Minds & Machines (2015) 25:321–338
DOI 10.1007/s11023-015-9384-8

How Sensorimotor Interactions Enable Sentence Imitation

Tzu-Wei Hung[1]

Received: 8 August 2015 / Accepted: 23 September 2015 / Published online: 28 September 2015

Abstract Despite intensive debates regarding action imitation and sentence imitation, few studies have examined their relationship. In this paper, we argue that the mechanism of action imitation is necessary and in some cases sufficient to describe sentence imitation. We first develop a framework for action imitation in which key ideas of Hurley's shared circuits model are integrated with Wolpert et al.'s motor selection mechanism and its extensions. We then explain how this action-based framework clarifies sentence imitation without a language-specific faculty. Finally, we discuss the empirical support for and philosophical significance of this perspective.

Keywords Sensorimotor interactions · Action imitation · Sentence imitation · Word-referent mapping · Syntactic abstraction

Introduction

We defend the central proposal that the sensorimotor mechanism for action imitation is necessary and in some cases sufficient to describe sentence imitation—the human capacity of producing a more or less exact copy of an observed sentence.

More specifically, sentence imitation may involve: (i) duplicating the entire sentence, (ii) copying the structure but changing some words, and (iii) changing both the structure and words but retaining the general meaning of the sentence. These different types of sentence imitation serve as important indicators of linguistic competence among children and adults with specific language impairments (Ratner and Sih 1987; Seeff-Gabriel et al. 2010; Silverman and Ratner 1997;

✉ Tzu-Wei Hung
htw@gate.sinica.edu.tw

[1] Institute of European and American Studies, Academia Sinica, No. 128, Sec. 2, Academia Rd., Nankang, Taipei 115, Taiwan

Verhoeven et al. 2011) and those without impairment (Miller 1973; Nelson et al. 1973).

The human capacity for understanding and producing instrumental actions (i.e., intentional actions with means-end structures) is highly relevant to the ability to understand and produce language (Byrne 2006; Garrod and Pickering 2008; Kiverstein and Clark 2008; Wolpert et al. 2003). Thus, the extent to which the mechanism that enables action imitation also facilitates language imitation is an interesting question. However, despite the intensive debates regarding action imitation and sentence imitation, relatively few reports have examined the relationship between their underlying mechanisms.

Among the recent studies related to this issue, Over and Gattis (2010) described verbal imitation with the *intention-based account of action imitation*, in which the distributed process patterns of an utterance are recombined when the hearer detects the speaker's intention. Their experiment indicated that children do not correct ungrammatical sentences (by not reproducing heard errors) until they recognize a speaker as an intentional agent. Pulvermüller and Fadiga (2010) contended that action and perception are functionally interdependent. Their data from transcranial magnetic stimulation showed that sensorimotor circuits offer a cortical basis for understanding phonemes, sentence structure, and grammar. Boza et al.'s (2011) artificial control system emulates general sociocognitive capacities, which is an engineering upgrade of Hurley's (2008) shared circuits model (SCM)—a functional model that specifies behavior-related skills in terms of dynamic sensorimotor interactions. Tourville and Guenther (2011; see also Guenther and Vladusich 2012) provided a nicely detailed neural network named DIVA to explain speech acquisition and production. Likewise, Glenberg and Gallese (2012) proposed an action-based theory of language acquisition, comprehension, and production. Pickering and Garrod (2013; see also Garrod et al. 2014) elucidated language comprehension and production in terms of action perception and production, in which grasping a speaker's words relies on the listener's forward model. Hickok (2014) also presented the hierarchical state feedback control (HSFC) model to specify the architecture of speech production.

The above interdisciplinary research sheds light on different aspects of sentence imitation. However, many significant issues remain to be investigated. Over and Gattis (2010) insightfully identified the role of intention in the sense of sentence imitation described in (iii) above, but they did not test the possibilities of (i) and (ii). Pulvermüller and Fadiga's (2010) brain imaging studies of motor activation are important, but they have not revealed the causal links among the relevant functional components of the mind. Boza et al.'s (2011) system does not concern a specific human skill, and it operates primarily at Marr's (1982) algorithmic level, which concerns the ways in which the SCM can be implemented within various algorithms. Hickok's (2014) model greatly elucidated speech production, but sentence imitation is not its focus. Finally, while their models are promising, Tourville and Guenther (2011), Glenberg and Gallese (2012), and Pickering and Garrod (2013) all ignored the extensive literature on the philosophy of language. Thus, none of them have differentiated the literal meaning of a word token from a speaker's unconventionally implicated meaning when examining language

comprehension and discussed conventional order and constituency when talking about syntax.

Therefore, to complement these valuable studies, we provide a philosophical investigation into the sense of sentence imitation described in (ii) above—copying the structure but changing some words (hereafter simply sentence imitation). We present a framework at the functional-computational level, in which action comprehension and production are clarified in terms of sensorimotor interactions ("The Framework" section). We then extend this framework to sentence imitation and show how it handles *word segmentation*, *syntactic abstraction*, *meaning interpretation* and *sentence reproduction* ("Sentence Imitations" section). We next argue for the framework's advantages and philosophical significance ("Advantages and Empirical Support" section) and conclude that the proposed framework is necessary and in some cases sufficient to describe sentence imitation.

The Framework

What is the relationship between action imitation and sentence imitation? Although the overlap of action and language capacities is assumed (Hurley 2008; Pickering and Garrod 2013), the premise of language *as* action is not explicitly justified. We have argued elsewhere that conversation is literally an action because verbal communication belongs to communicative action and hence instrumental action; therefore, a mechanism of understanding and producing the instrumental action

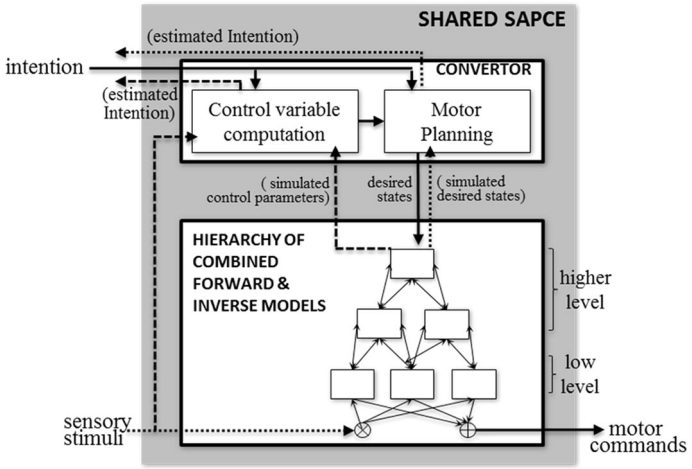

Fig. 1 The overall view of the proposed framework. It contains two main components: a *convertor* for transforming intention to desired motor states and a *hierarchy* for generating action in light of the desired states. It may operate in action production mode (*black lines*) and in action comprehension mode (information routes of the 1st method for action comprehension are marked with *dotted lines* and those of the 2nd method are marked with *dashed lines*)

should be applied to that of conversation (Hung 2014). The present paper advances this view via the following argument:

1. Action imitation involves segmenting continuous (visual) flow into constituents (e.g., movements), abstracting the sequence of the constituents, and reproducing the observed action with variation.
2. Sentence imitation also involves segmenting continuous (auditory) flow into constituents (e.g., words), abstracting the sequence of the constituents, and reproducing the heard sentence with variation.
3. Empirical data show that certain motor selection models functionally describe action imitation.
4. Therefore, these models can likely be used to functionally describe sentence imitation as well.

Based on the above argument, our framework borrows key concepts from Hurley's (2008) SCM and integrates them with Wolpert et al.'s (2003) *hierarchical modular selection and identification for control* (HMOSAIC) and its extensions (Haruno et al. 2003; Oztop et al. 2005). First of all, the framework contains a *shared space* (Fig. 1, grey box) between perception and action. This shared space receives the actor's *intention* and *sensory stimuli* as input and outputs *motor commands* for muscle contraction to generate actual behaviors. Within the space, there is a *convertor* that transforms the actor's intention (e.g., quenching thirst) to desired motor states (e.g., walk to a table, pick up the beer, and drink it)[1] and a *hierarchy of combined forward and inverse models* for executing the desired states.

More specifically, the convertor contains the mechanisms of *control variable processing* and *motor planning*. The former receives the intention and sensory stimuli as input and outputs control parameters (e.g., the distance between the beer and the actor's hand) to the latter. Motor planning synthesizes the parameters and intention and then outputs desired motor states for execution (see Appendix section "The Convertor" for the convertor's mathematic description).

Next, the desired motor states are sent to the hierarchy to determine what motor commands should be issued to complete the entire action of drinking the beer. Within each combined model, numerous pairs of predictors and controllers are working in a competitive manner. In each pair, the controller receives a desired motor state and an actual state, and it outputs a motor command (i.e., inverse model). The efference copy of this command is send to the paired predictor to simulate the possible next state (i.e., forward model). Only a motor command leading to a minimal difference between possible and actual states will be output by each model (see Appendix section "A Combined Forward and Inverse Model" for the mathematic description of a combined model).

These models are arranged hierarchically to enable more accurate motor control and prediction. Wolpert et al. (2003) argued that humans can produce a number of compensating movements to preserve the kinematics of writing across different

[1] Human intention is not flat but multi-layered. According to Hurley (2001), it can be roughly divided into *nonbasic intention* (the goal of an actor) and *basic intention* (desired means to achieve that goal).

instruments, suggesting the existence of high-level reference signals (e.g., intentions) that have many ways of activating low-level controllers. Accordingly, the framework's high-level models receive desired states and output motor commands in a certain sequential order (which determine the behavior of subordinate models). The subordinate level then outputs the commands of generating movements needed for completing the entire action (which determine the activation of the low-level models). The low-level models then compare the predictive state of the efference copy with the actual state to revise the next motor command.

The hierarchy's levels may have arbitrary depth, depending on the complexity of a task. When Bayesian terms are used to describe this cross-model communication, controllers at higher levels receive only posterior probabilities from models at subordinate levels. Predictors at higher levels generate prior probabilities for models at subordinate levels (see Appendix section "The Hierarchy of Models").

Finally, this framework operates in the mode of not only action production but also action comprehension. When seeing someone move a beer toward his or her mouth, an observer's low-level models will segment the action into constitutive movements;[2] these constituents are then sent to a higher level to determine whether they contradict any learned sequence.[3] The testing result will be sent to the highest models to identify the desired motor states.[4] If the actor's movements are in the observer's repertoire of motor states, the observer's motor planning can easily associate estimated desired states with the actor's possible intention through mirroring.[5] However, if the actor's action is unfamiliar to the observer (e.g., moving a hammer toward the mouth), a second method is needed to infer the actor's intention.

For example, research has shown that understanding the goal in an unfamiliar or complicated situation largely depends on the observer's inferential processes; otherwise, mirroring is more important (Brass et al. 2007; Hamilton 2013). However, how does the observer's framework identify the actor's intended goal by inference? According to Oztop et al. (2005), an actor's intention parameterizes the motor control system to generate actions. Hence, the observer can analyze observed actions to predict the actor's motor parameters and subsequently the goal. In the framework, through control variable processing, the observer receives sensory stimuli and encodes them into observed parameters. If the observed parameters match with the (prior) predictive parameters generated by the hierarchy, then the

[2] Low-level controllers will generate commands needed to complete an action similar to the observed one. If the paired predictions match with actual subsequent states, then these commands represent the appropriately segmented movements of the observed action (Wolpert et al. 2003).

[3] According to Haruno et al. (2003), the higher level can learn the pattern (sequence) of movements on a probabilistic base. Please see Appendix section "Abstraction of the Sequence of Constituents".

[4] To identify the actor's desired states X_t^*, the observer's highest level needs to issue predictive states $\hat{X}_t$ that have the least mismatch with actual states X_t. This final prediction's paired command U_t can be described by Eq. (5) in Appendix "The Hierarchy of Models".

[5] In the debate regarding what mirror neurons mirror, there are three main hypotheses (Oztop et al. 2005): mirror neurons encode (i) the detailed low-level motor parameters of the observed action; (ii) the higher-level motor plan; or (iii) the actor's intention. In this paper, we presuppose view (iii).

estimated intention of the actor can be derived from the comparison result (see Appendix section "Control Variable Processing for Estimated Intention"). The estimated intention is sent downward through the hierarchy for motor executing and calibration. If the low-level commands are inhibited, then the framework enables action comprehension. If the low-level commands are not inhibited, the observer duplicates the actor's action. These two methods can work both together and independently, especially when method is impaired.

To sum up, our framework is action oriented and based on several well-established models for action imitation. However, how is this framework extended to sentence perception and production, and how does it explain sentence imitation?

Sentence Imitations

Sentence imitation requires a hearer to understand the speaker's utterance before reproducing it. To comprehend an utterance, the hearer needs to segment the speaker's speech flow into phonetic words and map them onto appropriate referents (Hickok and Poeppel 2004).[6] Since a word's referent may depend on its position (e.g., indexicals) and its relationship with other words in the sequence (e.g., anaphora), the sequential order of words (i.e., syntax) should also be understood. The meaning and arrangement of words determine the semantic value of the entire sentence (Dever 2012). However, the semantic properties of a sentence are not the only clue to understand the sentence. People sometimes infer what speakers say (i.e., semantics) by grasping their intention (i.e., pragmatics), but they sometimes infer the intention based on what they say. Accordingly, we divide sentence imitation into following main explananda (Fig. 2) and explain each:[7] *segmenting phonetic words* ("Segmentation of Phonetic Words" section), *abstracting the syntax* ("Syntactic Abstraction" section), *comprehending semantics* and/or *pragmatics* ("Meaning Comprehension" section), and *reproducing the sentence* ("Sentence Reproduction" section).

Segmentation of Phonetic Words

First, suppose when a speaker utters "Ella saw Sue"; the utterance will be produced as a continuous speech flow (e.g., [ɛ]-[l]-[ə]-[s]-[ɔ]-[s]-[u]). Meanwhile, the framework of a competent hearer needs to decide how to segment the flow into appropriate constituents (e.g., whether [ɛlə]-[sɔ]-[su] or [ɛl]-[əs]-[ɔsu] is properly segmented). To do so, the hearer's framework produces different motor commands for phonation (e.g., to utter [ɛlə] or [ɛl]) at each round of segmentation. Each command results in a prediction of the next possible state, and each prediction will

[6] Hickok and Poeppel (2004) show that the brain's speech perception is realized in two processing streams: dorsal stream maps sound into articulation-based representation and ventral stream maps sound into meaning, which are interfaced by the posterior region of the middle temporal gyrus.

[7] For simplicity, although a hearer's visual perception of a speaker's lip movements might affect how the speaker's sound is perceived (e.g., the McGurk effect), we consider only auditory input, which by no means indicates that the framework is not applicable to the visual processing of written sentences.

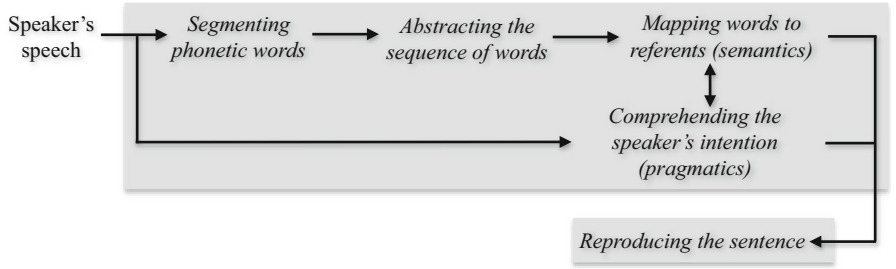

Fig. 2 Components of sentence imitation. The *big grey box* indicates the hearer's sentence comprehension, while the small one indicates the hearer's sentence reproduction

be tested with the actual next state. If they match and if the testing result shows no contradiction with higher models,[8] then that command (e.g., the command for producing [ɛlə]) is a properly segmented constituent (e.g., "Ella") of the utterance.[9] This testing process would be difficult if the speaker's words are not in the hearer's repertoire of words. For competent hearers, however, prior linguistic knowledge increases the accuracy of the initial prediction and hence reduces the testing time. Nonetheless, in some phonetically ambiguous cases (e.g., "ice cream" and "I scream"), lower-level prediction alone may be insufficient to determine the segmentation, so it has to work with a higher-level prediction of word meaning (see 3.3) to fix this problem.

Syntactic Abstraction

Outputs from the low level are then sent to a higher level to abstract the syntax of the utterance. Haruno et al. (2003) argued that the HMOSAIC's high level functions as a pattern generator with which to learn movement sequences by implementing recurrent neural networks. We argue that the same method can be used to abstract word sequences from continuous speech. Briefly, the higher level receives a sequence of input $[x_0, x_1, \ldots, x_t]$ from a subordinate level and generates a predictive sequence $[\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{t+1}]$. If the predictive sequence matches with the next actual sequence $[x_1, x_2, \ldots, x_{t+1}]$, then the predictive sequence is the correct one. The testing result (responsibility signal) can be sent to an even higher level to identify the speaker's desired motor state (see 3.3). It can also be used to revise higher-level

[8] As the hearer's perception of the number of syllables in a word is also determined by the sequential arrangement of phonemes (Mannell et al. 2014), the low level needs to check its output with sequence processing at a higher level.

[9] To describe this process computationally, the framework activates low-level controllers 1, 2, 3,…, n and generates commands $u_t^1, u_t^2, u_t^3, \ldots, u_t^n$. Each efference copy of a motor command is send to its paired predictor to generate a prediction $\hat{x}_{t+1}^i = \Phi\left(w_t^i, x_t, u_t\right)$. Each prediction is then compared with the sensory input to generate responsibility signal $\lambda_t^i = \frac{e^{-\left|x_{t+1}-\hat{x}_t^i\right|^2/\sigma^2}}{\sum_{j=1}^{n} e^{-\left|x_{t+1}-\hat{x}_t^j\right|^2/\sigma^2}}$. Each signal helps a controller to revise its motor command, and the final command of the entire framework can be generated through $u_t = \sum_{i=1}^{n} \lambda_t^i u_t^i$. This final motor commands is a properly segmented constituent (usually a word or free morpheme) of the utterance.

commands in order to regulate the behavior of models at the subordinate level (Appendix section "Abstraction of the Sequence of Constituents"). As the sequential order determines whether words are formed into sentences, it functions as syntax.

However, one might argue that syntax concerns not only word order but also constituency (i.e., "the bracketing of elements (typically words) into higher-order elements"; Evans and Levinson 2009, p. 440). Thus, a noun [apple] is a constituent of a noun phrase [[the] [apple]], which is a constituent of a sentence [[Ella] [[saw] [[the] [apple]]]]. Our reply is that the framework can also learn constituency. When uttering, the speaker's motor control system is parameterized by the speaker's intention, in which the syntactic conventions (e.g., constituency) of the speaker's speech community are also encoded. Thus, the hearer will encounter "Ella saw Sue" more frequently than "Ella Sue saw" because English language conventions forbid the latter. If the hearer's framework detects sentences such as "Ella saw Sue," "John saw Sue," and "Ella saw John," then it will predict that it is highly probable that the first element in the sequence (i.e., the subject) is changeable. Likewise, if the framework detects "Mary loves John," "Mary greets John," and "Mary calls John," it will statistically learn that the second element (i.e., the verb) is variable as well.[10] This process helps the framework to predict the replaceability of elements (words, phrases, or clauses) and capture constituency on a probabilistic basis.

Meaning Comprehension

The framework employs two methods to understand what the speaker intended to convey by uttering a sentence: (1) identify the meaning of the constituent words so that the entire sentence can be understood; (2) detect the speaker's intention by analyzing observed motor variables.

On the one hand, the output of the syntactic abstraction (i.e., constituents with a correct sequential order [ɛlə] → [sɔ] → [su]) can be used by the hierarchy's highest level to identify the actor's desired motor states (e.g., uttering "Ella saw Sue"). This level generates predictions of the actor's desired motor states and calibrates them against the next actual states. It then sends out the most likely predictions to the convertor. In the mode of production, the convertor receives intention d_t and actual state x_t, and it outputs desired state x^*_{t+1} (Appendix section "The Convertor"). However, in the mode of comprehension, the convertor can derive intention d_t from observed actual states and predictive desired states. When the predictive states exist in the hearer's repertoire, the convertor can easily associate a constituent ("Ella") with the referent to which the speaker intends to refer (a friend named Ella).[11] Hence, the entire sentence can be understood by comprehending its constituent words. This first method conforms to *the principle of compositionality* in the philosophy of language, which indicates that the meaning of a sentence depends on the meaning of its elements and their arrangement.

[10] Recursive processing is required for constituency and is presupposed by the framework.

[11] Following Hurley (2008), we also assume that the mechanism of an actor's intended goal can be used to identify a speaker's intended referent. Please see Hung (2014) for a relevant discussion.

On the other hand, Hickok and Poeppel's (2004) experiment on autism shows that the capacity of segmenting phonemes and that of sound-meaning mapping are double dissociative: one can function pretty well when the other is impaired. This finding seems to be incompatible with the conventional view that phonetic segmentation is a prerequisite to sound-meaning mapping. How can this incompatibility be explained? One merit of our framework is that it sorts out this incompatibility. In action comprehension, the framework may exploit a second method to derive the actor's intention from control variables of the actor's action. Likewise, processing the control variables may facilitate the computation of the parameters of heard speech (e.g., its tune and pressing) and nonlinguistic motor clues (e.g., whether the speaker is pointing/looking at an object) to infer the speaker's intended meaning. This inference can be implemented by Oztop et al.'s (2005) algorithm of mental state search (Appendix section "Searching for Meaning"). Thus, if the phonetic processing in the first method (semantics) is impaired, the framework may exploit the second method (pragmatics) to infer the meaning of the speaker's words. Conversely, only when both methods are impaired can the hearer's capacity of understanding the speaker's words be hindered, although low-level phonetic processing in the first method might remain intact. In other words, the framework clarifies why two valuable opinions (the conventional view and Hickok and Poeppel's (2004) finding) are compatible.

Nonetheless, skeptics might argue that the above account oversimplifies the dynamic essence of meaning comprehension. At the very least, for example, Austin (1962) and Grice (1975) have shown that speakers may intend to express something beyond what they actually utter (i.e., implicature) and that what a speaker actually utters can be interpreted *literally* or *contextually* (Cappelen and Lepore 2005). Thus, how does a hearer differentiate those meanings in real time? Here, we use four examples we have offered (Hung 2014) to show how the framework learns to distinguish them.

(i) Lexical meaning of *word type*: itis conventional and context-independent, and it refers to abstract entities (e.g., "she" is a third-person singular feminine pronoun, and "penguin" is the term for an aquatic bird living in the southern hemisphere);
(ii) Lexical meaning of *word token*: it is conventional but context dependent, and it refers to individual concretes (e.g., "she" can mean someone's mother or daughter, and "penguin" can refer to a species or an individual organism);
(iii) Speaker's *conventionally* implicated meaning: it is what a speaker intends to convey beyond his conventional use of words (e.g., replying, "I am married" to the query, "Can I have your number?" or answering, "I am Brazilian" to the question, "Do you play soccer?");
(iv) Speaker's *unconventionally* implicated meaning: it is what a speaker intends to convey beyond his unconventional use of words (e.g., a cleaner says to colleagues, "Check out the massive chocolate in the toilet," or someone who named his boat "Penguin" says, "My Penguin is sick").

Here, the framework can use its combined forward and inverse model to differentiate meanings through trial and error,[12] which is outlined graphically by Fig. 3. Suppose that during training, the framework detects a speaker by using "she" to indicate a woman on one occasion but a different woman on another occasion. On each occasion, the word "she" is successfully mapped onto a referent in the sense of (ii). If the word has been mapped to similar referents across a number of occasions, then the framework can fine tune the mapping to extrapolate (i) from (ii). The mapping described in (ii) is taken as the default prediction whenever the framework receives the same word on new occasions. For example, if default predictions match both input words (e.g., "I am Brazilian") and other contextual clues (e.g., the speaker who says this is, in fact, Brazilian), then the framework has made a correct prediction/confirmation of (ii). Nevertheless, default predictions can easily fail. If a default prediction contradicts an input word, as in (iv), then the framework must revise its predictions. The adjusted prediction, when matched with contextual clues, correctly predicts (iv). Alternatively, if a default prediction matches an input word but seems irrelevant with regard to the contextual clues, as in (iii), then the framework must modify its predictions again, which, if matched, correctly predict (iii). Therefore, the framework differentiates meanings (i)–(iv).[13]

Sentence Reproduction

Finally, the hearer's intention of preserving the structure but changing the words (i.e., sentence imitation) can be input into the convertor to generate desired motor states (e.g., uttering "Ella greets Sue"). These desired states are sent downward in the hierarchy for motor execution. The hierarchy generates and revises motor commands at different levels, ensuring that low-level controllers issue appropriate commands for uttering "Ella greets Sue." Thus, the hearer's sentence imitation is completed.

To sum up, sentence imitation would be impossible without the above mechanisms of *segmenting words*, *abstracting syntax*, *comprehending semantics* and/or *pragmatics*, and *reproducing the sentence*. In this sense, the framework is necessary for sentence imitation. In addition, the framework is sufficient, as it alone can achieve imitation of simple sentences in the case discussed above. Therefore, the central proposal holds.

Advantages and Empirical Support

What qualities make the framework better than previous architectures, such as those based on a domain-specific language faculty? What empirical support is there for the framework? The framework's merits are argued in terms of its philosophical significance, compatibility, explanatory power, and simplicity.

[12] Based on Eq. (2) in Appendix section "A combined Forward and Inverse Model", we may define a gradient learning rule of each controller, in which the desired command $\left(u_t^* - u_t^i\right)$ can be approximated by using the feedback command u_{fb}:$\Delta\alpha_t^i = \epsilon\lambda_t^i \frac{dy_{t^i}}{d\alpha_t^i}\left(u_t^* - u_t^i\right) = \epsilon\frac{du_t^i}{d\alpha_t^i}\lambda_t^i\left(u_t^* - u_t^i\right) \cong \epsilon\frac{du_t^i}{d\alpha_t^i}\lambda_t^i u_{fb}$

[13] However, when a fluent speaker intentionally utters a sentence, not all of his or her words are necessarily consciously selected or explicitly intended. Nevertheless, this does not prevent the framework from using the mechanism to understand the speaker's words because the words are linked to what the speaker would likely intend if she were aware of her word selection.

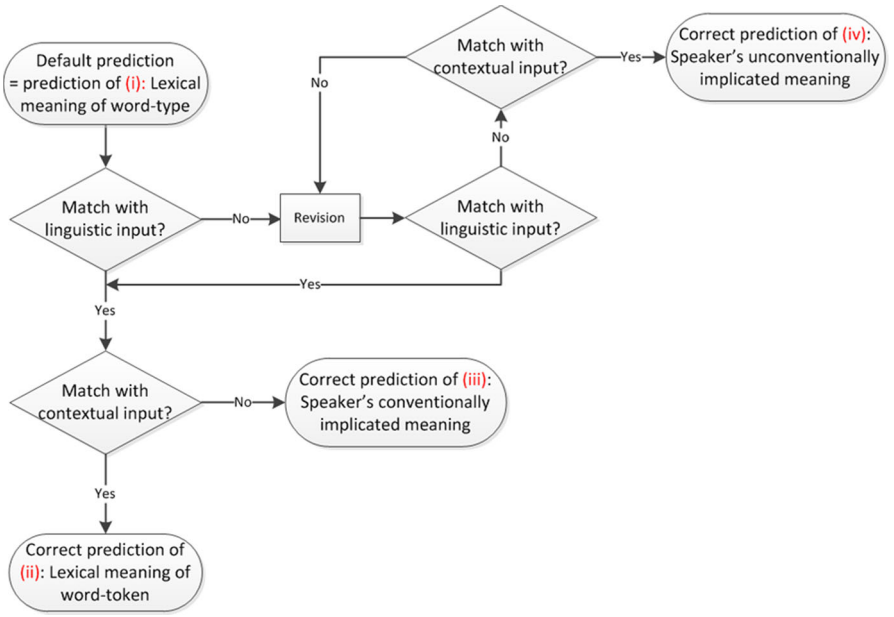

Fig. 3 Flowchart of meaning differentiation. This flowchart shows how a hearer is able to comprehend the meaning of words in diverse situations

In the debates regarding cognitive architecture, some theorists view the mind as a central processing system that regulates input and output systems. The central system, which usually processes symbolic representations according to formal rules, is highly modular for various higher cognitive tasks (Sperber 2002; Cosmides and Tooby 1992; Barrett and Kurzban 2006); for instance, there is a language faculty for linguistic abilities (Pinker 1994; Carruthers 2006). Conversely, behavior-based theorists hold that cognition emerges from the dynamic interaction between action, perception, and the world (Brooks 1999; Hurley 2008). Although they can better explain real-time and motor-related cognitive skills, owing to their rejection of a central system (and a language faculty), it is doubtful whether they can account for higher cognition, such as language. Moreover, because models under the labels of "situated cognition," "embodied cognition," and "extended mind" also lack a central processor and thus face a similar challenge, a solution to the behavior-based system may provide valuable input for the solution to all of these models. This explains why the framework is of philosophical significance.

The framework is developed from well-established models of behavior-related cognition (Haruno et al. 2003; Hurley 2008; Oztop et al. 2005; Wolpert et al. 2003). However, it also conforms to studies on language acquisition and processing. For example, this framework rejects the Chomskian view that the cognition system relies on prestored syntactical structures to handle language. Rather, its higher level functions as a pattern generator to capture syntax on a likelihood basis. This resembles Thompson and Newport's (2007) view that transitional probability plays

a role in the statistical learning of syntax and Clark and Lappin's (2010) demonstration that an artificial system with few domain-specific learning biases was capable of extracting syntax from a stream of linguistic stimuli by using a probabilistic learning method. Moreover, the framework is at the functional-computational level, but the claim that action and language processing share the same mechanism is also support by neural and cortical studies (Glenberg and Gallese 2012; Pickering and Garrod 2013).

Explanatory power is also a merit of the framework. We have seen in 3.3 that the framework clarifies why two seemly incompatible but valuable claims both hold (i.e., the conventional view that that phonetic segmentation is a prerequisite to sound-meaning mapping and Hickok and Poeppel's (2004) finding that the capacity of segmenting phonemes and that of sound-meaning mapping words are double dissociative). Moreover, the framework better explains pragmatics and dynamic motor clues that show what the speaker intends to convey by uttering a sentence. Such explanatory power arises from the fact that the framework does not assume classical computational theory of mind and hence avoids the globality problem of computation that Fodor (2008) noted, i.e., classical computation is merely sensitive to syntax, but the mind (and its language processing) is not.

Likewise, the framework is supported by pathological evidence. It implies that the mechanism of means-end mapping is necessary to link words with meaning. Hence, people in whom this mechanism is impaired are unlikely to have an intact word-meaning mapping capacity. In fact, although SPLD is a heterogeneous syndrome resulting from various causes, the deficit of meaning interpretation occurs frequently with the deficit of instrumental action comprehension. For example, patients with Rett syndrome do not exhibit means-end behavior beyond automatic responses to particular stimuli (Woodyatt and Ozanne 1992), and they have difficulty using words for functional communicative purposes. Cass et al. (2003) found that only 18 of 84 participants with Rett syndrome reported using words and that only six used words in meaningful ways. Likewise, a high proportion of children with autism do not develop the skill to form and manipulate symbolic material (Prior and Ozonoff 2007), and the acquisition of this skill is likely associated with means-end reasoning skills.

A prediction of the framework is that since the capacity of means-end associations is a prerequisite for the ability to form symbol-referent associations, the former is unlikely to develop after the latter. There is no direct evidence of this prediction yet, but it conforms to existing findings. Although infants can solve simple means-end problems such as pulling a cloth to retrieve a toy at as early as 6 months (Willatts 1999), they do not look at the correct portrait when hearing "Mommy" or "Daddy" until 6 months (Tincoff and Jusczyk 1999), and they can map meaning to newly segmented words only at 17 months (Graf Estes et al. 2007).

The framework's final merit lies in its simplicity. A model of language faculty is input specific; thus, it requires an additional module to handle nonlinguistic input. If two modules are responsible for different information types, then an interface for their communication must exist. In contrast, the framework assumes neither a language faculty nor an interface between linguistic and nonlinguistic cognitive

components. It explains both action and sentence imitations at once. Therefore, according to Occam's Razor, the framework is simpler and better than those that presuppose an extra language faculty.

Conclusions

To summarize, this paper justifies the view that a sensorimotor mechanism for action imitation also describes sentence imitation. We first propose a framework and show how it explains action comprehension and reproduction (with the mathematic description of the framework supplied in Appendix). We then divide sentence imitation into segmenting phonetic words, abstracting syntax, comprehending meaning, and reproducing the sentence and show how the framework clarifies each subtask individually. Finally, we provide empirical evidence in support of the framework.

If we suppose that our proposed linking of action and language is correct, all other things being equal, additional training in verbal skills might somewhat advance the motor skills of bilingual people. Confirming this view, recent studies show that bilingual children perform better than monolingual children with regard to domain-general control skills (Kovács and Mehler 2009) and executive control of spatial reasoning (Greenberg et al. 2013). In addition, because word learning requires intention detection, bilinguals exhibit advantages with regard to theory of mind (Kovács 2009).

In summary, the contribution of this framework lies in its illustration of sentence imitation with regard to the mechanism for action imitation, which partially clarifies the relationship between two significant human capacities. Nevertheless, because this study focuses on proposing a descriptive framework rather than reporting empirical results, experimental simulations, along with relevant issues not covered in this paper, should be the focus of future studies.

Acknowledgments This research was sponsored in part by the Ministry of Science and Technology, Taiwan under Grant No. 101-2410-H-001-100-MY2.

Appendix

The Convertor

The convertor receives the actor's intention d_t and actual state x_t at time t, and it outputs desired state x^*_{t+1} at time $t + 1$. We use $p_t = f(x_t, d_t)$ and $x^*_{t+1} = g(d_t, p_t)$ to describe the parameter generated by motor control processing and a desired motor state generated by motor planning, where f and g are functions with inverse relationship $x^*_{t+1} = g(d_t, f(x_t, d_t))$.

A Combined Forward and Inverse Model

Suppose each model activates multiple predictors 1, 2, 3,…, n at t, and select among their next state predictions $\hat{x}_{t+1}^1, \hat{x}_{t+1}^2, \hat{x}_{t+1}^3, \ldots, \hat{x}_{t+1}^n$ through testing (see Fig. 3). Each predictor receives actual feedback x_t and the efference copy of motor command u_t to generate a prediction. The prediction of the i-th predictor is $\hat{x}_{t+1}^i = \Phi\left(w_t^i, x_t, u_t\right)$, where w_t^i represents the parameters of the function approximator Φ. This predictive next state is compared with the actual next state. If an error occurs, then the wrong prediction is sent to a responsibility estimator to generate responsibility signal λ_t^i, which can be calculated by using the softmax activation function.

$$\lambda_t^i = \frac{e^{-\left|x_t + -\hat{x}_t^i\right|^2 / \sigma^2}}{\sum_{j=1}^n e^{-\left|x_t + -\hat{x}_t^j\right|^2 / \sigma^2}} \tag{1}$$

In Eq. (1), x_t is the framework's actual voice output, and σ is a scaling constant. The softmax activation function calculates the error signals and normalizes them into probability values between 0 and 1. Predictors with few errors receive higher responsibilities. Thus, responsibility signals can regulate predictor learning in a competitive manner. Moreover, a paired controller exists for each predictor, and it receives the desire next state x_{t+1}^* and outputs motor commands. Suppose that the framework activates controllers 1, 2, 3,…, n and generates $u_t^1, u_t^2, u_t^3, \ldots, u_t^n$. The motor command of the i-th controllers is $u_t^i = \psi\left(\alpha_t^i, x_{t+1}^*\right)$, where α_t^i is the parameter of a function approximator ψ. The summation of the motor commands generated by controllers 1, 2, 3,…, n is represented by Eq. (2).

$$u_t = \sum_{i=1}^n \lambda_t^i u_t^i = \sum_{i=1}^n \lambda_t^i \psi\left(\alpha_t^i, x_{t+1}^*\right) \tag{2}$$

The Hierarchy of Models

For simplicity, we describe only a two-level hierarchy, although it is extendable to an arbitrary number of levels. Suppose that the predictor of the i-th higher-level model receives actual state X_t and the efference copy of U_t at t, and suppose that it outputs the approximate prediction $\hat{X}_{t+1}^i$ without activating subordinate controllers:

$$\hat{X}_{t+1}^i = \Phi\left(W_t^i, X_t, U_t\right) = \left(P\left(1|W_t^i, X_t, U_t\right), \ldots, P\left(n|W_t^i, X_t, U_t\right)\right) \tag{3}$$

In Eq. (3), Φ refers to a vector-valued and nonlinear function approximator; W_t^i is the synaptic weight of the higher-level j-th pair; X_t is the current state (posterior probability); U_t is the higher-level command; and $P\left(j|W_t^i, X_t, U_t\right)$ refers to the posterior probability in which the j-th pair is selected under W_t^i, X_t, and U_t. Likewise, the i-th higher-level prediction $\hat{X}_{t+1}^i$ is compared with actual state X_t from the subordinate level to generate higher-level responsibility (i.e., prior probability) $\lambda_i^H(t)$ via the estimator

$$\lambda_i^H(t) = \frac{\hat{\lambda}_i^H(t)e^{-\left|x_t-\hat{x}_t^i\right|^2/\sigma^2}}{\sum_{j=1}^{N}\hat{\lambda}_i^H(t)e^{-\left|x_t-\hat{x}_t^i\right|^2/\sigma^2}} \tag{4}$$

$\lambda_i^H(t)$ can regulate the subordinate level in a competitive manner. Moreover, each higher-v predictor has a paired controller that receives the desired next state X_{t+1}^* and current state X_t from the subordinate level as input. X_{t+1}^* is an abstract representation that determines the selection and activation order of subordinate controllers. Each higher-level controller generates commands to the subordinate level, and the command of the i-th higher-level controller is $U_t^i = \Psi\left(\Lambda_t^i, X_{t+1}^*, X_t\right)$, where Λ_t^i is the parameter of a function approximator Ψ. Then, U_t, the summation of (prior probability) commands for the lowest pairs, is weighted by $\lambda_i^H(t)$:

$$U_t = \left(\hat{\lambda}_1^L(t), \ldots, \hat{\lambda}_n^L(t)\right) = \sum_{i=1}^{N}\lambda_i^H(t)U_t^i = \sum_{i=1}^{N}\lambda_t^H(t)\Psi\left(\Lambda_t^i, X_{t+1}^*, X_t\right) \tag{5}$$

Abstraction of the Sequence of Constituents

Suppose that the k-th higher-level model receives a sequence of actual input $[x_0, x_1, \ldots, x_t]$ (represented by X_t) and generates a prediction regarding a sequence of output $[\hat{x}_1^{\mathrm{k}}, \ldots, \hat{x}_t^{\mathrm{k}}]$ (represented by $\hat{X}_{t+1}^{\mathrm{k}}$). The task of this higher level is to determine the prediction that has the least mismatch with the next actual sequence of input. The comparison result can be represented as Eq. (4). The responsibility signal can be used to revise higher-level commands (see Eq. (5)), which determines the behavior of the subordinate level. The efference copy of the commands can also be used for further prediction (and revision). We also use a recurrent network to describe the higher-level prediction of sequence $\hat{X}_{t+1}^{\mathrm{k}} = f\left(W_t^k, X_t\right)$, in which f is a nonlinear function that can use weights W_t^k to predict a vector of posterior probabilities. The network dynamics can be described as:

$$\tau\frac{d}{d_t}a_i(t) = -a_i(t) + \sum_{j=1}W_{ij}^K b_j(t)$$

$$b_i(t) = \begin{cases} g(a_i(t)) & (\text{output}) \\ X_t^i & (\text{input}) \end{cases}$$

In the above differential equation, $g(X)$ is the sigmoid function with derivative $g(X)(1-g(X))$, a_i is the activation, and b_i is the output at the i-th node. In Haruno et al.'s (2003) simulation, their models successfully learned two sequences and determined the one that should be reproduced under a given context, even when 5 % noise was added.

Control Variable Processing for Estimated Intention

Suppose that the highest models generate predictive control parameters $\hat{p}_t^1, \hat{p}_t^2, \hat{p}_t^3, \ldots, \hat{p}_t^n$ at time t. Each predictive parameter will be compared with

actually observed parameter p_t from the control variable encoding (Fig. 4). The comparison result is represented by the responsibility signal:

$$\lambda_t^i = \frac{e^{-\left|p_t + -\hat{p}_t^i\right|^2 / \sigma^2}}{\sum_{j=1}^{n} e^{-\left|p_t + -\hat{p}_t^i\right|^2 / \sigma^2}} \tag{6}$$

The mechanism of intention/parameter matching in control variable processing generates simulated intention $\hat{d}_t^1, \hat{d}_t^2, \hat{d}_t^3, \ldots, \hat{d}_t^n$. The final estimated intention is represented by:

$$d_t = \sum_{i=1}^{n} \lambda_t^i d_t^i \tag{7}$$

Searching for Meaning

Here, we use Oztop et al.'s (2005) algorithm of mental state search to infer the meaning that the speaker intends to convey. To initialize the algorithm, we also set T_k and S_k to an empty sequence ($T_k = S_k = []$). T_k and S_k represent sequences of observed and mentally simulated vectors of control variables extracted under the mental state k. Next, repeat steps (1)–(5) from speech onset to speech end.

1. Pick next possible mental state (j) (which can be thought of as an index for the possible referent to which the speaker is referring).
2. *Observe*: Extract the relevant control variables based on the hypothesized mental state (j), x_j^i, and add them to T_j ($T_j = [T_j, x_j^i]$). Here, i indicates that the collected data were placed in ith position in the visual control variable sequence.
3. *Simulate*: Mentally simulate speech with mental state j while storing the simulated control variables x_j in S_j ($S_j = [x_j^0, x_j^1, \ldots, x_j^N]$, where N is the number of control variables collected during observation).
4. *Compare*: Compute the discounted difference between T_j and S_j, where N is the length of T_j and S_j. $D_N = \frac{(1-\gamma)}{(1=\gamma^{N+1})} \sum_{i=0}^{N} \left(x_{\text{sim}}^i - x^0\right)^T \mathbf{W} \left(x_{\text{sim}}^i - x^i\right) \gamma^{N-i}$, where x_{sim}^i

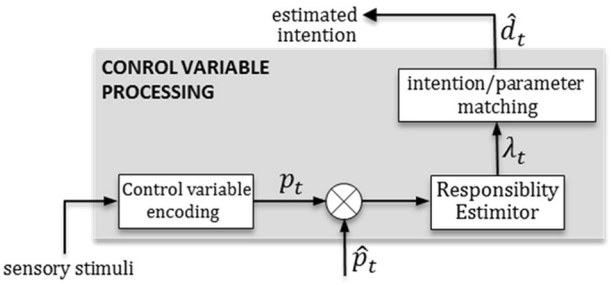

Fig. 4 Control variable processing in action comprehension mode

$\in Sj$ and $x^i \in Tj$ and **W** is a diagonal matrix normalizing components of x^i and γ is the discount factor.

5. If *DN* is smallest so far, set $j_{\min} = j$.

Return: $j_{\min}$ (the observer infers that $j_{\min}$ is the actor's intended meaning).

References

Austin, J. L. (1962). *How to do things with words*. Cambridge, MA: Harvard University Press.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: Framing the debate. *Psychological Review, 113*, 628–647.

Boza, A. S., Guerra, R. H., & Gajate, A. (2011). Artificial cognitive control system based on the shared circuits model of sociocognitive capacities. A first approach. *Engineering Applications of Artificial Intelligence, 24*(2), 209–219.

Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). Investigating action understanding: Inferential processes versus action simulation. *Current Biology, 17*, 2117–2121.

Brooks, R. A. (1999). *Cambrian intelligence: The early history of the New AI*. Cambridge, MA: The MIT Press.

Byrne, R. W. (2006). Parsing behaviour. A mundane origin for an extraordinary ability? In N. Enfield & S. Levinson (Eds.), *The roots of human sociality* (pp. 478–505). New York, NY: Berg.

Cappelen, H., & Lepore, E. (2005). *Insensitive semantics: A defense of semantic minimalism and speech act pluralism*. Oxford: Blackwell.

Carruthers, P. (2006). *The architecture of the mind*. Oxford: Oxford University Press.

Cass, H., Reilly, S., Owen, L., Wisbeach, A., Weekes, L., Slonims, V., & Charman, T. (2003). Findings from a multidisciplinary clinical case series of females with Rett syndrome. *Developmental Medicine and Child Neurology, 45*(5), 325–337.

Clark, A., & Lappin, S. (2010). *Linguistic nativism and the poverty of the stimulus*. Oxford: Wiley Blackwell.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. New York, NY: Oxford University Press.

Dever, J. (2012). Compositionality. In G. Russell & D. Graff Fara (Eds.), *The Routledge handbook to the philosophy of language* (pp. 91–102).

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32*(5), 429–492.

Fodor, J. (2008). *LOT 2: The language of thought revisited*. Oxford: Oxford University Press.

Garrod, S., Gambi, C., & Pickering, M. J. (2014). Prediction at all levels: forward model predictions can enhance comprehension. *Language, Cognition and Neuroscience, 29*(1), 46–48.

Garrod, S., & Pickering, M. J. (2008). Shared circuits in language and communication. *Behavioural and Brain Sciences, 31*(1), 26–27.

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

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18*, 254–260.

Greenberg, A., Bellana, B., & Bialystok, E. (2013). Perspective - taking ability in bilingual children: Extending advantages in executive control to spatial reasoning. *Cognitive Development, 28*(1), 41–50.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York, NY: Academic Press.

Guenther, F. H., & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics, 25*(5), 408–422.

Hamilton, A. F. (2013). The mirror neuron system contributes to social responding. *Cortex, 49*(10), 2957–2959.

Haruno, M., Wolpert, D. M., & Kawato, M. (2003). Hierarchical MOSAIC for movement generation. *International Congress Series, 1250*, 575–590.

Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language, Cognition and Neuroscience, 29*(1), 2–20.
Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition, 92*(1–2), 67–99.
Hung, T. -W. (Ed.). (2014). What action comprehension tells us about meaning interpretation. *Communicative action: selected papers of the 2013 IEAS conference on language and action*. Singapore: Springer.
Hurley, S. (2001). Perception and action: Alternative views. *Synthese*, *129*, 3–40.
Hurley, S. (2008). The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mind reading. *Behavioral and Brain Sciences, 31*(1), 1–22.
Kiverstein, J., & Clark, A. (2008). Bootstrapping the mind. *Behavioural and Brain Sciences, 31*(1), 41–52.
Kovács, Á. M. (2009). Early bilingualism enhances mechanisms of false-belief reasoning. *Developmental Science, 12*(1), 48–54.
Kovács, Á. M., & Mehler, J. (2009). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences, 106*, 6556–6560.
Mannell, R., Cox, F., & Harrington, J. (2014). An introduction to phonetics and phonology. Macquarie University. Retrieved 26 Sep 2015 from http://clas.mq.edu.au/speech/phonetics/index.html.
Marr, D. (1982). *Vision*. San Francisco, CA: W.H. Freeman.
Miller, J. F. (1973). Sentence imitation in pre-school children. *Language and Speech, 16*(1), 1–14.
Nelson, K. E., Carskaddon, G., & Bonvillian, J. D. (1973). Syntax acquisition: Impact of experimental variation in adult verbal interaction with the child. *Child Development, 44*(3), 497–504.
Over, H., & Gattis, M. (2010). Verbal imitation is based on intention understanding. *Cognitive Development, 25*(1), 46–55.
Oztop, E., Wolpert, D., & Kawato, M. (2005). Mental state inference using visual control parameters. *Cognitive Brain Research, 22*(2), 129–151.
Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioural and Brain Sciences, 36*(4), 329–347.
Pinker, S. (1994). *The language instinct: How the mind creates language*. New York, NY: Harper Collins.
Prior, M., & Ozonoff, S. (2007). Psychological factors in autism. In F. R. Volkmar (Ed.), *Autism and pervasive developmental disorders* (pp. 69–128). New York, NY: Cambridge University Press.
Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience, 11*, 351–360.
Ratner, N. B., & Sih, C. C. (1987). Effects of gradual increases in sentence length and complexity on children's dysfluency. *Journal of Speech and Hearing Disorders, 52*(3), 278–287.
Seeff-Gabriel, B., Chiat, S., & Dodd, B. (2010). Sentence imitation as a tool in identifying expressive morphosyntactic difficulties in children with severe speech difficulties. *International Journal of Language and Communication Disorders, 45*(6), 691–702.
Silverman, S. W., & Ratner, N. B. (1997). Syntactic complexity, fluency, and accuracy of sentence imitation in adolescents. *Journal of Speech, Language, and Hearing Research, 40*(1), 95–106.
Sperber, D. (2002). In defense of massive modularity. In E. Dupoux (Ed.), *Language, brain and cognitive development: Essays in honor of Jacques Mehler*. Mass: MIT Press.
Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development, 3*(1), 1–42.
Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science, 10*(2), 172–175.
Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes, 26*(7), 952–981.
Verhoeven, L., Steenge, J., van Weerdenburg, M., & van Balkom, H. (2011). Assessment of second language proficiency in bilingual children with specific language impairment: A clinical perspective. *Research in Developmental Disabilities, 32*(5), 1798–1807.
Willatts, P. (1999). Development of means–end behavior in young infants: Pulling a support to retrieve a distant object. *Developmental Psychology, 35*(3), 651–667.
Wolpert, D., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London B, 358*(1431), 593–602.
Woodyatt, G., & Ozanne, A. (1992). Communication abilities and Rett syndrome. *Journal of Autism and Developmental Disorders, 22*(2), 155–173.