mental representation of scenes (Altmann and Kamide 2009), and the role of event knowledge in predictive processing (Milburn et al. 2016). Furthermore, the VWP is easy for participants to complete: participants are often required only to "look and listen" or to follow simple instructions, making it ideal for use in populations, such as older adults (Hayes et al. 2016), people with aphasia (Mack et al. 2013), young children (Borovsky et al. 2012) or deficit populations (Norbury 2017; Vulchanova et al. 2019).

Critically, eye movements in the VWP reflect complex, systematic, interactions between linguistic and visual contexts, and untangling the factors involved in these interactions and the ways in which these interactions motivate eye movements has proven to be challenging (Huettig et al. 2011; Magnuson 2019). One way of thinking about the relationship between linguistic input and visual context in the VWP is to consider the visual context as a frame onto which the linguistic input is projected: participants interpret the linguistic input within the context of the visual display. To illustrate this, consider studies examining how scene information and event knowledge drive (predictive) eye movements. In a VWP study using naturalistic scenes, Milburn et al. (2016) examined eye movements to targets driven both by information contained in the auditory verb stimulus and world knowledge conveyed by the scene depicted in the visual display. Participants viewed scenes and listened to sentences containing either constraining or unconstraining verbs. For example, the constraining verb *drink* is most likely to specify liquid items as direct objects (verified by cloze norming). When hearing the sentence *Someone will drink the___*, participants can predict that something liquid will follow the verb *drink* even without a visual display. In contrast, the unconstraining verb *throw* can have a wide range of possible fillers in the direct object position (again verified by cloze norming). However, when accompanied by a scene depicting a bride at a wedding, participants showed rapid eye movements to the bouquet of flowers despite the presence of many throwable objects in the scene, because in a wedding context the bouquet of flowers is the most appropriate direct object for *throw*. Participants therefore used the visual context to constrain their interpretation of the linguistic stimulus. Most critically for the current discussion, however, Milburn and colleagues as part of their stimulus norming showed the visual stimuli to participants and asked them to briefly describe what might happen next in the scene. They found that participants gave a wide range of answers, only referring to the target objects about a third of the time. This suggests that, during the eyetracking portion of the experiment, participants in turn used the linguistic stimuli to constrain their interpretation of the scenes: although each scene contained rich non-linguistic semantic information pointing to a wide variety of possible events, participants used the linguistic information provided in the auditory stimulus to constrain and guide their interpretation of this semantic information, leading them to the appropriate direct object. Thus, the relationship between linguistic and visual context in the visual world is both complex and bi-directional.

This complex relationship raises two critical questions for researchers interested in the interactions between language and perception. First, psycholinguistic research using the VWP depends on the tight linkage between eye movements and language processing, but the mechanism underlying this linkage is unclear, and often left implicit (Magnuson 2019). In this volume, Magnuson reviews four possible linking hypothesis underlying fixations in the VWP, pointing out that the nature of the linking hypothesis assumed by the researcher has implications for the interpretation of eye movement patterns, and that fixations to one item in a display over another may be driven by competition, co-activation, or even facilitation. Lack of formal linking hypotheses can therefore result in experimental results being attributed to the wrong processing level.

Paralleling this work, a second critical question for researchers using the VWP is the nature of the mechanisms underlying context effects on eye movements and language processing. In this issue, Knoeferle discusses the factors that may predict stronger or weaker context effects during language processing, paying particular attention to how language may be grounded in a visual context with reference to the comprehender's knowledge of the world. This perspective is congruent with that proposed by Altmann and Kamide (2007), in which language processing reflects an increasingly complex mental world as the comprehender draws on their knowledge of scenes and

events [see also work by McRae and Matsuki (2009)showing rapid effects of world knowledge on comprehension]. Critically, Knoeferle (2019) proposes that not all effects of visual context may be caused by the same underlying mechanism—distinct language-world relations elicit distinct context effects—and demonstrates a strong role for comprehender-specific characteristics modulating context effects during comprehension.

Taken together, the questions raised in these reviews collectively call for more detailed models of the bidirectional interface between vision and language. Given that visual context affects language processing—and therefore drives eye movements—in complex ways, fuller conceptions of how specific language-world relationships elicit specific context effects can inform the development of clear linking hypothesis between eye movements and language processing. More broadly, this work can deepen our understanding of linguistic and perceptual interactions.

Language and gesture

Within the growing field of gesture studies, gestures that accompany speech are seen as part of the language system. Research over the past three decades has accumulated evidence suggesting that language and gesture are part of an integrated common system of communication (McNeill 1992, 2015). For example, brain imaging studies in adults demonstrate that the process of meaning integration between speech and co-occurring gesture involves classic language areas in the left frontal and temporal lobes and their right hemisphere homologues (Andric and Small 2012; Dick et al. 2014).

Gesture has been shown to play an important role in communicative development. Gesture development, and the production of deictic gesture pointing in particular, both predates and predicts later language development in typically developing children (Iverson and Goldin-Meadow 2005). In contrast, delayed and atypical gesture development has been documented in children with autism (see Ramos-Cabo et al. 2019 for a qualitative review, and Ramos-Cabo et al. in preparation), thus suggesting a possible break-down in an integrated gesture-language communication system. Furthermore, the functional neuro-anatomy of gesture-speech integration has been shown to vary depending on individual differences in how gesture is processed in children and in the course of development. Thus, the gesture-speech integration network was differentially activated to meaningful gestures accompanying stories only in the children who found meaning in the gestures they were shown, but not for children who did not show this behaviourally (Demir-Lira et al. 2018). Those studies also demonstrate that the neuro-anatomy of gesture-speech integration becomes more refined and less widely distributed over the course of development.

The close interaction between the visual world and co-speech gestures is especially apparent in iconic gestures. Iconic gestures depict something concrete resembling an action, event or object (McNeill 1992). From among iconic gestures, motion event gestures that refer to events like “rolling down” or “jumping up” are the most comprehensively studied. In gesture-speech production studies of this type, participants typically watch a cartoon which they then have to retell to an interlocutor. This method is most appropriate for gesture elicitation, especially when comparing video stimuli to static stimuli. Those comparisons reveal that, unlike dynamic stimuli, static stimuli appear to reduce gesture frequency (McNeill 2005; Hostetter and Hopkins 2002), thus confirming that the communication system interacts closely with the perceptual system. Importantly, research using the method of retelling a video stimulus has shown that the content of a gesture is shaped by both the visual input and the language being produced.

Studies on gestures produced in the context of motion events found that gestural encoding of a motion event depended on how the event was syntactically encoded, even when participants watched the same stimulus cartoon (e.g., Fritz et al. 2019; Akhavan et al. 2017; Kita and Özyürek, 2003). These findings highlight the close relationship between gesture and speech production. More relevant from a perception perspective are studies showing that co-speech gestures can depict aspects of a stimulus which are not present in the spoken modality. For example, Kita and Özyürek (2003) looked at whether the direction of the lateral movement of path gestures matched the direction of this movement experienced through the visual stimulus (e.g., A cat swings across the street; from the right side of the screen to the left). Indeed, the direction encoded in the gesture predominantly

matched the one participants viewed in the video, although this information was never verbalised. Observations like that led to the Interface Hypothesis, proposing that gesture and speech are tightly linked throughout the production process. However, gestures originate outside of the speech production system, which allows them to encode aspects of the visual world that are not present in speech (Kita 2000; Kita and Özyürek 2003), and thus supplement the message conveyed by language (Iverson and Goldin-Meadow 2005). It is exactly iconic gestures which add information, e.g., for disambiguation, that activate the classical language areas (Dick et al. 2014).

Further evidence of the tight relationship between speech and gesture comes from neuro-physiological research on iconic gestures. EEG/ERP studies have shown that information provided via co-speech gestures is processed similarly to speech input (e.g., Kelly et al. 2004; Holle and Gunter 2007; Özyürek et al. 2007), indicating that people can extract meaning from gestures. Moreover, research has found that co-speech gestures become part of a comprehender's discourse model, indicating that information from both gesture and speech is integrated and used to form a unified meaning representation of the input. This suggests that comprehenders do not distinguish between information perceived via the visual or auditory channel (see Özyürek 2014, for a review), and support the speech-gesture integration neuro-anatomy findings (Dick et al. 2014; Demir-Lira et al. 2018). Behavioural studies further highlight the communicative function of gestures. For example, information exclusively conveyed via gesture in a stimulus can be picked up in a participant's retelling of the stimulus (e.g., Cassell et al. 1999). Furthermore, in a production study, Alibali et al. (2001) manipulated whether participants could see their interlocuter or not. They found that visibility of the interlocuter resulted in the production of more iconic gestures.

Despite an increasing interest in multimodal language comprehension, there is little research on the eye-gaze of comprehenders during multimodal interactions. Generally, listeners' eyes fixate on gestures very rarely. Eye-tracking studies demonstrate that as much as 90–95% of the time within a face-to-face interaction, listeners fixate on the speaker's face (Gullberg and Holmqvist 2006). Thus, gestures are mainly perceived through peripheral vision. However, fixations on gestures increase under certain circumstances. Gullberg and Kita (2009) found that this is the case for gestures that include a hold before a next gesture begins. Increased fixations on such gesture holds might be the result of the sudden change in the visual field. Alternatively, holds may be a challenge for peripheral vision, because if there is no motion that the comprehender can perceive, then a shift in eye-gaze is necessary to take up any new visual information. Furthermore, the same study found increased gesture fixations for gestures that were first fixated by the speaker. The influence of the speaker's gaze on the attention the comprehender is paying to gestures suggests a social component of gestural fixations.

The distribution of attention between gestures and other visual input would be interesting to study within the visual world paradigm. Although this paradigm has been used in a co-speech gesture study (Silverman et al. 2010), the authors did not report the proportions of fixations to the centre-screen gesture stimulus. Generally, sign language studies using the visual world paradigm have demonstrated that this paradigm can be used to investigate language processing in the visual modality (see for example Wienholz and Lieberman, 2019). Future studies of multimodal language comprehension could employ the visual world paradigm to answer questions about gesture-speech integration and whether gestures play a role in predictive language processing.

Summary

We have reviewed three specific domains where the tight interface between language and perception are most evident. In early cognitive and language development, the perceptual affordances of objects, and specifically, object shape play an important role in the acquisition of object labels and drive vocabulary growth. Visual context constrains the way listeners interpret spoken language, while speech serves to guide listeners' attention to visually present entities, as revealed in studies employing the Visual World Paradigm (VWP). In turn, gestures, which originate in the visual and motor systems, interact systematically with language, either complementing or supplementing the verbal message. While research has provided evidence of the interaction, the exact levels at which these two systems interact, the time-course of

the interaction, and more importantly, what is driving the interaction, remain largely open questions in need of future research.

References

Abdelaziz, A., Kover, S. T., Wagner, M., & Naigles, L. R. (2018). The shape bias in children with autism spectrum disorder: potential sources of individual differences. *Journal of Speech, Language, and Hearing Research, 61*(11), 2685–2702.

Akhavan, N., Nozari, N., & Göksun, T. (2017). Expression of motion events in Farsi. *Language, Cognition and Neuroscience, 32*(6), 792–804.

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language, 44*(2), 169–188.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419–439.

Altmann, G. T. (2001). The language machine: Psycholinguistics in review. *British Journal of Psychology, 92*(1), 129–170.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247–264.

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*(4), 502–518.

Altmann, G. T., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition, 111*(1), 55–71.

Andric, M., & Small, S. (2012). Gesture's neural language. *Frontiers in Psychology, 3,* 99.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1281–1289.

Bedny, M., & Caramazza, A. (2011). Perception, action, and word meanings in the human brain: the case from action verbs. *Annals of the New York Academy of Sciences, 1224*(1), 81–95.

Bedny, M., McGill, M., & Thompson-Schill, S. L. (2008). Semantic adaptation and competition during word comprehension. *Cerebral Cortex, 18*(11), 2574–2585.

Biederman, I. (1995). *Visual object recognition* (vol. 2). Cambridge, MA: MIT press.

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112*(4), 417–436.

Cassell, J., McNeill, D., & McCullough, K.-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition, 7*(1), 1–34.

Chatterjee, A. (2010). Disembodying cognition. *Language and Cognition, 2*(1), 79–116.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*(1), 84–107.

Coventry, K. R., Christophel, T. B., Fehr, T., Valdés-Conroy, B., & Herrmann, M. (2013). Multiple routes to mental animation: Language and functional relations drive motion processing for static images. *Psychological Science, 24*(8), 1379–1388.

Dawson, M., Soulières, I., Gernsbacher, M. A., & Mottron, L. (2007). The level and nature of autistic intelligence. *Psychological Science, 18*(8), 657–662. https://doi.org/10.1111/j.1467-9280.2007.01954.

Demir-Lira, O. E., Asaridou, S., Beharelle, A. R., Holt, A., Goldin-Meadow, S., & Small, S. (2018). Functional neuroanatomy of gesture-speech integration in children varies with individual differences in gesture processing. *Developmental Science, 21*(5), e12648.

Dick, A. S., Mok, E. H., Raja Beharelle, A., Goldin-Meadow, S., & Small, S. L. (2014). Frontal and temporal contributions to understanding the iconic co-speech gestures that accompany speech. *Human Brain Mapping, 35*(3), 900–917.

Field, C., Allen, M. L., & Lewis, C. (2016). Are children with autism spectrum disorder initially attuned to object function rather than shape for word learning? *Journal of Autism and Developmental Disorders, 46,* 1210–1219. https://doi.org/10.1007/s10803-015-2657-5.

Fritz, I., Kita, S., Littlemore, J., & Krott, A. (2019). Information packaging in speech shapes information packaging in gesture: The role of speech planning units in the coordination of speech-gesture production. *Journal of Memory and Language, 104,* 56–69.

Gershkoff-Stowe, L., & Smith, L. B. (2004). Shape and the first hundred nouns. *Child Development, 75*(4), 1098–1114.

Glenberg, A. M., & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex, 48*(7), 905–922.

Gullberg, M., & Holmqvist, K. (2006). What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics and Cognition, 14*(1), 53–82.

Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behavior, 33,* 251–277.

Hartley, C., & Allen, M. L. (2014a). Brief report: Generalisation of word–picture relations in children with autism and

typically developing children. *Journal of Autism and Developmental Disorders, 44*(8), 2064–2071.
Hartley, C., & Allen, M. L. (2014b). Intentions versus resemblance: Understanding pictures in typical development and autism. *Cognition, 131*(1), 44–59.
Hayes, R. A., Dickey, M. W., & Warren, T. (2016). Looking for a location: Dissociated effects of event-related plausibility and verb—argument information on predictive processing in aphasia. *American Journal of Speech-Language Pathology, 25*(4S), S758–S775.
Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience, 19,* 1175–1192.
Hostetter, A. B., & Hopkins, W. D. (2002). The effect of thought structure on the production of lexical movements. *Brain and Language, 82,* 22–29.
Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*(2), 151–171.
Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture Paves the way for language development. *Psychological Science, 16*(5), 367–371.
Jackendoff, R., & Jackendoff, R. S. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.
Jones, S. S., & Smith, L. B. (2005). Object name learning and object perception: A deficit in late talkers. *Journal of Child Language, 32*(1), 223–240.
Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89*(1), 253–260.
Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 162–185). Cambridge: Cambridge University Press.
Knoeferle, P. J. (2019). Predicting (variability of) context effects in language comprehension. *Journal of Cultural Cognitive Science*. https://doi.org/10.1007/s41809-019-00025-5.
Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language, 48*(1), 16–32.
Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*(3), 299–321.
MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception and Psychophysics, 24,* 253–257. https://doi.org/10.3758/BF03206096.
Mack, J. E., Ji, W., & Thompson, C. K. (2013). Effects of verb meaning on lexical integration in agrammatic aphasia: Evidence from eyetracking. *Journal of Neurolinguistics, 26*(6), 619–636.
Magnuson, J. S. (2019). Fixations in the visual world paradigm: where, when, why? *Journal of Cultural Cognitive Science*. https://doi.org/10.1007/s41809-019-00035-3.
Mani, N., & Plunkett, K. (2010). In the infant's mind's ear: Evidence for implicit naming in 18-month-olds. *Psychological Science, 21*(7), 908–913.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748. https://doi.org/10.1038/264746a0.
McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.
McNeill, D. (2005). *Gesture and thought*. Chicago: The University of Chicago Press.
McNeill, D. (2015). Gesture in linguistic. In: *International encyclopedia of the social and behavioral sciences* (2nd edn, Vol. 10, pp. 109–120. Oxford: Elsevier.
McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass, 3*(6), 1417–1429.
Milburn, E., Warren, T., & Dickey, M. W. (2016). World knowledge affects prediction as quickly as selectional restrictions: Evidence from the visual world paradigm. *Language, Cognition and Neuroscience, 31*(4), 536–548.
Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Cambridge: Belknap Press.
Norbury, C. (2017). Eye-tracking as a window on language processing in autism spectrum disorder. In L. Naigles (Ed.), *Innovative investigations of language in autism* (pp. 13–33). New York: APA Books. https://doi.org/10.1037/15964-002.
Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1651), 213–296.
Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience, 19*(4), 605–616.
Pereira, A. F., & Smith, L. B. (2009). Developmental changes in visual object recognition between 18 and 24 months of age. *Developmental Science, 12*(1), 67–80.
Poulin-Dubois, D. (1999). Infants' distinction between animate and inanimate objects: The origins of naive psychology. In P. Rochat (Ed.), Early social cognition: Understanding others in the first months of life (pp. 257–280). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
Pulvermüller, F. (2012). Meaning and the brain: The neurosemantics of referential, interactive, and combinatorial knowledge. *Journal of Neurolinguistics, 25*(5), 423–459.
Pylyshyn, Z. W. (1978). What has language to do with perception? Some speculations on the Lingua Mentis. In Theoretical Issues in Natural Language Processing-2.
Quine, W. (1960). *Word and object*. Cambridge: MIT Press.
Ramos-Cabo, S., Vulchanov, V., & Vulchanova, M. (2019). Gesture and language trajectories in early development: An overview from the autism spectrum disorder perspective. *Frontiers in Psychology, 10,* 1211.
Ramos-Cabo, S., Vulchanov, V., & Vulchanova, M. (in preparation). Non-verbal communication patterns in typically developing children, children with autism and children at high risk for autism in a gesture elicitation interactive task.
Raven, J. C. (1998). *Raven's progressive matrices and vocabulary scales*. Oxford: Oxford Pyschologists Press.
Silverman, L. B., Bennetto, L., Campana, E., & Tanenhaus, M. K. (2010). Speech-and-gesture integration in high functioning autism. *Cognition, 115*(3), 380–393.

Smith, L. B. (2003). Learning to recognize objects. *Psychological Science, 14*(3), 244–250.

Smith, L. B. (2009). From fragments to geometric shape: Changes in visual object recognition between 18 and 24 months. *Current Directions in Psychological Science, 18*(5), 290–294.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science, 13*(1), 13–19.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology, 45*(4), 447–481.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632–1634.

von Koss Torkildsen, J., Syversen, G., Simonsen, H. G., Moen, I., & Lindgren, M. (2007). Brain responses to lexical-semantic priming in children at-risk for dyslexia. *Brain and Language, 102*(3), 243–261.

Vulchanova, M., Chahboun, S., Galindo-Prieto, B., & Vulchanov, V. (2019). Gaze and motor traces of language processing: Evidence from autism spectrum disorders in comparison to typical controls. *Cognitive Neuropsychology*. https://doi.org/10.1080/02643294.2019.1652155.

Vulchanova, M., Talcott, J. B., Vulchanov, V., Stankova, M., & Eshuis, H. (2012). Morphology in autism spectrum disorders: Local processing bias and language. *Cognitive Neuropsychology, 29*(7–8), 584–600.

Wienholz, A., & Lieberman, A. M. (2019). Semantic processing of adjectives and nouns in American Sign Language: Effects of reference ambiguity and word order across development. *Journal of Cultural Cognitive Science*. https://doi.org/10.1007/s41809-019-00024-6.

Yee, M. N., Jones, S. S., & Smith, L. B. (2012). Changes in visual object recognition precede the shape bias in early noun learning. *Frontiers in Psychology, 3,* 533.