TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 00 (2024) 1–11

ISSN: 1756-8765 online
DOI: 10.1111/tops.12728

This article is part of the topic "Parallelism in the Architecture of Language," Giosuè Baggio, Neil Cohn, and Eva Wittenberg (Topic Editors).

Extending the Architecture of Language From a Multimodal Perspective

Peter Hagoort,[a,b] Aslı Özyürek[a,b]

[a]*Max Planck Institute for Psycholinguistics, Nijmegen*
[b]*Donders Institute for Brain, Cognition and Behaviour, Nijmegen*

Received 24 September 2023; received in revised form 26 February 2024; accepted 27 February 2024

Abstract

Language is inherently multimodal. In spoken languages, combined spoken and visual signals (e.g., co-speech gestures) are an integral part of linguistic structure and language representation. This requires an extension of the parallel architecture, which needs to include the visual signals concomitant to speech. We present the evidence for the multimodality of language. In addition, we propose that distributional semantics might provide a format for integrating speech and co-speech gestures in a common semantic representation.

Keywords: Co-speech gestures; Distributional semantics; Memory, Unification, Control (MUC) model; Multimodal; Speech; Unification

1. The unimodal parallel architecture

Jackendoff's parallel architecture (PA) proposal (2002, 2007) is an important correction on the syntactocentric views within the Chomskyan tradition in linguistics (e.g., Chomsky, 1995;

Correspondence should be sent to Peter Hagoort and Aslı Özyürek, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands. Email: peter.hagoort@mpi.nl; asli.ozyurek@mpi.nl

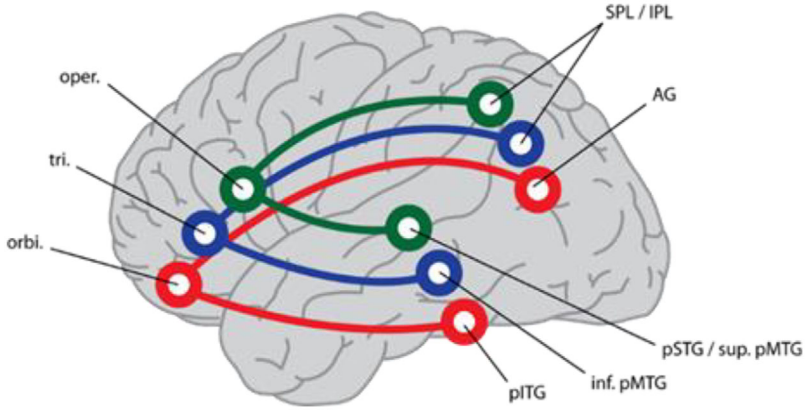


Fig. 1. A schematic drawing of the topographical connectivity pattern between the frontal and temporal/parietal cortex in the perisylvian language network, as revealed by resting-state fMRI (after Xiang et al., 2010). The strongest connections to the pars opercularis (oper.), pars triangularis (tri.), and pars orbitalis (orbi.) of Broca's region are shown. SPL/IPL, superior parietal lobule/inferior parietal lobule; AG, angular gyrus; pSTG: posterior superior temporal gyrus; sup. pMTG, superior posterior middle temporal gyrus; inf. pMTG: inferior posterior middle temporal gyrus; pITG, posterior inferior temporal gyrus.

Seuren, 2004). It was an important source of inspiration for the neurobiological memory, unification, and control model (Hagoort, 2005, 2014, 2017). In line with the PA architecture, this model assumes lexically specified building blocks for phonology, syntax, and semantics, while unification operations assemble larger structures from the lexical building blocks. The lexical building blocks are subserved by memory circuits in different parts of the temporal and parietal cortices. Unification, on the other hand, requires a dynamic network interaction of these areas with part of the left prefrontal cortex (Broca's area and adjacent cortex). Based on resting-state functional magnetic resonance imaging (fMRI) data (Xiang, Fonteijn, Norris, & Hagoort, 2010; Hagoort, 2017), the unification gradient in prefrontal areas is determined by the functional connectivity profile in left perisylvian cortex, with the semantic unification network most ventrally (in red), the syntactic unification network more dorsally (in blue), and the phonological unification network most dorsally (in green; see Fig. 1).

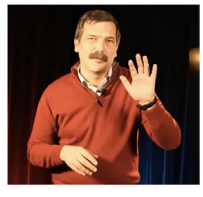
The unification areas in the prefrontal cortex are domain-general and not language-specific, in agreement with claims made by Jackendoff and Audring (2020): "this combinatorial procedure is uniform across the grammar, … all the differentiation lies in the declarative templates, which specify what is to be combined … it is a plausible combinatorial principle not only for linguistic structure but also for a variety of other cognitive domains" (p. 29).

An important aspect of the PA is the processing dynamics. Overwhelmingly, event-related potential (ERP) and Magneto-encephalography (MEG) studies on language processing show that the different information types (lexical, syntactic, phonological, pragmatic) are processed in parallel and influence the interpretation process incrementally, that is, as soon as the relevant pieces of information are available (Marslen-Wilson, 1989; Zwitserlood, 1989; Jackendoff, 1999, 2002; for speech and gesture, see Chu and Hagoort, 2014). We have referred to this PA feature as the immediacy principle (Hagoort & van Berkum, 2007; Hagoort, 2008).

Neurobiological evidence further supports this view. For instance, it has been found that information in spoken language and co-speech gestures are unified at the same moments in time (Özyürek, Willems, Kita, & Hagoort, 2007) and supported by the same brain structures for unification (Willems, Özyürek, & Hagoort, 2007) as the so-called purely linguistic types of information.

However, in one major aspect, the PA version proposed by Jackendoff (1999, 2002, 2007) is too limited and needs extension to be adequate in light of the linguistic facts and the processing details. The current PA model is inherently unimodal, in which all information is extracted from a unimodal visual (i.e., in reading) or auditory (i.e., in speech) input stream. However, as we will argue below, the human language faculty is inherently multimodal, in which visuospatial signals are continuously and effortlessly combined with the speech input (see also Cohn, 2016; Chon & Schilperoord, 2021) for extending this view to text-image relations). These visuospatial signals are not just add-ons to the linguistically relevant unimodal structures but instead are an integral part of the linguistic structures themselves. Linguistic theories missing out on the multimodal contributions might, therefore, be insufficient. For example, speakers of Turkish signal a negation in gesture well before the morpheme for negation (*-mi-ma-me*) appears in the sentence co-occurring mostly on the verb or even the noun before:

Ben	brokoli	sev-**mi**-yor-um
I	broccoli	like- **not**- PROG-1person



The negation case in Turkish is relevant for the processing account of word order. The negation gesture comes earlier (mostly on the verb) than the negation marker in speech. The listener hence "knows" that the statement is denied well before this becomes clear by the language marker itself. One could argue that this is relevant for a linguistic account of negation marker positioning in Turkish. A linguistic analysis based only on the position of the spoken negation morpheme in Turkish can, therefore, be misguided if the gestural marker for negation is not included in the analysis. Likewise, facial expressions can influence the truth value of the propositional content. If someone says "I like broccoli" one believes this to be a true proposition. However, if this sentence is accompanied by an ironic facial expression, then it is clear that in fact the speaker might not like broccoli. Hence, the truth value of the multimodal proposition is the opposite of what the sentence itself declares (see Ebert, 2024; Schlenker, 2019; for different types of implicatures and presuppositions gestures are considered to trigger in relation to information in speech).

Below we will, therefore, argue that the PA of language needs to be extended to include visuospatial markers of language in our linguistic analyses and processing models. This then

raises the issue of how the interface between speech and co-speech gestures should be thought of. One possibility is derived from distributional semantics (Boleda, 2020), in which spoken words and co-speech gestures could be represented by vectors in a multidimensional space. This common representational format enables the integration of information from speech and gestures. But before getting there, we will present the arguments in support of language as a multimodal system.

2. Arguments in favor of language as a multimodal system

There is growing consensus in language studies that language, in its primary face-to-face context is multimodal (e.g., Holler & Levinson, 2019; Perniss, 2018). That is, expressions in the visual modality, such as visible communicative movements (i.e., manual and facial gestures) universally accompany spoken languages. In addition, expressions in sign languages of deaf communities are as intrinsic to the nature of language as expressions in the vocal modality (e.g., Goldin-Meadow et al., 2015; Özyürek, & Woll, 2019; Perniss, Özyürek, & Morgan, 2015; Kita & Emmorey, 2023). As far as we know, all human languages that are used in face-to-face communication involve the visual modality (co-speech gestures in spoken languages and sign languages). No known language uses only speech for its expression. Furthermore, even though print can be used to represent speech, human communication prefers new technologies that allow visual graphemes to be integrated into text (e.g., emojis). Thus, these days even print is used multimodally.

Despite the prominent and universal omnipresence of visible bodily expressions in language, most theories of language and the so-called "design features" of language are based on characteristics of speech or text. These include the arbitrary form meaning mappings, the sequential, categorical, and single-channel features of language (e.g., Hockett, 1960). These features have influenced (psycho)linguistic, neurobiological, and computational models of language.

This theoretical bias notwithstanding, all spoken and sign languages embody the visible iconic and indexical (i.e., pointing), simultaneous and multichannel aspects of language; that is, speech and gesture can be expressed at the same time; in sign language, different articulators can represent different arguments simultaneously (Slonimska, Özyürek, & Capirci, 2022). Moreover, gestures and signs can represent meaning in an analog and iconic rather than in a purely categorical way. These phenomena were mostly not considered to be among the core defining features of language, as they contrast with arbitrary, categorical, and linear design features. Below, we first define what these visual features look like and also present evidence showing that they *are* an integral part of the language system.

3. Visual expressions in co-speech gestures

Co-speech gestures commonly accompany spoken language and differ in their forms, meanings, and functions (Clark, 1996; Kendon, 2004; McNeill, 1992, 2005). Some gestures like iconic gestures and beats are designed to go along with speech and can be hard to interpret without. Others, like emblems, can stand alone (e.g., an OK-gesture).

Emblems have an arbitrary relationship with the meaning they convey, similar to lexicalized words. Their interpretation can vary across cultures. Conversely, iconic gestures maintain a visually based connection between their form and the subject or action they denote, albeit with some cultural conventionalization of their form-meaning mapping. For instance, a stirring motion while talking about cooking resembles the actual stirring action. Yet, these gestures, though visually motivated, can be ambiguous without speech. Their meaning becomes clearer when paired with relevant speech, forming a co-expressive unit. Iconic gestures differ in their semiotic characteristics, representing objects, actions, events, or even locations in unique ways, emphasizing visual perspectives, spatial locations, relations, shapes, and sizes (Debreslioska, Özyürek, Gullberg, & Perniss, 2013; McNeill, 1992). They communicate depictively, visually reenacting referents in shared spaces.

Pointing gestures accompany demonstrative forms and pronouns, indicating referents or locations (e.g., Peeters & Özyürek, 2016; Azar, Backus, & Özyürek, 2019). These gestures can highlight concrete objects or abstract locations in the speaker's gesture space, reinforcing relationships between the gesture and speech. Beats, rhythmic hand movements, synchronize with speech, marking distinctions between new and old information, often aligning with prosodic emphasis (Rohrer, Delais-Roussarie, & Prieto, 2023).

Gestures occur universally across speaking communities, although frequency and cultural values associated with gestures might vary (Chu and Hagoort, 2014; Kita, 2009). The connection between speech and gestures seems innate; even congenitally blind individuals use gestures while speaking, indicating their deep-rooted nature in human communication (Iverson & Goldin-Meadow, 1998; Özçalışkan, Lucero, & Goldin-Meadow, 2016; Mamus, Speed, Özyürek, & Majid, 2022).

4. Integration of speech and co-speech gestures in production and across languages

Gestures, while frequently used, might not be deemed fundamental to our understanding or producing of language. However, research demonstrates that gestures differ systematically among typologically distinct languages, in terms of timing and co-expressive meaning alignment with speech (Defina, 2016; Floyd, 2016; Gu, Mol, Hoetjes, & Swerts, 2017; Kita & Özyürek, 2003). They serve functions in language similar to spoken language components. A primary example of cross-linguistic gesture variation is seen in how different languages gesture about motion. For instance, languages like Japanese and Turkish do not usually package information about path and manner of motion within a single linguistic clause, as English does. Accordingly, speakers of these languages tend to represent motion components in separate gestures (Kita & Özyürek, 2003). English speakers, conversely, often do express both components within a single gesture, as the language allows for combined expression at the clause level (Unal, Manhradt, and Özyürek, 2022).

This integration remains consistent even among blind speakers. Blind Turkish and English speakers display speech and gesture patterns similar to their sighted counterparts (Özçalışkan et al., 2016).

Another key area is pointing gestures. These gestures are influenced by language-specific demonstrative systems (Cooperrider, Fenlon, Keane, Brentari, & Goldin-Meadow, 2021) and differ in encoding spatial characteristics and joint attention. Differences also arise in combining pointing and demonstratives. For example, Turkish speakers often use pointing gestures with particular demonstratives based on the listener's attention. Turkish has a three-way demonstrative system. While two of the demonstratives (*bu; o*) encode distance (close and far to a speaker) the third demonstrative (*su*) is used to attract the listener's attention, regardless of the spatial position of the referent. This attention-getter demonstrative *su* is also found to be more frequently accompanied by a pointing gesture than the other two (Peeters & Özyürek, 2016; Küntay & Özyürek, 2006; Azar et al., 2019).

In conclusion, gestures serve roles analogous to the lexical, syntactic, semantic, and pragmatic facets of spoken languages. They help convey messages, in visually iconic or indexical ways, which cannot be easily transmitted through spoken language.

5. Integration of speech and co-speech gestures in comprehension and brain

Speech and gestures are tightly integrated not only in production but also in comprehension. Research indicates that listeners derive semantic information from gestures that accompany speech. Such gestures play a substantial role in enhancing comprehension.

Kelly and Barr (1999) highlighted that participants often incorporated gestural information into their recall of speech. Similarly, Beattie and Shovelton (1999) found that descriptions aided by gestures improved listeners' accuracy about size and positional information. Studies by Kelly, Özyürek, and Maris (2010) further accentuated the significance of congruence between speech and gesture in comprehension. Evidence suggests that gestures are not processed independently; rather, there is semantic integration between representational gestures and speech (see also Trujillo and Holler, 2023, for an extended processing model).

Neurocognitive studies have substantiated the semantic correlation between gestures and speech. For instance, Wu and Coulson (2007) discovered that gestures that are semantically incongruous with preceding visual stimuli influenced the brain's processing, as evidenced by ERP effects. Furthermore, neural evidence shows that gestures can disambiguate upcoming speech, demonstrating the power of gestures in shaping comprehension (Holle & Gunter, 2007; Özyürek et al., 2007).

fMRI studies shed light on brain activations during gesture perception. Brain regions, such as the left inferior frontal gyrus (IFG), medial temporal gyrus (MTG), and superior temporal gyrus/sulcus, which are integral to linguistic processing, are activated during gesture comprehension and are also found to be relevant for integrating information from speech and gesture (Dick, Goldin-Meadow, Solodkin, & Small, 2012). Straube, Green, Weis, and Kircher (2012) found overlap in brain regions activated by meaningful speech and gestures, particularly in the left IFG and bilateral MTG (see also Xu, Gannon, Emmorey, & Braun, 2009).

Recent studies underscore the potency of gestures in refining comprehension, especially in challenging auditory conditions. Viewing iconic gestures alongside unclear speech can clarify

ambiguous speech content, sometimes even more effectively than lip reading. The power of alpha and beta oscillations in the gesture-relevant areas changes as a function of the semantic fit between the two modalities in both clear and noisy speech, underscoring the dynamic interplay between speech and gesture (Drijvers, Özyürek, & Jensen, 2018).

In conclusion, gestures are pivotal in enhancing and shaping speech comprehension. They are not mere accessories to speech but play a cardinal role in how listeners interpret and assimilate spoken content. Neuroscientific evidence further illustrates the link between speech and gesture, highlighting the neural networks involved in multimodal language processing.

6. The integration of speech and gesture: A distributional semantics framework

As we have seen above, observational, behavioral, and neuroscientific research has clearly shown that gestures are an integral part of the multimodal capacity for language. An extension beyond the current version of the PA is needed to accommodate this central feature of the cognitive architecture for language. However, this raises the following important question: What is the common format in which meanings derived from speech and co-speech gestures can be integrated? The auditory and visuospatial input signals have quite different characteristics. Nevertheless, the meaning features extracted from the different input channels have to result in a common semantic representation, including and integrating the semantic contributions of the individual sources of input. Here the distributional semantics framework might offer a solution. In this framework, the basic idea is that the meaning of a word is derived from the company it keeps; that is, from the context in which it is embedded. This results in a semantic space with a very large number of dimensions. The meaning of a word is a vector in this multidimensional vector space (Boleda, 2020). The spatial proximity of word vectors in semantic space defines their similarity in meaning. However, these vectors are usually derived from large text corpora and do not include information from another channel. The question is how to fuse the inputs from different channels. This problem has been solved for the integration of text and visual objects (Baroni, 2016). The latter can be recognized by convolutional networks such as ImageNet and result in visual vectors for recognized objects. A multimodal fusion function takes the textual and visual vectors as input, returning a new set of vectors in which textual and visual information is integrated (Baroni, 2016). This works not only for static objects but also for moving objects. Regneri et al. (2013) showed that multimodal integration of information extracted from videos and action-depicted verb phrases is also possible. Wahlster (2002, 2023) argues for the need to integrate multimodal input on a semantic and pragmatic level (see Fig. 2).

Although multimodal vector fusion between speech and co-speech gestures seems one way to integrate the meanings extracted from both channels into a common semantic representation, practical constraints limit the current possibilities. One major limitation is that training data from a large corpus of co-speech gestures are not yet available. This is a requirement for the automatic semantic feature extraction from the gestures. Work in this direction is needed, but not impossible and already partly ongoing (Pouw, Dingemanse, Motamedi, & Özyürek, 2021).

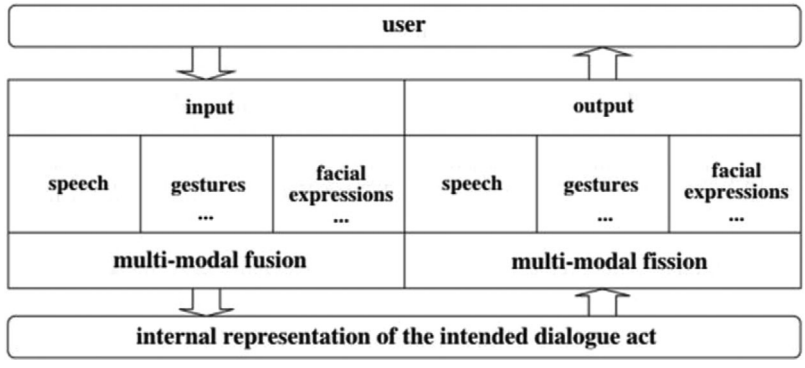


Fig. 2. Multimodal fusion and fission. Based on the multimodal input an internal representation of the speech act can be generated by vector fusion of the vectors from the different input types. In production, the multimodal representation results in fission into the relevant dialogue dimensions (source: Wahlster, 2023).

Given the ambiguous nature of the co-speech gestures, the resulting vector might be semantically underspecified. However, the fusion with the vector for the spoken word input will result in a more specific semantic representation than the spoken input alone can contribute. For instance, a circular movement in space in the presence of the spoken word table will result in the semantic representation "*round table*." In the similarity space for visuospatial movements, the particular gesture will occupy a vector in the area of rounded motions. Fusion of this vector with the vector for the word *table* results in an output vector that is most similar in semantic space to tables with the particular form feature *round*.

Even though in the past such vectors for visual gestures would be hard to imagine, with current kinematic recognition tools such as *Open Pose* it is possible to extract kinematic movement patterns of gestures (e.g., size, distance, speed) from video-based images (Cao et al., 2017). A recent study has shown that it is possible to find kinematic similarities between gestures that are also semantically related (Pouw et al., 2021).

In conclusion, language is inherently multimodal in nature. This key feature is missing in current accounts of the PA (but see Cohn, 2016; Cohn & Schilperoord, 2021). These models, therefore, need to be extended to incorporate the multimodal characteristics of language. However, given the differences in the formats of representation, this raises the issue of how to translate the format of the visuospatial input into that of the spoken input or vice versa. Here we propose that, in principle, distributional semantics might provide a possible answer. Multimodal fusion of the semantic vectors from both input domains results in the joint semantic representation that integrates the information from both speech and gesture.

Acknowledgments

Open access funding enabled and organized by Projekt DEAL.

References

Azar, Z., Backus, A., & Özyürek, A. (2019). General- and language-specific factors influence reference tracking in speech and gesture in discourse. *Discourse Processes*, *56*, 553–574.

Baroni, M. (2016). Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, *10*, 3–13.

Beattie, G., & Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, *18*, 438–462.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, *6*, 213–234.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. 1017 IEEE Conference on computer vision and pattern recognition (CVPR). *Honolulu*, (HI), pp 1302–1310. https://doi.org/10.1109/CVPR.2017.143

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Chu, M. Y., & Hagoort, P. (2014). Synchronization of speech and gesture: Evidence for interaction in action. *Journal of Experimental Psychology-General*, *143*, 1726–1741.

Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

Cohn, N. (2016). A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, *146*, 304–323.

Cohn, N., & Schilperoord, J. (2021). Remarks on multimodality: Grammatical interactions in the parallel architecture. *Frontiers in Artificial Intelligence*, *4*(4), 778060. https://doi.org/10.3389/frai.2021.778060

Cooperrider, K., Fenlon, J., Keane, J., Brentari, D., & Goldin-Meadow, S. (2021). How pointing is integrated into language: Evidence from speakers and signers. *Frontiers in Communication*, *6*, 567774.

Debreslioska, S., Özyürek, A., Gullberg, M., & Perniss, P. (2013). Gestural viewpoint signals and referent accessibility. *Discourse Processes*, *50*, 431–456.

Defina, R. (2016). Do serial verb constructions describe single events? A study of co-speech gestures in avatime. *Language*, *92*, 890–910.

Dick, A. S., Goldin-Meadow, S., Solodkin, A., & Small, S. L. (2012). Gesture in the developing brain. *Developmental Science*, *15*, 165–180.

Drijvers, L., Özyürek, A., & Jensen, O. (2018). Alpha and beta oscillations index semantic congruency between speech and gestures in clear and degraded speech. *Journal of Cognitive Neuroscience*, *30*, 1086–1097.

Ebert, C. (2024). Semantics of gesture. *Annual Review of Linguistics*, *10*, 169–189

Floyd, S. (2016). Modally hybrid grammar? Celestial pointing for time-of-day reference in Nheengatu. *Language*, *92*, 31–64.

Goldin-Meadow, S., Brentari, D., Coppola, M., Horton, L., & Senghas, A. (2015). Watching language grow in the manual modality: Nominals, predicates, and handshapes. *Cognition*, *136*, 381–395.

Gu, Y., Mol, L., Hoetjes, M., & Swerts, M. (2017). Conceptual and lexical effects on gestures: The case of vertical spatial metaphors for time in Chinese. *Language, Cognition and Neuroscience*, *32*, 1048–1063.

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, *9*, 416–423.

Hagoort, P. (2014). Nodes and networks in the neural architecture for language: Broca's region and beyond. *Current Opinion in Neurobiology*, *28*, 136–141.

Hagoort, P. (2017). The core and beyond in the language-ready brain. *Neuroscience and Biobehavioral Reviews*, *81*, 194–204.

Hagoort, P. (2008). The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. *Philosophical Transactions of the Royal Society B*, *363*, 1055–1069.

Hagoort, P., & van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society B*, *362*, 801–811.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*, 88–111.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, *19*, 1175–1192.

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

Iverson, J. M., & Goldin-Meadow, S. (1998). Why people gesture when they speak. *Nature*, *396*, 228–228.

Jackendoff, R. (1999). The representational structures of the language faculty and their interactions. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 37–79). Oxford, England: Oxford University Press.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford, England: Oxford University Press.

Jackendoff, R. (2007). A parallel architecture perspective on language processing. *Brain Research*, *1146*, 2–22.

Jackendoff, R., & Audring, J. (2020). *The texture of the lexicon: Relational morphology and the parallel architecture*. Oxford, England: Oxford University Press.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge, England: Cambridge University Press.

Kelly, S. D., & Barr, D. J. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, *40*, 577–592.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*, 260–267.

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

Kita, S., & Emmorey, K. (2023). Gesture links language and cognition for spoken and signed languages. *Nature Reviews Psychology*, *2*, 407–420

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*, 16–32.

Kuntay, A. C., & Özyürek, A. (2006). Learning to use demonstratives in conversation: What do language specific strategies in Turkish reveal? *Journal of Child Language*, *33*, 917–917.

Mamus, E., Speed, L. J., Özyürek, A., & Majid, A. (2022). The effect of input sensory modality on the multimodal encoding of motion events. *Language Cognition and Neuroscience*, *38*, 711–723.

Marslen-Wilson, W. D. (1989). Access and integration: Projecting sound onto meaning. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 3–24). Cambridge, MA: MIT Press.

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

McNeill, D. (2005). Gesture, gaze, and ground. In S. Renals & S. Bengio (Eds.), *Machine learning for multimodal Interaction*, *3869*, 1–14.

Özçalışkan, S., Lucero, C., & Goldin-Meadow, S. (2016). Is seeing gesture necessary to gesture Like a native speaker? *Psychological Science*, *27*, 737–747.

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*, 605–616.

Özyürek, A., & Woll, B. (2019). Language in the visual modality: Co-speech gesture and sign language. In P. Hagoort (Ed.), *Human language: From genes and brain to behavior* (pp. 67–83). Cambridge, MA: MIT Press.

Perniss, P. (2018). Why we should study multimodal language. *Frontiers in Psychology*, *9*, 1109.

Perniss, P., Özyürek, A., & Morgan, G. (2015). The influence of the visual modality on language structure and conventionalization: Insights from sign language and gesture. *Topics in Cognitive Science*, *7*, 2–11.

Peeters, D., & Özyürek, A. (2016). This and that revisited: A social and multimodal approach to spatial demonstratives. *Frontiers in Psychology*, *7*, 222.

Pouw, W., Dingemanse, M., Motamedi, Y., & Özyürek, A. (2021). A systematic investigation of gesture kinematics in evolving manual languages in the lab. *Cognitive Science*, *45*, e13014.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., & Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, *1*, 25–36.

Rohrer, P. L., Delais-Roussarie, E. & Prieto, P. (2023). Visualizing prosodic structure: Manual gestures as highlighters of prosodic heads and edges in English academic discourses. *Lingua*, *293* 103583.

Schlenker, P. (2019). Gestural Cosuppositions within the transparency theory. *Linguistic Inquiry*, *50*(4), 873–884.

Seuren, P. A. M. (2004). *Chomsky's minimalism*. Oxford, England: Oxford University Press..

Slonimska, A., Özyürek, A., & Capirci, O. (2022). Simultaneity as an emergent property of efficient communication in language: A comparison of silent gesture and sign language. *Cognitive Science*, *46*, e13133.

Straube, B., Green, A., Weis, S., & Kircher, T. (2012). A supramodal neural network for speech and gesture semantics: An fMRI study. *PLoS One*, *7*, e51207.

Trujillo, J. P., & Holler, J. (2023). Interactionally embedded gestalt principles of multimodal human communication. *Perspectives on Psychological Science*, *18*(5), 1136–1159. https://doi.org/10.1177/17456916221141422

Unal, E., Manhardt, F., & Özyürek, A. (2022). Speaking and gesturing guide event perception during message conceptualization: Evidence from eye movements. *Cognition*, *225*, 105127.

Wahlster, W. (2002). SmartKom: Fusion and fission of speech, gestures, and facial expressions. Proceedings of the 1st International Workshop on Man-Machine Symbiotic Systems, Kyoto, Japan (pp. 213–225).

Wahlster, W. (2023). Understanding computational dialogue understanding. *Philosophical Transactions A*, *381*, 20220049.

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, *17*, 2322–2333.

Wu, Y. C., & Coulson, S. (2007). Iconic gestures prime related concepts: An ERP study. *Psychonomic Bulletin & Review*, *14*, 57–63.

Xiang, H. D., Fonteijn, H.M., Norris, D.G., & Hagoort, P. (2010). Topographical functional connectivity pattern in the Perisylvian language networks. *Cerebral Cortex*, *20*, 549–560.

Xu, J., Gannon, P. J., Emmorey, K., & Braun, A. R. (2009). Symbolic gestures and spoken language are processed by a common neural system. *Proceedings of the National Academy of Sciences*, *106*, 20664–20669.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, *32*, 25–64.