Envisioning intention-oriented brain-to-speech decoding

Leon Li and Jared Vasil

Duke University

Serban Negoita

University of Maryland School of Medicine

Key words: brain-to-speech, speech decoding, brain-machine interface, neurolinguistics, neuropragmatics

CORRESPONDING AUTHOR:

Leon Li

Email: leon.li@duke.edu

110 Sociology Psychology Building

Duke University

Durham, NC. 27708

Abstract

The typical approach to decoding speech from the brain (using brain-machine interfaces) is to decode low-level linguistic units (e.g., phonemes, syllables) from motor articulation areas (e.g., premotor cortex) with the aim of assembling these low-level units into higher-level discourse. We propose that brain-to-speech decoding may benefit from adopting a functional view of language, which conceives of language as an instrumental tool for interacting with others' intentions in order to fulfill one's own intentions. This functional view of language motivates adopting usability (i.e., the decoder's usefulness as a tool for achieving goals), in addition to decoding accuracy, as a criterion for assessing decoder performance. Decoders may achieve gains in usability by incorporating data about communicative situations and speaker intentions in order to generate and fill in speech act templates (e.g., when the speaker wishes to make a request, the decoder generates an imperative speech act template and then fills in the contents of the template based on situational and intentional data). We suggest that this intention-oriented, template-based, and functionally inspired view of brain-to-speech decoding may facilitate efforts to achieve naturalistic speech decoding.

Envisioning intention-oriented brain-to-speech decoding

1. Introduction

Words, like nails and hammers, are tools that we use to achieve goals. We use nails and hammers to achieve physical goals like fastening planks of wood together. We use words, constructions, and other linguistic form-meaning pairings to achieve social goals like establishing a meeting of minds. Thus, we begin this paper by emphasizing the functional definition of language as *a tool that we use* (e.g., Tylén, Weed, Wallentin, Roepstorff, & Frith, 2010). By looking beyond the surface structures that we ordinarily equate with "language" (e.g., the tangible letters and words that we see in print and hear at the marketplace), we discover that the heart of language is essentially its functional usage (Tomasello, 2003). Language is, in essence, a means for aligning mental states, a tool for building bridges between our scattered islands of consciousness.

The aim of this paper is to leverage insights from this functional view of language towards the design of brain-machine interfaces (BMIs) for decoding speech from the brain. Our essay is primarily geared towards the conceptualization of BMIs. However, our arguments will have important implications for both the engineering of BMIs and the use of BMIs as a scientific tool for exploring questions about human cognition. In previous work, researchers have used methods such as electrocorticography (ECoG) and functional magnetic resonance imaging (fMRI) to create BMIs capable of decoding units of language from brain activity (Brumberg, Wright, Andreasen, Guenther, & Kennedy, 2011; Moses, Leonard, & Chang, 2018; Pei, Barbour, Leuthardt, & Schalk, 2011). Because ECoG is an invasive procedure whose use must be independently justified by non-research purposes, research using ECoG to decode speech has generally been conducted in two types of clinical contexts: 1) with patients who are in an

"locked-in" state and unable to communicate due to injuries to their motor articulatory processes, and 2) with epilepsy patients (who have normal communication abilities) who are undergoing surgical electrode placement on their cerebral cortex in the course of identifying seizure loci.

These previously developed BMIs, which have been sufficiently accurate to decode simple constrained responses (e.g., phonemes, binary "yes" or "no" responses), have demonstrated benefits for patients who would otherwise have no motor ability to communicate. In promising recent research, BMIs have even enabled typing rates of around 25 characters per minute (Pandarinath et al., 2017). The ultimate aspiration of BMI design, however, is to create a "naturalistic" BMI that can rapidly and intuitively produce speech on par with the fluency of natural language (that is, around 150 words per minute). Thus, by "naturalistic," we mean a kind of BMI that can enable fast, accurate, and fluent expressions of a speaker's intended meaning—a BMI whose learning and inference properties would make it as though one were thinking out loud. We contend that the design of such a BMI will only be possible by taking into consideration functional, top-down aspects of language use in addition to lower-level phonemic data.

This paper is organized as follows. Section 2 relays the theoretical background on the dichotomy central to this paper, namely, the difference between bottom-up and top-down modes of speech decoding. Section 3 considers important aspects of the functional view of language that, we suggest, may be important when considering the feasibility and utility of top-down BMIs. Section 4 clarifies the notion of a *speech act template*, which we repeatedly employ in the ensuing discussion of top-down BMIs. Sections 5 and 6 relay eight types of data that will be instrumental to the design of top-down BMIs. In particular, Section 5 relays four types of data inherent to communicative situations, and Section 6 identifies four types of data inherent to

speakers' communicative goals and intentions. We summarize these eight types of data in Figure 1. It is useful to note at the outset that these eight types of data are not privileged over other possible kinds of data; rather, we simply intended to provide a useful way to structure and elaborate on our central arguments. We acknowledge that additional kinds of relevant data may exist, and other researchers may seek to categorize our eight kinds of data differently. Finally, Section 7 notes potential future directions.

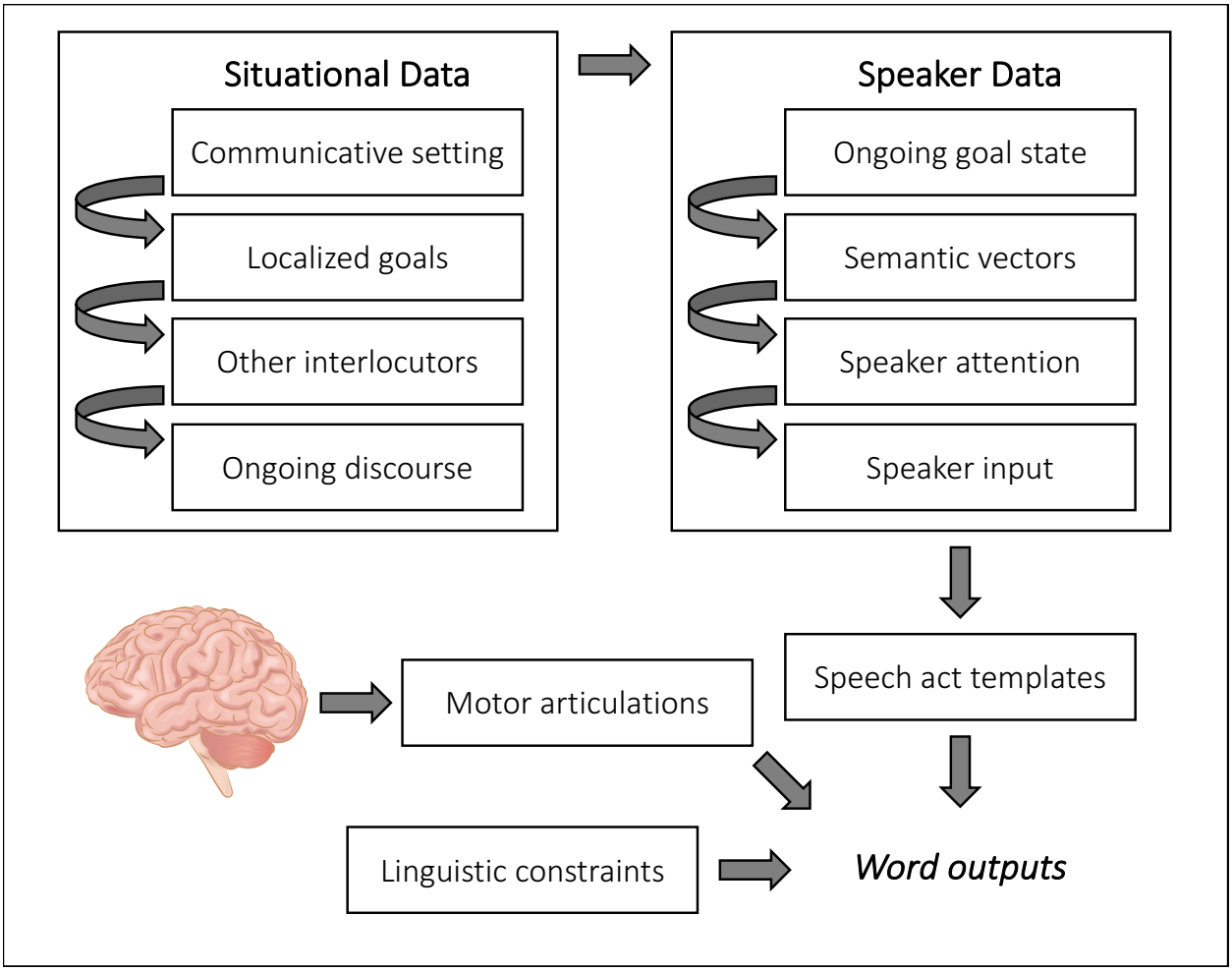

Figure 1. Various types of information that may aid the decoding of speech from the brain.

2. Bottom-Up and Top-Down Decoding

The typical approach in brain-to-speech BMI design is to decode low-level units of language (e.g., phonemes, syllables) from motor articulation areas (e.g., premotor cortex). The intuition guiding this "bottom-up" approach is that low-level units, once decoded, could then be assembled—from the bottom-up—into higher-level units such as words and sentences. The bottom-up approach has achieved promising success. For example, one recent ECoG study attained over 70% accuracy when decoding which phoneme (from a hypothesis space of 4 options) a person was saying (Ramsey et al., 2018). In addition, earlier studies achieved accuracies of around 20% to 30%, which are substantially higher than expected by chance, when decoding which phoneme (from hypothesis spaces of 38 to 39 options) a person was hearing or intending to produce (Brumberg et al., 2011; Moses, Mesgarani, Leonard, & Chang, 2016).

Thus, the bottom-up approach has proven to be highly promising and deserving of further development. Despite the impressive achievements of the bottom-up approach, however, BMIs following the bottom-up approach have not yet attained naturalistic levels of speed, accuracy, and fluency. The overall effort to attain naturalistic speech decoding may stand to benefit from employing a "top-down" approach, as described by Li and Negoita (2018). The guiding intuition of the top-down view is that high-level data about the communicative exchange could be decoded to help constrain the hypothesis space for the subsequent decoding of low-level units. The potential utility of taking a top-down approach is supported by previous theory and research on predictive coding (e.g., Knill & Pouget, 2004) and the hierarchical organization of the cortex (Bassett et al., 2008; Felleman & Van Essen, 1991). Namely, researchers have proposed that the neural dynamics of "higher" layers of cortex exert recurrent constraints on lower layers (e.g., Tajima et al., 2017). This top-down regulation in effect enables the dynamics of higher layers to

function as a set of overhypotheses or Bayesian priors on the dynamics of the lower layers (Friston, 2008). In other words, lower layers characteristically "conform" to the expectations encoded in higher layers.

This top-down approach motivates a formal scheme (see Buckley, Kim, McGregor, & Seth, 2017; Friston, 2008) for understanding how low-level dynamics such as phoneme selection may be entrained by higher-level dynamics such as expectations about phoneme sequences (Kiebel, Daunizeau, & Friston, 2008; Rabinovich, Simmons, & Varona, 2015; Yildiz, von Kriegstein, & Kiebel, 2013). Moreover, this type of approach has proven successful in accounting for an array of cognitive and neuroscientific phenomena in adults (Clark, 2013, 2016; Friston, 2010) as well as children (Perfors, Tenenbaum, Griffiths, & Xu, 2011; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). In the present setting, the implication of this top-down approach for BMI construction is that data on the dynamics of high-level contextualizing layers (e.g., intentions) may effectively function to minimize the hypothesis space over the flow of low-level (e.g., phonemic) dynamics, thereby rendering that flow predictable for the purposes of speech decoding.

In addition to neural data, other helpful types of top-down information, as described by Li and Negoita (2018), may include the communicative intentions of the speaker, the communicative pragmatics of the situation, the lexical and syntactic constraints of the speaker's language, and the speaker's prior speech history. Here, we build on Li and Negoita's (2018) top-down proposal by describing how a functional view of language—with a special focus on communicative intentions—may inform BMI design.

3. The Functional View of Language

Intention is the central concept in the functional view of language. After all, the ultimate function of language is to help us interact with others' intentions (Tomasello, 2003). We use language to achieve social goals such as expressing our own intentions, discovering others' intentions, influencing others' intentions, and aligning shared intentions during joint activities (Fusaroli, Gangopadhyay, & Tylén, 2014; Tomasello, 2008). These uses of language have been catalogued by influential theories of *speech acts* (e.g., Searle, 2001), which conceive of utterances as instrumental actions for achieving social goals. The social goal of discovering another person's intentions, for instance, may be achieved by uttering an *interrogative* speech act (e.g., asking a question). The social goal of influencing another person's intentions may be achieved by uttering an *imperative* speech act (e.g., issuing an order).

According to this perspective, a communicative exchange may be considered to be "successful" when the right kind of intentional interaction has been achieved (Rakoczy & Tomasello, 2009). For example, communicative success may be said to be attained when a speaker succeeds in expressing their intentions with a *declarative* speech act, or when two speakers succeed in aligning their intentions for a joint plan of action. However, this intention-oriented definition of success is not the criterion that has typically been used when assessing decoder performance. Instead, the typical standard of decoder performance has been accuracy, that is, the decoder's likelihood of selecting the correct unit (e.g., the correct phoneme) from a pool of possible choices (e.g., of multiple phonemes; see Herff et al., 2015; Kellis et al., 2010; Moses et al., 2016; Mugler, et al., 2014; Wehbe et al., 2014). Accordingly, efforts to improve BMI design have focused on improving performance as it relates to accuracy.

We propose that adopting what we term *usability* as a measure of decoder performance

may helpfully inform BMI design. By usability, we mean the BMI's functional usefulness as a means for achieving communicative goals. From the perspective of usability, we may ask questions such as: Could I use the BMI to ask a question (i.e., to issue an interrogative speech act)? Could I use the BMI to comment on a topic (i.e., to issue a declarative speech act)? Could I use the BMI to make a request (i.e., to issue an imperative speech act)?

We suggest that one way to make BMIs more functionally useful is to design BMIs to be able to decode speech acts. Within this paper, we characterize individual speech acts in terms of their particular illocutionary force (i.e., as manifestations of well-defined categories, such as imperative, declarative, etc.) and their particular semantic content (closely, the Austinian "locutionary force"). Thus, a statement such as "Hand me the screwdriver, please" is characterized by its particular illocutionary force (as an imperative speech act) as well as its particular semantic content (e.g., concepts pertaining to [SCREWDRIVER], such as the actions that one may undertake with such devices; see Langacker, 1986). The reason for making this dual characterization owes to its mapping nicely onto existing work in the emerging field of *neuropragmatics*, which is a subfield of the broader area of study known as *neurolinguistics* (see Pulvermüller, 2015, for a recent review). Research within the theoretical framework of neuropragmatics has attempted to associate various pragmatic and semantic properties of language with possible neural instantiations. For instance, the understanding of communicative intent (i.e., the intention to modulate, via communication, another's intentional state; Tomasello, 2003) is a central component of pragmatic inference (Sperber & Wilson, 1986). Neuropragmatic work suggests that communicative intent may be associated with unique neural signatures (Enrici, Adenzato, Cappa, Bara, & Tettamanti, 2011).

Relatedly, the possibility of detecting the illocutionary force of speech acts from brain

dynamics was suggested by a study by Egorova, Shtyrov, and Pulvermüller (2016). It was found that different patterns of neural activity were associated with observations of naming (i.e., declarative) versus requesting (i.e., imperative) speech acts. Whereas the naming speech acts generated more activity in the left angular gyrus, the requesting speech acts led to more activity in the left inferior frontal gyrus, bilateral premotor cortex, right posterior superior temporal sulcus, and left anterior parietal cortex (Egorova et al., 2016). In addition, another fMRI study examined not only the observation but also the production of communicative gestures such as imperative or declarative points; this study found that pointing production was associated with regions such as bilateral ventral premotor cortex, anterior midcingulate cortex, right presupplementary motor area, right temporoparietal junction, and middle insula (Committeri et al., 2015). Moreover, the possibility of detecting the particular semantic content of a speech act from neural dynamics has been suggested as well, as reviewed by Pulvermüller (2013). In his comprehensive article, Pulvermüller (2013) reviews substantial evidence for the unique neural circuits (and combinations of circuits) subserving both concrete and abstract semantic knowledge. For instance, concrete action words and phrases (e.g., “lick” and “kick”) tend to be associated with increased activation of inferior frontal cortex and frontocentral regions (Bak, O’Donovan, Xuereb, Boniface, & Hodges, 2001). This is in contrast to abstract words and phrases (e.g., “freedom” or “beauty”), which tend to be associated with increased activation of anterior cingulate cortex and limbic structures (Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011; see Pulvermüller, 2013).

Taken collectively, these results suggest that non-phonetic data may aid existing BMIs in decoding language from brain dynamics (see, e.g., Cooney, Folli, & Coyle, 2018; Noordzij et al., 2009). This motivates the question of how various sources of pragmatic data could be measured,

organized, and leveraged for the ultimate goal of decoding communicative intentions. In other words, what might an intention-oriented BMI look like?

4. Speech Act Templates

We envision that the intention-oriented BMI would produce an output in the form of a speech act "template" with open "slots" for the particular contents of a given speech act. Suppose that a BMI user wishes to ask someone for a glass of water. Through its decoding of the speaker's neural activity, the BMI could potentially predict that the speaker has a communicative need that they wish to fulfill using an imperative speech act. On the basis of this prediction alone, without any further specifications of *what* in particular the speaker wishes to request, the BMI may already be able to generate a "speech act template" of the intended output, which may take the form of: [decoded speech act: imperative] + [direct object of the speech act: to-be-decoded].

This template-based decoding scheme may turn out to be more flexible—as well as more useful—than a decoding scheme that is solely accuracy-based. In our example, the speaker's goal is to obtain water. The speaker's means for achieving this goal is to influence someone else's intentions using an imperative speech act. From the speaker's perspective, it may not matter much which particular words the BMI selects to fill in the speech act template, so long as the overall point is conveyed. Indeed, any request-related word (e.g., "get," "give," or even "want") for the first "speech act slot"—in combination with any relevant content word (e.g., "water," "glass," or "drink") for the second "direct object slot"—may suffice to convey the speaker's intention. For instance, for the speaker who wishes to satiate their thirst, whether the low-level components of the BMI read (specifically) "get water," "give glass," or "want drink"

may be largely unimportant. What matters is that the BMI can effectively convey the intended communicative force of the speech act along with the intended proposition or state of affairs.

One may imagine how this template-based decoding scheme would be more useful and flexible than the traditional bottom-up decoding scheme that emphasizes accuracy. An accuracy-oriented decoder that aims solely to decode low-level units would have to arrive at precisely the right combination of phonemes and syllables—in the right order, as well—for expressing only one particular phrasing of the request. The intention-oriented BMI, in contrast, has more flexibility regarding which content words it selects. This emphasis on decoding meaning at the level of constructions and speech acts (as opposed to the level of individual words) is based on the psycholinguistic recognition that words often have multiple or ambiguous meanings (Li & Slevc, 2017), such that the compositional meanings of individual words may not be as important as the overall meanings of larger constructions (Goldberg, 2003; Tomasello, 2003). Of course, we are not suggesting that the bottom-up approach is not important to pursue. Indeed, we believe that the effort to attain naturalistic speech decoding will require considerable advances in both bottom-up decoding accuracy and top-down usability. Given that researchers have already made substantial advances using the bottom-up approach, further suggestions for the top-down, intention-oriented approach may now be warranted.

In the following sections, we turn to various types of data that may help BMIs generate and fill in speech act templates. Four types of situational data that may be helpful include:

1. The communicative situation itself.
2. The speaker's goals and previous history of speech localized to particular situations.
3. The speaker's relationships with the other people in the situation.
4. The contents of the ongoing discourse.

Additionally, four types of data about communicative intentions that may be helpful include:

1. The neural signatures of speakers' goal states.
2. The neural signatures of semantic vectors.
3. The speaker's attention.
4. The speaker's volitional input.

We acknowledge that a considerable complication in considering so many sources of data is the challenge of integrating the data coherently. In what order should the different types of data be collected? What weights (i.e., priorities) should be assigned to the different types of data? For example, should speaker intention data be prioritized over situational data in the BMI's prediction scheme, or vice versa? Even within the category of situational data alone, which type of situational data (e.g., the speaker's relationships or the contents of the discourse) should be given priority?

In light of this considerable ambiguity, our preliminary suggestion is for the BMI to decode data in a temporally sequential fashion that is ordered by specificity (namely, from most broad to most specific). In other words, we suggest that BMIs first decode broad types of data. Then, once the broad types of data have been decoded, the BMI could go on to consider more specific types of data. Another way to frame our eight-step decoding scheme is that we first decode four types of data (listed above) that are more "external" to the BMI user. These four types of data are "external" to the speaker in the sense that they may be valuable to the BMI's predictions independent of whether the speaker herself is aware of them (e.g., it may be helpful for the BMI to "know" that the speaker tends to talk about certain topics at certain locations even if the speaker herself is not aware of her own tendency). The latter four types of data may be said to be more "internal" to the speaker's own awareness and thereby more subject to her volition

and control. We acknowledge that the temporal order in which the data are decoded may not be equivalent to the order of priority of the data's importance; after all, some types of data that are decoded later in the process may actually be more informative than other types of data that are decoded earlier. Undoubtedly, more theory and research will be needed to clarify the process of combining different data sources.

5. Situational Data

An important type of high-level data to consider is the speaker's current situation (see Figure 1). As proposed by Li and Negoita (2018), brain-to-speech decoders may benefit from exploiting the informational richness of the communicative setting itself. Communicative exchanges are often constrained by enriched cultural scripts and expectations that guide how people act within recurrent communicative situations (Fusaroli et al., 2014; Tomasello, 2003). In some cases, the situational data alone may be sufficient for predicting what speakers are saying. For example, in studies that employed the human simulation paradigm (HSP), adults who watched muted clips of parent-child interactions were often able to guess what word the parent had said (Cartmill et al., 2013). Below, we identify four different types of situational data (although they may certainly overlap).

5.1. The Communicative Setting

First, the BMI could track the communicative situation itself. Given that information about communicative situations is found *outside* the brain, how can brain-to-speech BMIs extract and leverage this information? One possibility is for the decoder to track the speaker's location using GPS. Once the location has been identified, the decoder may go on to consider the

speaker's past history of speech at that location. This information about speaker history (localized to the particular location) would be highly useful, given that speakers often repeat what they have said at particular places (Tomasello, 2003). For example, the decoder may learn that a speaker tends to order one kind of drink at one restaurant and another kind of drink at another restaurant. Information about the time of day may also be helpful (e.g., speakers may tend to order different things for lunch and dinner).

5.2. The Speaker's Goals at Particular Settings

Data about speakers' locations and speakers' previous speech at various locations provide a general hypothesis space for which content words are likely to arise in the speech stream (e.g., words related to drinks are common at restaurants). However, as emphasized by the functional view of language, words are simply the means by which we accomplish goals; the primary emphasis is on the decoding of goals, not the decoding of linguistic units. Indeed, speaker goals are often localized to particular settings. That is, the cultural common ground shared by interlocutors within culturally recognized settings constrains the kinds of goals that speakers and listeners may have in such settings (e.g., interactions with culturally recognized sets of affordances; van Dijk & Rietveld, 2017). For example, we tend to focus on food-related goals at restaurants, sports-related goals at stadiums, and money-related goals at the bank. Thus, a second suggestion for increasing the usability of BMIs is to include data concerning not only the content words that are recurrent at a location but also the speaker's recurrent goals at that location. This suggestion aligns with classical psychological theories on scripts (e.g., Schank & Abelson, 1977), which broadly propose that human represent their lives in terms of recurrent event sequences.

5.3. The Speaker's Relationships with Interlocutors

A third suggestion is for decoders to identify not only the communicative setting itself but also the other people present. If the decoder is able to identify (and remember) particular interlocutors, then it could make predictions on the basis of the speaker's previous histories of interactions with those interlocutors. To build on the previous suggestions, the decoder could represent the content words and goals that are associated with particular interlocutors (e.g., we tend to say imperative speech acts to waiters at restaurants). In other words, the decoder could represent aspects of the speaker's *relationships* with interlocutors. This suggestion is motivated by the emerging shift in social cognitive neuroscience towards investigating the neural responses of actual social interactions, not just the neural responses associated with observing social interactions from a third-party stance (Gallotti & Frith, 2013; Pfeiffer, Timmermans, Vogeley, Frith, & Schilbach, 2013; Schilbach, 2010). Relatedly, an interesting proposal by Bara, Enrici, and Adenzato (2016) is to monitor the brains of both of the participants in a conversation simultaneously (but much more research will be needed to discover how to analyze and leverage this kind of data).

5.4. The Contents of the Ongoing Discourse

The previous suggestions pertain to high-level types of situational data. At the proximal level of ongoing conversation, decoders may also track the trajectory of a conversation as it unfolds. What has been said so far? What questions have been raised? What information has been exchanged up to now? Decoders may benefit from leveraging the contents of the ongoing discourse to predict upcoming utterances (Herff et al., 2015). Utterances are often contextually

situated (even in cases where a person is talking only to themselves). This suggestion builds on the proposal by Li and Negoita (2018) for BMIs to consider lexical collocations, i.e., regularities in how semantically and topically related words co-occur in discourse (Heylen, Wielfaert, Speelman, & Geeraerts, 2015). Beyond just lexical collocations, it would also be helpful for more abstract types of textual analysis to be paired with BMIs. For instance, latent semantic analysis could be employed to track concepts and meanings, not just lexical word forms. As well, syntactic analysis could help predict useful types of information such as recursions or verb arguments (e.g., if the decoder encounters a transitive verb, it may expect that the subsequent word will be a noun). In sum, various types of data about the communicative situation could be decoded and integrated in a temporally coherent way, from most broad to most specific.

6. Speaker Goals and Intentions

The decoding of situational data, as described above, may help constrain the hypothesis space for the BMI's subsequent decoding of actual, proximal utterances. When decoding utterances, it will be important to decode not only the utterances themselves (via the bottom-up approach) but also the intentions underlying the utterances (see Figure 1). To reiterate, the functional view of language conceives of utterances as instrumental tools for interacting with others' intentions in order to fulfill one's own intentions. Thus, it will be important for the BMI to directly decode the neural signatures of speakers' goals and intentions. Below, we describe four different types of speaker intention data, which, again, may certainly overlap.

6.1. The Speaker's Goal State

With respect to decoding goals, the neural signatures of tangible physiological needs

(e.g., the neural signature for the thirst response) may be relatively straightforward to localize (e.g., in the hypothalamus). Moreover, the BMI could potentially also detect the neural signatures of less physiologically oriented goals, such as socially oriented desires. For decoding socially oriented goals, the BMI may benefit from tracking the neural activity in regions such as medial prefrontal cortex and temporoparietal junction. These regions have been considered crucial parts of an Intention Processing Network, which is involved in humans' considerations of others' mental states (Bara et al., 2016; Tettamanti et al., 2017). Given that the Intention Processing Network is involved when people consider others' intentions, the activation of regions within this network may be associated with socially oriented goals, such as the desire to affiliate with others. This remains only a preliminary speculation, and more research will be needed to explore this possibility. Broadly, it will be important for future BMI design to find ways for BMIs to decode more abstract types of goals (e.g., socially affiliative goals), not just concrete goals.

6.2. Semantic Vectors

Secondly, once the speaker's goal state has been decoded, the BMI may then attempt to decode the particular contents of the goal state. One step in this direction would be to decode the topic that the speaker has in mind (e.g., once the BMI has decoded that the speaker is thirsty, the BMI may try to decode what kind of drink in particular the speaker would like). To that end, the BMI may benefit from decoding the neural signatures of *semantic vectors*. In one recent fMRI study, a large semantic space consisting of 29,805 words was analyzed to generate 180 interpretable clusters of words, with each cluster representing a region of semantic space consisting of words that tend to appear in similar contexts (Pereira et al., 2018). Impressively, the

decoding system could detect correspondences between these 180 semantic clusters and the neural activity patterns of participants as they read text inside the fMRI scanner (Pereira et al., 2018). As such, the decoding of semantic vectors may help the BMI predict what topic a person has in mind.

6.3. The Speaker's Attention

Semantic vectors can only give the decoder a general impression of what the speaker has in mind. That is, semantic vectors may not be precise enough to specify which particular object the speaker desires (e.g., even if the decoder has detected that the speaker wishes to talk about soft drinks, it may still be unclear which particular soft drink the speaker would like). A third suggestion, then, is to decode the speaker's *attention*. Once the decoder has decoded the speaker's goal as well as an impression of the contents of the goal, further predictions about the goal's contents may be refined by assessing the speaker's attention. One avenue for assessing attention is for the BMI to be paired with cameras for eye-tracking and object-detection (Li & Negoita, 2018). Gaze is a robust predictor of both attention and intention because speakers often talk about what they are directly attending to—and attend to what they are directly gazing at (Tomasello, 2003).

To return to our example of the thirsty speaker, suppose that the BMI has detected that the speaker wishes to request something (and, accordingly, has prepared the imperative speech act template) and is now in the process of predicting what in particular the speaker will request. The BMI's prediction that it is water in particular that the speaker wishes to request could receive convergent support from both the speaker's neural activity associated with water (e.g., the articulatory signatures of the word "water" from the premotor cortex, as well as the neural

signatures of the semantic vector associated with beverages) and the speaker's gaze towards a water source. In a previous example of the integration of BMIs with eye-tracking, Zander, Gaertner, Kothe, and Vilimek (2011) found that users could use a BMI to signify their "selection" of a target stimulus in their visual field, which supports the idea of using BMIs in conjunction with eye-tracking to identify the referent of a user's attention. Aside from assessing gaze data, other speculative possibilities for decoding a speaker's attention are to decode the contents of the speaker's episodic memory, given that episodic memory may include spatiotemporal object representations that could be relevant to what the speaker has in mind (Negoita, Boone, & Anderson, 2016), or to decode the contents of the speaker's "inner voice," that is, the private speech stream within one's mind (Steels, 2003).

6.4. The Speaker's Volitional Input

Lastly, perhaps the most direct way for a BMI to detect a speaker's goals and intentions is for the speaker to simply inform the BMI. This method is advantageous for several reasons. First, the speaker's hypothesis space may already be highly constrained by the time the decoder presents the speaker with options from which to choose. For example, if the decoder has already detected that the speaker is at a certain restaurant—where the speaker regularly orders a certain drink—from a certain waiter, then all the decoder may have to do is ask the user whether they would like to order that drink again, and all the speaker may have to do is manually indicate a "yes" or "no" response. As research has shown, binary responses may be decoded with high accuracy. For example, Naci, Cusack, Jia, and Owen (2013) achieved 90% accuracy when using fMRI to decode whether participants were selectively attending to "yes" or "no" in a stream of sounds. In the event that speakers respond with a "no" to the BMI's suggestion, the BMI could

then scroll through presenting other options for the speaker to confirm or deny.

Manual selection is also advantageous because it does not require the BMI to draw inferences from underdetermined and highly complex neural data. Instead, the speaker may simply spell out—perhaps by using manual methods such as cursor selection on a keyboard interface—what their goal will be for an upcoming communicative exchange. Note that even "manual" selection can be achieved by the speaker's manipulation of their brain activity for decoding by the BMI (e.g., Sellers, Ryan, & Hauser, 2014), as opposed to using one's hands to point and click.

Philosophically, our suggestion to assess the speaker's volitional inputs is motivated by the recognition that speech is grounded in a "first-person" ontology (Searle, 2001). That is, speech is an expression of a speaker's volitional intentions from the speaker's own first-person perspective. The mere fact that the neural signatures of certain words are active at a given time—and observable from a third-party stance by the decoder—does not necessarily entail that the speaker actually has the first-person intention to speak about those contents at that time (Carota et al., 2010). This is to say that the neural signatures of words should not be equated with the actual intentions to speak those words. After all, we often entertain thoughts in our heads that we would rather not say out loud.

Hypothetically, speakers could potentially even use manual inputs from start to finish, so to speak, without relying on the help of predictions based on other kinds of data. To return to the example of the speaker who wants water, the speaker could simply tell the BMI that they wish to make a request, at which point the BMI could prepare the imperative speech act template. Next, after confirming that the BMI has selected the correct template for imperative speech acts, as opposed to other kinds of speech acts, the speaker could then manually specify the content of the

request. Thus, the speaker could fill in the template's open slot for a direct object. Once the BMI's specification of the direct object has been confirmed by the speaker to be correct, the speaker could then permit the BMI to produce the acoustic output (e.g., "get water").

This stepwise procedure seems to be more flexible and potentially even more accurate than attempting to decode the speaker's entire speech stream at once from the bottom-up. Additionally, if the BMI could covertly communicate with its user—such as by confirming with the user, through a private earpiece, that a to-be-produced output is as intended—then users could potentially also inform the BMI about their communicative goals even during the course of a live conversation. The drawback to such a procedure is that it may be slow and cumbersome. However, one may envision that increased practice and familiarity with the stepwise procedure could make it more intuitive for speakers over time. At the very least, this stepwise procedure could facilitate the design of experimental paradigms for improving and evaluating BMIs.

7. Future Directions

An important direction for future research will be to discover how to optimize the integration of different types of data for predicting utterances. The suggestion that we initially raised is for decoders to consider the different types of data in a temporally sequential way—by order of specificity from most broad to most specific (Figure 1). We acknowledge that this temporally sequential strategy has limitations, one being that we are not certain, *a priori*, about how to assign different weights to different data sources.

A second possible solution for integrating the various types of data is to train deep learning networks (e.g., Hinton, 2007; Lecun, Bengio, & Hinton, 2015) to consider the assorted types of broad and specific data (see Sejnowski, 2018). These networks statistically recapitulate

the structures of the processes generating their training data so as to generate novel data from the same process. Interestingly, this possibility links up with predictive coding schemes (described in Section 2) that view neural dynamics as a hierarchically organized, predictive ensemble of component parts (Badcock, Friston, & Ramstead, in press). For instance, by training hierarchical networks on a speaker's actual usages of specific communicative constructions (e.g., Bannard, Lieven, & Tomasello, 2009), the BMI could potentially "learn" to produce the kinds of constructions that the speaker tends to employ. In response to the BMI's suggested constructions, speakers could subsequently simply respond "yes" or "no" as to whether the outputs suggested by the BMI are successful expressions of the speaker's intentions; this user-generated feedback could, in turn, inform the BMI's subsequent predictions. Thus, speakers' inputs could help BMIs undergo a kind of semi-supervised learning. Over time, as more and more feedback is gathered from speakers, deep learning algorithms could potentially learn how to best integrate the various types of information in a data-driven way (Sejnowski, 2018), thereby obviating the need for researchers to provide decoding systems with *a priori* weighting schemes.

This flexible approach may prove to be effective by virtue of its correspondence with the fluid and noisy nature of the mind itself. As proposed in a review by Clark (2006), the mind may not consist of one "Central Meaner" (p. 370) who straightforwardly translates inner cognitive processes into public expressions of language in a temporally stepwise fashion. Rather, language itself may be one of many cognitive influences on a relatively disorganized *coalition* of cognitive processes that influence one another in parallel (Clark, 2006). Language, in this view, may help stabilize the fluid workings of the cognitive network, allowing us to follow deep trajectories through representational space (Clark, 2006). If Clark's (2006) proposal is correct, and the relationship between thought and language is fluid and flexible rather than stable and

hierarchical, then it stands to reason that BMIs may benefit from employing a fluid decoding scheme—one that matches the fluid nature of what is being decoded. Such a scheme might, for instance, allow the BMI to determine the appropriate complexity of the speech act templates required to adequately capture and track the relationships between surface level productions and their underlying causes in mental states.

Importantly, advances in brain-to-speech BMI design will have immediate clinical applications in restoring communication for patients who are unable to produce speech or movement due to damage to their motor processes. These patients may include those with a complete loss of communication, such as those with locked-in syndrome, as well as a broader population of patients with debilitating dysarthrias or aphasias that remain resistant to speech therapy. Beyond the clinical realm, the development of intention-oriented BMIs may also open up novel avenues for the field of human-computer interaction at large. One may envision that software engineering and code writing may be greatly facilitated if computers could predict their users' intentions. In fact, all kinds of technologies (e.g., text-to-speech decoders) could potentially better achieve their intended uses by incorporating data about humans' goals, cognitive states, and values. One intriguing potential application of the decoding of intentions pertains to the emerging commercial industry of "smart" home technologies (e.g., the "Internet of things"). These "smart" technologies include home devices (e.g., lights, kitchen appliances, and even domestic robots) that are responsive to user inputs via digital interfaces. Designing these technologies to be amenable to users' intentions, not just to users' manual inputs, could greatly increase the standard of living for people with motor disabilities who wish to live independently at home. Thus, our intention-oriented view of neural decoding may be informative not just to intention-oriented speech decoding but to intention-oriented user technologies in

general.

8. Conclusion

Ordinarily, when we think about language, we think about the letters, words, and sentences that so tangibly pierce the silences of our inner and outer lives and so beautifully embellish the pages and banners of our civilization. We note, however, that words are only one among many possible means of communication; humans have, in fact, created many other means of conveying meaning, including hand gestures, signal flags, Braille markings, and blips of Morse code. It may not be long until the electric ripples of our brains, as assessed by brain-to-speech decoders, will be counted among our most commonplace methods of communicating. For this to be achieved, we argue, brain-to-speech BMIs should be designed with the insights of a functional view of language in mind. The contributions of theoretical linguistic insights—in combination with further improvements in bottom-up decoding accuracy as well as further advances in neurolinguistics, neuropragmatics, and the overall neuroscience of language (e.g., Catani & Bambini, 2014; Cooney et al., 2018; Tettamanti et al., 2017)—may, in time, suffice to bring the secret musings of our brains to life—to sound—to speech—to significance.

Acknowledgements

We are very thankful to Erin Campbell, Joshua Perlin, Ryan Simmons, and Wouter Wolf for helpful discussions and commentary on the manuscript and the ideas presented therein. We are very grateful to Michael Tomasello, L. Robert Slevc, and Susan Courtney for their formative research training on topics related to language and neuroscience.

References

Badcock, P. B., Friston, K. J., & Ramstead, M. J. D. (in press). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews*.

Bak, T. H., O'Donovan, D. G., Xuereb, J. H., Boniface, S., & Hodges, J. R. (2001). Selective impairment of verb processing associated with pathological changes in Brodmann areas 44 and 45 in the motor neurone disease–dementia–aphasia syndrome. *Brain*, *124*, 103–120.

Bara, B. G., Enrici, I., & Adenzato, M. (2016). At the core of pragmatics: The neural substrates of communicative intentions. In G. Hickok & S. L. Small (Eds.), *Neurobiology of language* (pp. 675–685). London, UK: Academic Press.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*, 17284–17289.

Bassett, D. S., Bullmore, E., Verchinski, B. A., Mattay, V. S., Weinberger, D. R., & Meyer-Lindenberg, A. (2008). Hierarchical organization of human cortical networks in health and schizophrenia. *Journal of Neuroscience*, *28*, 9239–9248.

Brumberg, J. S., Wright, E. J., Andreasen, D. S., Guenther, F. H., & Kennedy, P. R. (2011). Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. *Frontiers in Neuroscience*, *5*, 65.

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55–79.

Carota, F., Posada, A., Harquel, S., Delpuech, C., Bertrand, O., & Sirigu, A. (2010). Neural dynamics of the intention to speak. *Cerebral Cortex*, *20*, 1891–1897.

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

Catani, M., & Bambini, V. (2014). A model for Social Communication And Language Evolution and Development (SCALED). *Current Opinion in Neurobiology*, *28*, 165–171.

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, *10*, 370–374.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–253.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York, NY: Oxford University Press.

Committeri, G., Cirillo, S., Costantini, M., Galati, G., Romani, G. L., & Aureli, T. (2015). Brain activity modulation during the production of imperative and declarative pointing. *NeuroImage*, *109*, 449–457.

Cooney, C., Folli, R., & Coyle, D. (2018). Neurolinguistics research advancing development of a direct-speech brain-computer interface. *iScience*, *8*, 103–125.

Egorova, N., Shtyrov, Y., & Pulvermüller, F. (2016). Brain basis of communicative actions in language. *NeuroImage*, *125*, 857–867.

Enrici, I., Adenzato, M., Cappa, S., Bara, B. G., & Tettamanti, M. (2011). Intention processing in communication: A common brain network for language and gestures. *Journal of Cognitive Neuroscience*, *23*, 2415–2431.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, *4*, e1000211.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.

Fusaroli, R., Gangopadhyay, N., & Tylén, K. (2014). The dialogically extended mind: Language as skilful intersubjective engagement. *Cognitive Systems Research*, *29–30*, 31–39.

Gallotti, M., & Frith, C. D. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences*, *17*, 160–165.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*, 219–224.

Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: Decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience*, *9*, 217.

Heylen, K., Wielfaert, T., Speelman, D., & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, *157*, 153–172.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*, 428–434.

Kellis, S., Miller, K., Thomson, K., Brown, R., House, P., & Greger, B. (2010). Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of Neural Engineering*, *7*, 056007.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology*, *4*, e1000209.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*, 712–719.

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*, 14–34.

Langacker, R. W. (1986). An introduction to cognitive grammar. *Cognitive Science*, *10*, 1–40.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Li, L., & Negoita, S. (2018). Brain-to-speech decoding will require linguistic and pragmatic data. *Journal of Neural Engineering*, *15*, 063001.

Li, L., & Slevc, L. R. (2017). Of papers and pens: Polysemes and homophones in lexical (mis)selection. *Cognitive Science*, *41*, 1532–1548.

Moses, D. A., Leonard, M. K., & Chang, E. F. (2018). Real-time classification of auditory sentences using evoked cortical activity in humans. *Journal of Neural Engineering*, *15*, 036005.

Moses, D. A., Mesgarani, N., Leonard, M. K., & Chang, E. F. (2016). Neural speech recognition: Continuous phoneme decoding using spatiotemporal representations of human cortical activity. *Journal of Neural Engineering*, *13*, 056004.

Mugler, E. M., Patton, J. L., Flint, R. D., Wright, Z. A., Schuele, S. U., Rosenow, J., . . . Slutzky, M. W. (2014). Direct classification of all American English phonemes using signals from functional speech motor cortex. *Journal of Neural Engineering*, *11*, 035015.

Naci, L., Cusack, R., Jia, V. Z., & Owen, A. M. (2013). The brain's silent messenger: Using selective attention to decode human thought for brain-based communication. *Journal of Neuroscience*, *33*, 9385–9393.

Negoita, S., Boone, C., & Anderson, W. S. (2016). Directionality of medial prefrontal cortex and hippocampal interactions is task-dependent. *Neurosurgery*, *79*, N22–N24.

Noordzij, M. L., Newman-Norlund, S. E., de Ruiter, J. P., Hagoort, P., Levinson, S. C., & Toni, I. (2009). Brain mechanisms underlying human communication. *Frontiers in Human Neuroscience*, *3*, 14.

Pandarinath, C., Nuyujukian, P., Blabe, C. H., Sorice, B. L., Saab, J., Willett, F. R., . . . Henderson, J. M. (2017). High performance communication by people with paralysis using an intracortical brain-computer interface. *eLife*, *6*, e18554.

Pei, X., Barbour, D. L., Leuthardt, E. C., & Schalk, G. (2011). Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of Neural Engineering*, *8*, 046028.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., . . . Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*, 963.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*, 302–321.

Pfeiffer, U. J., Timmermans, B., Vogeley, K., Frith, C. D., & Schilbach, L. (2013). Towards a neuroscience of social interaction. *Frontiers in Human Neuroscience*, *7*, 22.

Pulvermüller, F. (2013). How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, *17*, 458–470.

Pulvermüller, F. (2015). Language, action, interaction: Neuropragmatic perspectives on symbols, meaning, and context-dependent function. In A. K. Engel, K. J. Friston, & D. Kragic (Eds.), *The pragmatic turn: Toward action-oriented views in cognitive science* (pp. 139–157). Cambridge, MA: The MIT Press.

Rabinovich, M. I., Simmons, A. N., & Varona, P. (2015). Dynamical bridge between brain and mind. *Trends in Cognitive Sciences*, *19*, 453–461.

Rakoczy, H., & Tomasello, M. (2009). Done wrong or said wrong? Young children understand the normative directions of fit of different speech acts. *Cognition*, *113*, 205–212.

Ramsey, N. F., Salari, E., Aarnoutse, E. J., Vansteensel, M. J., Bleichner, M. G., & Freudenburg, Z. V. (2018). Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids. *NeuroImage*, *180*, 301–311.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schilbach, L. (2010). A second-person approach to other minds. *Nature Reviews Neuroscience*, *11*, 449.

Searle, J. R. (2001). *Rationality in action*. Cambridge, MA: The MIT Press.

Sellers, E. W., Ryan, D. B., & Hauser, C. K. (2014). Noninvasive brain-computer interface enables communication after brainstem stroke. *Science Translational Medicine*, *6*, 257re7.

Sejnowski, T. J. (2018). *The deep learning revolution*. Cambridge, MA: The MIT Press.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.

Steels, L. (2003). Language re-entrance and the 'inner voice'. *Journal of Consciousness Studies*, *10*, 173–185.

Tajima, S., Koida, K., Tajima, C. I., Suzuki, H., Aihara, K., & Komatsu, H. (2017). Task dependent recurrent dynamics in visual cortex. *eLife*, *6*, e26868.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Tettamanti, M., Vaghi, M. M., Bara, B. G., Cappa, S. F., Enrici, I., & Adenzato, M. (2017). Effective connectivity gateways to the Theory of Mind network in processing communicative intention. *NeuroImage*, *155*, 169–176.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2008). *Origins of human communication*. Cambridge, MA: The MIT Press.

Tylén, K., Weed, E., Wallentin, M., Roepstorff, A., & Frith, C. D. (2010). Language as a tool for interacting minds. *Mind & Language*, *25*, 3–29.

van Dijk, L., & Rietveld, E. (2017). Foregrounding sociomaterial practice in our understanding of affordances: The skilled intentionality framework. *Frontiers in Psychology*, *7*, 1969.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE*, *9*, e112575.

Yildiz, I. B., von Kriegstein, K., & Kiebel, S. J. (2013). From birdsong to human speech recognition: Bayesian inference on a hierarchy of nonlinear dynamical systems. *PLoS Computational Biology*, *9*, e1003219.

Zander, T. O., Gaertner, M., Kothe, C., & Vilimek, R. (2011). Combining eye gaze input with a brain-computer interface for touchless human-computer interaction. *International Journal of Human-Computer Interaction*, *27*, 38–51.