rlp

The importance of studying prosody in the comprehension of spontaneous spoken discourse

La importancia de estudiar la prosodia en la comprensión del discurso hablado espontáneo

Recibido: Agosto de 2012
Revisado: Febrero de 2013
Aceptado: Marzo de 2013

Jazmín Cevasco
National Scientific and Technical Research Council, Argentina
University of Buenos Aires, Argentina

Fernando Marmolejo Ramos
University of Adelaide, Australia

Correspondence concerning this article should be addressed to: Jazmín Cevasco, Instituto de Investigaciones, Facultad de Psicología, UBA. Independencia 3065, Piso 3º Oficina 8 (1225), Buenos Aires. Tel: (05411) 4779-2055. Email: jazmincevasco@psi.uba.ar

Abstract

The study of the role of prosodic breaks and pitch accents in comprehension has usually focused on sentence processing, through the use of laboratory speech produced by both trained and untrained speakers. In comparison, little attention has been paid to their role in the comprehension and production of spontaneous discourse, or to the interplay between prosodic cues and pitch accents and the generation of inferences. This article describes studies which have focused on the effects of prosodic boundaries and pitch accents in sentence comprehension. Their results suggest that prosody has an early influence in the parsing of sentences as well as the processing of the information structure of a statement. It also presents a new model of spontaneous discourse comprehension that can accommodate paralinguistic factors, like pitch and prosody, and other communication channels and their relation to cognitive processes. Stemming from the model presented, future research directions are suggested as well as the importance of including spontaneous spoken discourse

Resumen

La investigación acerca del rol de los límites prosódicos y los acentos en la comprensión del lenguaje se ha centrado tradicionalmente en el estudio de la comprensión de oraciones, a través de la utilización del discurso producido por hablantes expertos o no expertos en el laboratorio. Comparativamente, se ha prestado menor atención al estudio de la comprensión del discurso oral espontáneo y del interjuego entre las claves prosódicas, los acentos, y la generación de inferencias. Esta revisión realiza un recorrido a través de investigaciones que han estudiado el efecto de los límites prosódicos y los acentos en la comprensión de oraciones. Los resultados de estos estudios sugieren que la prosodia tiene un efecto temprano en la segmentación de oraciones y en el procesamiento de la estructura de información de un enunciado. Se presenta también un modelo para la comprensión del discurso espontáneo que tiene en cuenta factores paralingüísticos, como prosodia y acento, y otros canales comunicativos y sus relaciones con procesos cognitivos. Partiendo del modelo propuesto,

materials and examining the role of prosodic cues and pitch accents in the establishment of connections among spoken statements is highlighted.

Key words: Prosodic Cues, Discourse Comprehension, Inference Generation, Spontaneous Spoken Discourse, CoSESM model.

se sugieren futuras líneas de investigación y se destaca la importancia de utilizar materiales de discurso espontáneo y de examinar el rol de las claves prosódicas y los acentos en el establecimiento de conexiones causales entre los enunciados.

Palabras clave: claves prosódicas, comprensión del discurso, generación de inferencias, discurso oral espontáneo, modelo CoSESM

Discourse is one of the behaviours which make us human (Graesser, Millis, & Zwaan, 1997). It can be defined as sequences of sentences that are coherently related (Halliday & Hasan, 1976). Discourse comprehension has been extensively investigated with respect to written discourse (e.g., Cain & Oakhill, 1999; van den Broek, 1990, 2010; van den Broek & Kremer, 2000). In comparison, little attention has been paid to the processing of spoken discourse (Carlson, 2009; Cevasco, 2008; Schafer, Carter, Clifton & Frazier, 1996; Speer & Blodgett, 2006). This gap is such important given that there are substantial differences between written and spoken discourse which could lead to differences in the cognitive processes involved in their comprehension (Cevasco & van den Broek, 2008; Lau & Ferreira, 2005; Speer & Blodgett, 2006). For instance, while written language allows for the readers to control the rate at which they acquire information (they can, for example, fixate on or skip words), spoken language has to be processed at the rate that it is produced. Also, while written discourse provides the reader with word boundaries clearly marked by spaces and sentences by periods, this is not the case with spoken language (Ferreira & Anes, 1994).

Spontaneous spoken discourse can be defined as spoken discourse produced in response to immediate situational demands (Ochs, 1979; Rico, Cohen, & Gil, 2006; Stubbs, 1983). Therefore, it is common for speakers to hesitate, correct errors publicly, repeat words and abandon phrases (Brennan & Shober, 2001; Fox Tree, 1995; Fox Tree & Schrock, 1999; Ochs, 1979). As a result, unplanned speech is frequently characterized by simple active sentences, juxtaposition of clauses with no explicit link at all, deletion of referents, etc. The comprehension of spontaneous discourse, thus, requires the ability to maintain continuity in speech and comprehension, respond immediately to unexpected utterances, repair speech errors and make changes of topic in real time (Stubbs, 1983). Yet, the comprehension of spontaneous speech is facilitated through information delivered paralinguistically and kinesically, employing means such as intonation, pitch, loudness, voice quality, speech rate (Cameron, 2001; Chafe, 1994;), gesturing (Hostetter & Alibali, 2008), and facial movements (Flecha-García, 2010). For instance, kinesic factors like beat gestures can alter which syntactic structure is assigned to ambiguous sentences (Holle et al., 2012) and modulate speech processing early in utterances (Wu & Coulson, 2010). Paralinguistic factors like prosody can favour spoken discourse by enriching the verbal message (Chafe, 1994; Gumperz, 1982). It is therefore assumed in this paper that spoken discourse accounts for communicative situations which have a dialogue- or conversational-like structure (see Flecha-García, 2010).

This article starts by reviewing several of the studies that have examined the role of prosody in sentence comprehension by normal adult speakers and listeners. These studies have focused on the effect of prosodic breaks in the parsing of sentences, and the influence of pitch accents on the processing of the information structure of a statement. In this article, prosody is defined as those acoustic-phonetic properties of a statement that are not the result of the choice of lexical items (Wagner & Watson, 2010), while prosodic breaks involve a pause, a boundary tone before this pause, and a lengthening of the word which precedes the pause (Kjelgaard & Speer, 1999). Pitch accents, on the other hand, are defined as changes, in fundamental frequency, that involve increased duration and intensity (Ladd, 1996). Thus, the goal of this paper is to highlight the need for more studies that use spontaneous discourse segments as materials to study prosody, so they focus on the discourse connections that the listener establishes among spoken statements in order to generate inferences.

A new model of spontaneous discourse comprehension that includes paralinguistic factors, like prosodic breaks and pitch accents, and their effect on cognitive processes is also presented. However, such a model also includes other communication channels and cognitive processes that are relevant to the construction of shared representations between conversational parties. Finally, several conclusions are presented regarding a potential research program that, framed within the proposed model, can account for the effects of prosody cues and pitch accents on the generation of inferences.

Prosodic boundaries and syntactic processing

Studies on the role of prosodic breaks have tended to examine how they affect the parsing of a sentence into its syntactic constituents. Allbritton, McKoon and Ratcliff (1996) investigated this by focusing listeners' ability to parse and disambiguate sentences produced by trained and untrained speakers. For example:

For our parties, we invite David and Pat or Bob, but not all three.

(A) *We invite David, and we also invite either Pat or Bob.*
(B) *We invite both David and Pat, or else we invite Bob.*

A and B are two possible interpretations of the ambiguous statement. Only about 2% of the pair of utterances produced by the untrained naïve speakers was rated as having disambiguating prosody. These sentences revealed some prosodic breaks, such as speakers' lengthening of phrases before the critical boundaries (e.g., lengthening of *'David'* for A and *'Pat'* for B). In the case of the pairs that had been rated as acceptable, a group of listeners was able to choose the appropriate meaning that the speaker meant. A second group, this time trained subjects (actors and broadcasters), was asked to read the passages without instructions to provide disambiguating cues. Ratings for appropriateness of prosody were again low. Those sentences which had been rated as having been produced with appropriate prosody also revealed some prosodic breaks at the critical boundaries. When a group of trained speakers informed about the purpose of the study was asked to produce the sentences, there was a greater number of acceptable statements. These statements also showed prosodic breaks at the critical boundaries. Listeners' meaning judgments on the acceptable pairs for the trained informed and naïve conditions indicated that they had been able to use the prosodic breaks provided; yet, the effect was larger for the sentences produced with explicit instructions. These results suggest that speakers are able to produce disambiguating prosodic breaks, if they are asked to do so.

In order to explore the interplay between prosody and context, Fox Tree and Meijer (2000) asked non-professional speakers to memorize and produce three sentence passages in which the middle sentence was ambiguous:

(1) *Toni went deep sea diving in the Pacific Ocean. She saw a man-eating fish. It scared her.*
(2) *Jenny went to a seafood restaurant. She saw a man eating fish. He seemed to like it.*

When listeners were presented with the ambiguous middle sentences in isolation, they could not accurately match them to their original contexts. The ambiguous sentences that received more correct responses were presented to a new group, who had to paraphrase them in writing. Results indicated that this time listeners were able to provide different paraphrases based on prosody. In order to examine whether people would still use prosodic breaks when they had contextual cues available, the ambiguous sentences were spliced into the alternative context. For example, the ambiguous sentence *'She saw a man-eating fish'* from context (1) was spliced into context (2) *'Jenny went to the Seafood restaurant…He seemed to like it.'* In the same way, the test sentence coming from context (2) was spliced into context (1) *'Toni went deep sea diving in the Pacific Ocean…It scared her.'* When prosody and context were incongruent, answers to content questions were scored relative to context. Therefore, it, appears that listeners can make use of prosody, but they rely on context when it is available.

In order to study prosody in a context that was closer to meaningful conversation, Schafer, Speer, Warren and White (2000) used a cooperative game task. This task involved the use of a predetermined set of utterances to negotiate moves around game boards. Some of them contained syntactic ambiguities:

(1) *When that moves the square will encounter a cookie.*
(2) *When that moves the square it should land in a good spot.*

The ambiguity could be resolved by prosodic or nonprosodic information (the word following *'square'*, the preceding discourse, etc). Phonetic analyses revealed that word durations increased as prosodic boundary strength

increased. That is, sentences such as (1) tended to show more prosodic boundaries following *moves* than *square*, and sentences such as (2) tended to show more prosodic boundaries at *square* rather than at *moves*. A group of listeners was presented with the ambiguous portions of the statements (i.e., *'When that moves the square…'*), and asked to choose between the two possible continuations. Results indicated high percentages of disambiguation. Contrary to what had been suggested by previous studies, naïve speakers were able to disambiguate syntactic ambiguity for listeners, even when it had already been disambiguated by context.

Snedeker and Trueswell (2003) studied the role of prosodic boundaries through a referential communication task in which a speaker instructed a listener to perform actions on the other side of a screen. During the target trials, the instructions could be ambiguous such as *"Tap the frog with the flower"*, in which *"with the flower"* could be understood as indicating what instrument to use to do the tapping or which frog to tap. The first group of speakers and listeners were provided with referential contexts that supported both meanings. Results indicated that prosodic breaks were a highly effective means of syntactic disambiguation. Phonological analyses revealed that when speakers were indicating which instrument to use to do the tapping, they tended to lengthen the word *'frog'* and pause between the *'frog'* and the with-phrase. When speakers were indicating which frog to tap, they tended to lengthen the word *'tap'* and pause after it. In order to test if speakers were providing prosodic breaks because they were aware of the alternative meanings, a new group was provided with a context that supported only the intended meaning. Results indicated that speakers were poor at producing prosodic breaks, and listeners were not able to distinguish between the two interpretations. These findings suggest that speakers' knowledge of the referential situation affects whether they use prosodic breaks to disambiguate utterances.

Kraljic and Brennan (2005) attributed the mixed evidence on speakers' ability to provide prosodic breaks to the different degrees of naturalness and interaction that the tasks used in previous studies had allowed. In order to provide speakers and listeners with the opportunity to interact freely, they used a task in which a director spontaneously instructed a matcher to move objects on a display. The matcher then provided feedback as to whether he or she had understood the reference. The critical target instructions could be ambiguous. For example, in the utterance *'Put the dog in the basket on the star'*, *'in the basket'* could be interpreted as specifying which dog, or as specifying where to put the dog. The object display could be ambiguous (supporting both interpretations) or unambiguous (supporting only one interpretation). In order to account for the possibility that speaker's awareness of the ambiguity makes him or her provide stronger prosodic breaks, members of each pair switched roles halfway through the task. Results indicated that directors did disambiguate the syntactic boundaries prosodically by lengthening the word before the prosodic boundary (*'dog'* or *'basket'*). Matchers were able to interpret the instructions, and directors produced disambiguating cues for both syntactically ambiguous and unambiguous utterances. Previous experience in the matcher role did not lead directors to mark boundaries more strongly. Whether the display supported only one or both interpretations did not matter either. In order to address the possibility that directors had been aware of the needs for disambiguation of their listeners, a new group was asked to produce instructions to move objects in the same ambiguous situation and in displays with unique objects (unambiguous). Results indicated that speakers again marked syntactic boundaries by lengthening them in the case that the display was ambiguous or unambiguous. Matchers were again able to follow the instructions. Given these findings, the authors concluded that prosodic marking of syntactic boundaries emerges from planning and articulating syntactic structure, and not from being aware of an addressee's need for disambiguation. Other studies which have examined prosodic breaks have also found a role for them in syntactic disambiguation (Bögels, Schriefers, Vonk, Chwilla, & Kerkhofs, 2010; Dede, 2010; Kerkhofs, Vonk, Schriefers, & Chwilla, 2008; Roll, Horne, & Lindgren, 2011).

To summarize, studies on the role of prosody in the disambiguation of spoken utterances suggest that it can allow listeners to decide between two alternative interpretations of an ambiguous utterance. They also suggest that results obtained with researcher constructed speech or trained speakers might not represent what participants do when they have a clear communicative goal and a co-present addressee. When participants do not have a clear goal, they appear not to mark prosodic boundaries to the same extent as when they do. In consequence, it appears that researchers need to provide participants with tasks

that allow them to spontaneously produce utterances in order to approximate how they would communicate outside the laboratory.

Pitch accents

As mentioned before, studies on the effects of pitch accents have tended to examine their influence on the processing of the information structure of statements. Accents are expected to increase attention to particular words or syllables, and to signal which information is new and which one has already been mentioned or given (Birch & Clifton, 1995, 2002; Bögels, Schriefers, Vonk, Chwilla, 2011; Cutler, Dahan & van Donselaar, 1997; Ladd, 2008; Fraundorf, Watson, & Benjamin, 2012; Speer & Blodgett, 2006; Watson, 2008).

Birch and Clifton (1995) examined the effects of pitch accents on sentence comprehension by asking listeners to rate question-answer utterances for appropriateness of intonation. For example:

Isn't Kerry good at math?

a. *Yes, she TEACHES math.*
b. *She teaches MATH.*

When answers accented the new information in the sentences (*'teaches'* in *a*), listeners provided more *"makes sense"* judgments than when answers accented information that had already been mentioned (*'math'* in *b*). These results suggest that listeners are sensitive to accent placement, and that they consider that *given* information should not be highlighted.

Dahan, Tanenhaus and Chambers (2002) examined the role of pitch accents through the use of an eye-tracking technique. Participants were presented with visual displays which included a *candle*, a *candy* and a *triangle*. Their eye movements were tracked as they listened to instructions to move those objects:

"Put the candle/ candy below the triangle"
"Now put the CANdy/CANdle…"

When the second instruction accented *CAN*, participants' looks to the new, non-mentioned item (*'candy'* if the first instruction had mentioned *'candle'*, and *'candle'* if the first instruction had referred to *'candy'*) increased, even before the word was fully articulated. In other words, listeners seem to be able to make early use of prosody in order to predict upcoming referents.

Ito and Speer (2008) also studied the effect of pitch accents through the use of an eye-tracking technique. In their study, listeners had to follow pre-recorded instructions to decorate a Christmas tree. Ornaments included: *candies, stockings, balls, angels, bells,* etc. Pitch accents in the instructions varied. They could be assigned to words that conveyed contrastive information, such as:

"First, hang the blue ball."
"Next, hang the GREEN ball."

In this case, only the accented colour of the object contrasted with the previous instruction. Accents could also be assigned to words that did not convey contrastive information:

"First, hang the blue ball"
"Next, hang the green BALL"

Results indicated that looks occurred earlier and more often to the target object when accents were placed on contrastive words. These findings provide converging evidence that pitch accents can help listeners predict upcoming words, and can contribute to sentence comprehension.

Weber, Braun, and Crocker (2006) also found an effect of pitch accents on the identification of references. They asked participants to follow two consecutive instructions to click on objects on a computer display while they monitored eye movements. The first instruction always introduced one member of a contrast pair (*purple scissors*). The second instruction referred to either the other member of the contrast pair (*red scissors*), or to an object differing in form but not colour from the other member of the pair (*red vase*). In half of the trials, the adjective was unaccented and the accent was on the noun (*red SCISSORS*). In the other half, the adjective carried a contrastive accent (*RED scissors*). Since only the red scissors contrasted in colour with another displayed object (the *purple scissors*), the accent was a cue for the upcoming referent. Results indicated there were earlier looks toward the correct object when the adjective was accented compared with when it was not. Once more, these findings suggest that listeners rapidly exploit prosodic information to interpret referential expressions.

In conclusion, studies on pitch accents repeatedly suggest that they have a role in listeners' identification of references (see also Arnold, 2008; Heim & Alter, 2006;

Toepel, Pannekamp, & Alter, 2007). They seem to facilitate the processing of new information even before words have been completed.

Towards a comprehensive model of spontaneous discourse comprehension

The studies that we have considered so far allow us to reach some conclusions. Prosody seems to have a role in the processing of spoken sentences. It allows listeners to parse them in order to decide between alternative syntactic interpretations, and it prompts them to focus on specific words that should be highlighted. Still, these studies do not allow us to draw conclusions about the establishment of discourse connections beyond the identification of references.

Discourse connections which need to be established for comprehension include causal relations (van den Broek, 1990, 1994; Zwaan & Rapp, 2006). The establishment of these relations requires for the comprehender to generate *connective, reinstatement, elaborative* and *predictive* inferences (van den Broek, 1990; 1994). Connective inferences are made when the reader identifies a causal relation between the statement that he or she is reading or hearing, and information that remained activate in working memory after processing the immediately previous statement. For example, the comprehender needs to establish a causal connection if he or she hears or reads: *"Murray poured water on the bonfire. The fire went out"* (Singer & Halldorson, 1996). Reinstatement inferences are made when the reader reactivates information presented previously (before the immediately previous statement), in order to maintain sufficient causal justification for the statement that he or she is processing. For example, in *"Murray poured water on the bonfire. He heard his cell phone ringing. The fire went out"* the comprehender needs to reactivate that *"Murray poured water on the bonfire"* to understand why *"the fire went out"*. Causal elaborative inferences draw on the readers' background knowledge to identify a causal antecedent that is not explicitly mentioned or to anticipate events. Among them, there are emotional inferences and predictive inferences. Emotional inferences involve the activation of knowledge about fictional characters' emotional states, as a consequence of story events (Gernsbacher, Goldsmith, & Robertson, 1992; Molinari, Barreyro, Cevasco, & van den Broek, 2011). For example, if we read or hear that *"While shooting a film, the actor accidentally fell out the 14th floor window. His friends went to the funeral"*, we will probably infer that his friends were sad. Predictive inferences draw on the readers' background knowledge to anticipate upcoming statements (see Marmolejo-Ramos, Elosúa de Juan, Gygax, Madden, & Mosquera, 2009). For example, if we read or hear that *"While shooting a film, the actor accidentally fell out the 14th floor window"* we probably will infer that he died.

Studies that which have examined the role of prosody so far have not considered the generation of these inferences, or how prosodic breaks and pitch accents might interact with them. It would be therefore interesting to consider whether accenting particular words in a statement might facilitate or hinder the establishment of causal connections and emotional inferences, and whether prosodic breaks would interact with them in the disambiguation of sentences. That is, given that appropriate prosodic breaks disambiguate ambiguous statements, they should allow the comprehender to move on to establish connections with other statements, and generate inferences. Inappropriate prosodic breaks might make it more difficult to parse a spoken statement, and thus not allow the listener to move on to establish causal connections. For example, if we hear *"While shooting a film, the actor accidentally fell out the 14th floor # window"*, with a prosodic break after *"floor"* this might make us slowdown in order to process why that break has been placed there, and prevent us from being able to anticipate upcoming events so easily. Also, inappropriate pitch accents could make causal inferences more difficult to generate. For instance, if we hear *"Murray poured water on the bonfire. THE fire went out"*, it might slow us down so as to process why the accent is placed on *"the"* instead of on the new information. Consequently, we would be unable to move on to generate the connective inference.

Thus far, the focus has been on the role of pitch accents and prosodic breaks in the comprehension of sentences. Thus, current research agenda should account for how paralinguistic factors from one speaker affect the generation of specific inferences in the other. However, we believe that this issue is just one of the potential set of researchable interactions that are part of a more comprehensive model of spontaneous discourse comprehension and production.

A brief prolegomenon of a model of spontaneous discourse comprehension which has as its ultimate goal the construction of shared embodied situation models is

presented: the CoSESM model. The CoSESM model is generated by merging current theories from the embodied cognition framework in relation to language comprehension with classic theories in linguistics. Specifically, current embodiment theories argue that sensorimotor experience is crucial in the comprehension of language (see Glenberg & Gallese, 2012), and that memory, inference generation, and simulation systems need to interact for the representation of real and vicarious bodily, affective, and cognitive states (Marmolejo-Ramos, 2007; Mishra & Marmolejo-Ramos, 2010; Marmolejo-Ramos & Cevasco, 2012). Current embodiment theories also acknowledge that comprehension of language requires the use of sensorimotor systems in the brain for the generation of embodied situation models (see Barsalou, 2008; Glenberg, 1999). However, current theorisation in the embodiment of language has the problem of defining language as strings of spoken or written words, sentences, or textoids disconnected from their kinesic and paralinguistic co-systems. In order to address this problem, the CoSESM model adopts the Basic Triple Structure (Poyatos, 1984) in which language, kinesics, and paralanguage are integrated in a unitary communicative system. Under the Basic Triple Structure, kinesics is defined as "gestures, manners, and postures", and paralanguage as "voice modifications and independent word-like utterances coded in the vocal/narial-auditory channel" (Poyatos, 1984, p. 307). More importantly, the CoSESM model includes the conversational context as a factor that has an effect on the communicative act (see Poyatos, 1984) (see Figure 1).

The CoSESM model relies on three basic assumptions: parity of the representations, effects of alignments at different levels, and influence of kinesics and paralanguage in the comprehension and production of language[1]. The CoSESM model shares the parity of representations principle of current dialogue models, like the interactive-alignment model (IA model) (Menenti, Pickering, & Garrod, 2012), in that the construction of a *SESM* requires that it be "coded in the same form irrespective of whether a person is speaking or listening" (Menenti et al., 2012, p. 3); thus the word "shared" in the current model. The *SESM* is a situation model in that it represents the scenarios, entities, characters, and actions described in the linguistic stream. It is embodied in that it

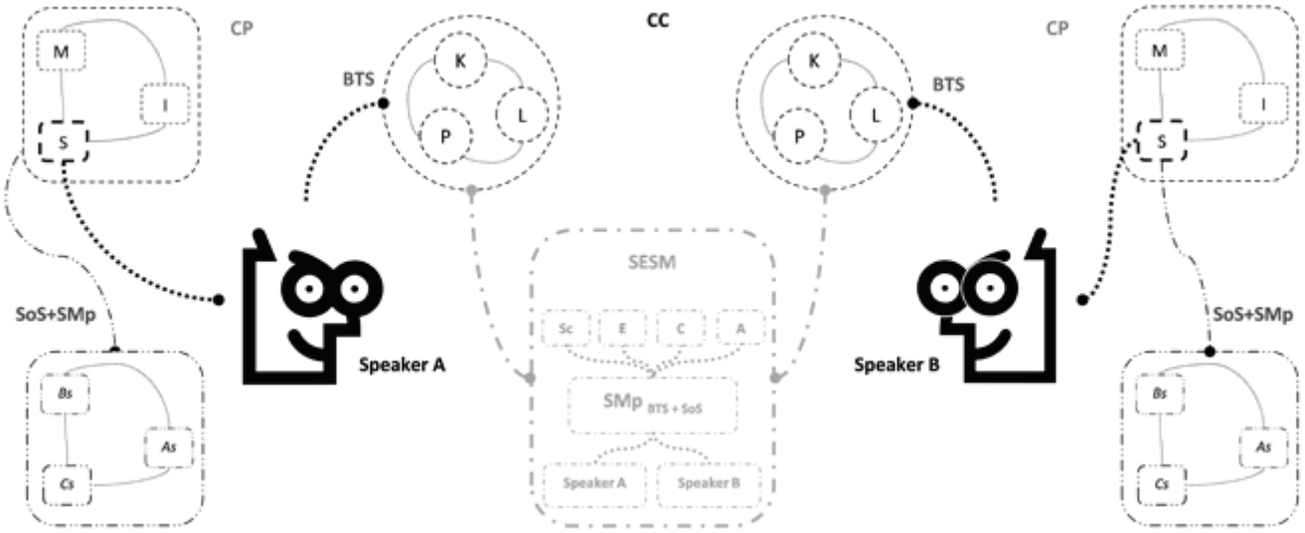

Figure 1. Model of the construction of shared embodied situation models during the comprehension and production of spontaneous discourse. CP = cognitive processes (M = memory, I = inferences, and S = simulation), SoS+SMp = simulation of states and associated sensorimotor properties (Bs = bodily states, As = affective states, and Cs = cognitive states), BTS = Basic Triple Structure (K = kinesics, L = language, and P = paralanguage), SESM = shared embodied situation model (Sc = scenario, E = entities, C= characters, A= actions, SMp BTS + SoS = sensorimotor properties associated to the BTS and the SoS), and CC = conversational context.

1. The proposed model assumes that discourse comprehension and production are one integrated process in that similar brain areas are used (Price, 2010), thus rendering comprehension and production into reverse processes of each other (see Glenberg & Gallese, 2012; Hostetter & Alibali, 2008). However, for simplicity purposes, the term comprehension will be mostly used.

also contains sensorimotor properties associated with the scenarios, entities, characters, and actions described and sensorimotor properties of the conversationalists parties (e.g., possible emotional and cognitive states, sensorimotor properties of the kinesic and paralinguistic channels, etc.).

The CoSESM model, however, assumes a modified version of the alignment at different levels principle. Whereas the IA model assumes that "alignment at one level of representation affects alignment at another" (Menenti et al., 2012, p. 3), the CoSESM assumes a more relaxed and flexible stance in that mis-alignment at one level of representation does not necessarily affect the alignment at another. For instance, a mis-alignment at the phonetic level does not necessarily affect the alignment at the pragmatic level and vice versa. The common situation of non-native speakers of English talking to native speakers of English presents an example of the mis-alignment/alignment principle. There is evidence showing that the comprehension of sentences is relatively high despite them being spoken with an accent (e.g., Adank & Janse, 2010), and even over a noisy background (e.g., Adank, Davis, & Hagoort, 2012). However, comprehension of spoken language not only relies on the linguistic channel alone but also on the other two co-systems (*K* and *P*) in order to compensate for deficiencies in this channel. In the specific case of the kinesic channel, it has been shown that gestures and speech mutually and obligatorily interact in order to enhance language comprehension (Kelly, Özyürek, & Maris, 2010). It has even been indicated that gestures influence the processing of figurative language in non-native speakers (Ibáñez et al., 2010). The inverse example is one in which both parties cannot reach a *SESM* despite not having problems at lower representational levels. A simple case would be that in which *Speaker A* explains a process to *Speaker B* and, even though both speakers share the same L1, there is a conceptual misunderstanding between them. In this case, although both parties had an alignment at the phonetic level, there is a mis-alignment at the highest level of representation, i.e., the *SESM*. In this conversational situation, assuming that both speakers were skilfully using their *BTS*s, the reason why a *SESM* cannot be constructed can be traced back to the *CP* system, e.g., differences in episodic memory capacity. For instance, there is evidence showing that working memory capacity determines the comprehension of spoken discourse, particularly when the spoken stream is distorted (Sörqvist & Rönnberg, 2012). This situation is exemplified when two speakers are holding a conversation in a "cocktail party" context, but one of them has low working memory capacity. In this case, it is likely that a conceptual mis-alignment might occur since one of the parties cannot hold up information for long periods of time when there is high background noise, therefore affecting his/her inferencing and simulation processes.

A third prediction is that kinesic and paralinguistic systems influence the comprehension and modulate the production of language. That *K* and *P* factors play a key role in language production and comprehension is an assumption that has not been explicitly acknowledged by current models, and it is therefore a unique characteristic of the present model. Gestures, prosody, and pitch are part of the *K* and *P* toolkit used during language comprehension/production. A recent framework called Gestures as Simulated Action (GSA_f) proposes that gestures physically represent the content of the situation models being constructed during language production/comprehension (see Hostetter & Alibali, 2008; see also Glenberg & Gallese, 2012). That is, gestures are the manifestation of the simulations occurring during the generation of sensorimotor properties associated with what is being conversed. Some of the predictions of the GSA_f have been tested to show that the simulation of motor properties is gesticulated using the point of view of the character in the narration, while the simulation of visual imagery is gesticulated using the point of view of an observer (e.g., Parrill, 2010). Although, to the best of our knowledge, the effects of the *P* and *K* systems have not been tested in the context of spontaneous discourse comprehension/production, there is recent research that throws light in this direction. Flecha-García (2010) found that eye brow raising (a *K* component) was more common during the utterance of instructions than requests and it was aligned with pitch accents (a P component). This finding can be interpreted as evidence regarding interactions among components within the *BTS*, i.e., interactions among *K*, *P*, and *L*. However, there is no research studying how pitch accents and prosody breaks, when coupled with body movements, affect the generation of inferences and the retention of information[2].

2. Another *P* device is onomatopoeia (a word that mimics the source of the sound it describes). To our knowledge, there is no study devised to understand the relationship between onomatopoeia and the generation of inferences. More importantly, given that onomatopoeia has a closer link to the concepts it refers to in that it enhances the representation of concepts' sensorimotor properties (particularly ideophones), it would be interesting to investigate its role in the construction of *SESM*s.

Discussion

The aim of this article was to address how the study of the role of prosodic breaks and pitch accents in sentence comprehension has been approached so far, in order to suggest what the consideration of the key results of these investigations can contribute to the exploration of their role in the establishment of discourse connections that are necessary for the construction of a coherent discourse representation. Also, it was argued that a comprehensive study of spontaneous discourse comprehension should account for the modulatory effects that *P* factors, namely pitch accents and prosodic breaks, have on cognitive processes, particularly on the generation of inferences.

Studies on prosody and sentence comprehension have suggested that prosodic breaks can allow a listener to decide between alternative interpretations of a statement. They also suggest that pitch accents can facilitate the processing of new information. The consideration of these results can be useful to think about studies that explore what the role of these prosodic cues can be, for example, in the generation of causal inferences. That is, even though our understanding of the role of prosody in sentence comprehension has been enriched by the existing studies, it still has not been established what the role of these breaks and accents is in the processing of discourse connections, such as causal links among statements, and even emotional inferences. This gap is important, given that statements are not presented in isolation in natural conversation, but rather in the context of other statements with which they are causally connected. In other words, one of the aims of this article was to highlight that if we want to understand how speakers and listeners behave in everyday language comprehension, we need to take discourse connections such as these into account (Speer & Blodgett, 2006).

In addition, as suggested by the proposed CoSESM model, paralinguistic factors such as pitch accents and prosodic breaks can be coupled with kinesic factors in order to effectively communicate the intended message. Pitch accents might reinforce words' salience by their association to kinesic tools like eye brows (Flecha-García, 2010), and prosody can do the same by being linked to gestures that have a pointing-like function (Lœvenbruck, Dohen, & Vilain, 2009). This type of evidence lends support to the model proposed herein in that paralinguistic and kinesic factors interact for the effective delivery of linguistic streams. For the particular case of inference generation, it would be important to unveil how vocal pitch pairs with body pitch in order to augment or reduce the amount of inferences generated. For instance, it has been found that when there is a gestural apex there is also a pitch accent (but not the other way around) and that studies in laboratory and spontaneous speech have reached the same conclusion (Jannedy & Mendoza-Denton, 2005; see also McClave, 1998). More recently, in a spontaneous speech study, it has been found that gesture phrases align with intermediate phrases regarding their timing, structure, and pragmatic meaning (Loehr, 2012). It could thus be entertained that if such couplings are attenuated or reduced, then the generation of inferences could be affected. Although this idea remains merely speculative and needs to be addressed empirically, some tentative research scenarios are worth considering.

Prosody and pitch are paralinguistic tools that are necessarily connected to the *L* and *K* systems in order to support effective communication. As the evidence reported above suggests, *P* and *K* factors work in conjunction during the production of language. Furthermore, it is tenable to entertain that the linkage between *P* and *K* factors not only facilitates language production but also language comprehension in that both production and comprehension processes might share similar brain areas (see footnote 1). Given the linkage among the *K*, *P*, and *L* systems, it is likely that when one of those systems is compromised, the other two systems, or at least one of them, should compensate for the deficiency in the affected system. Thus, it could be considered that the construction of emotional inferences in the listener could be affected by imbalances within the *BTS* in the speaker and since inferences are needed in the simulation of language, the final *SESM* can be affected too. Moreover, the conversational context (*CC* in the model) can also influence the level of inferencing made by the listener, thus ultimately affecting the *SESM*.

The consideration of the presented model can help us think about some possible research directions, which can contribute to the exploration of language comprehension For example, investigating the effects that P components of the *BTS* can have on the elaboration of emotional inferences during spoken discourse is a new research question. Specifically, different levels of mastery in *L*, *K*, and, particularly, *P* systems in the speaker might lead to different levels of emotional inferencing in the listener. In turn, since inferences combine with memory systems in order to simulate cognitive,

emotional, and bodily states (Marmolejo-Ramos, 2007), it is likely that the final embodied situation model reflects the effects of *P* systems on emotional inferencing.

Another possible future direction involves the use spontaneous discourse segments as materials in new studies. As previously suggested, spontaneous discourse tends to include disfluencies (such as repairs, repetitions, filled pauses, etc), which have not tended to be part of the materials used to study spoken discourse. On the contrary, these studies have used speech produced in the laboratory by trained and untrained speakers, with or without a communicative goal. A problem with these statements is that they may not reflect those which speakers would produce outside the laboratory, or the processes in which listeners may engage to comprehend them (Cutler et al, 1997; Speer & Blodgett, 2006). Indeed, it has been suggested that information encoded in disfluencies such as repairs can facilitate the processing of syntactic ambiguities (Lau & Ferreira, 2005).

On the other hand, it would be interesting to take into account that most studies on prosody have focused on the comprehension of the literal meanings of statements. The processing of prosodic breaks and pitch accents in the comprehension of ironic statements has not received the same attention (Carlson, 2009). Neither has the processing of prosody in languages other than English (Speer & Blodgett, 2006).

In conclusion, it seems that the study of the role of prosody in language comprehension is an open topic which can contribute to our understanding of the unique aspects of spoken discourse and the cognitive processes involved in its comprehension. The CoSESM model depicts a comprehensive agenda for the study of this topic. It is expected that future research using this model will ultimately provide empirical evidence as to how the model operates and to the strength of relationship among its components. For instance, methods like structural equation modelling and multilevel linear modelling would be extremely helpful in this front. This model, although initially conceived for the study of spontaneous discourse processes, can be adapted to suit laboratory conditions. While laboratory-set studies have provided useful insights as to the mechanisms behind discourse processing (see Xu, 2010), spontaneous discourse studies require engineering clever methodological and analytical approaches in order to corroborate and extend laboratory-set claims.

References

Adank, P., & Janse, E. (2010). Comprehension of a novel accent by young and older listeners. *Psychology and Aging, 25*(3), 736-740.

Adank, P., Davis, M., & Hagoort, P. (2012). Neural dissociation in processing noise and accent in spoken language comprehension. *Neuropsychologia*, *50*(1), 77-84.

Allbritton, D., McKoon, G., & Ratcliff, R. (1996). Reliability of prosodic breaks for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22*(3), 714-735.

Arnold, J. E. (2008). Reference production: production-internal and addressee-oriented processes. *Language and Cognitive Processes, 23*(4), 495-527.

Barsalou, L.W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-645.

Birch, S., & Clifton, Jr., C. (1995). Focus, accent, and argument structure: Effects on language comprehension. *Language and Speech, 38*(4), 365-391.

Bögels, S., Schriefers H., Vonk W., & Chwilla D. J. (2011) Prosody and sentence processing: The role of prosodic breaks investigated by ERPs. *Language and Linguistics Compass, 5*(7), 424-440.

Bögels, S., Schriefers, H., Vonk, W., Chwilla, D. J., & Kerkhofs, R. (2010). The interplay between prosody and syntax in sentence processing: The case of subject- and object-control verbs. *Journal of Cognitive Neuroscience, 22*(5), 1036-1053.

Brennan, S., & Schober, M. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language, 44*(2), 274-296.

Cain, K., & Oakhill, J.V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing: An Interdisciplinary Journal, 11*(5-6), 489-503.

Cameron, D. (2001). *Working with Spoken Discourse*. Thousand Oaks, CA: Sage Publications.

Carlson, K. (2009). How prosody influences sentence comprehension. *Language and Linguistics Compass, 3*(5), 1188-1200.

Cevasco, J. (2008). La importancia del estudio de la comprensión de discurso a través de la utilización de materiales de discurso natural. *Investigaciones en Psicología, 13*, 45-60.

Cevasco, J., & van den Broek, P. (2008). The importance of causal connections in the comprehension of spontaneous spoken discourse. *Psicothema, 20*(4), 801-806.

Chafe, W. (1994). *Discourse, Consciousness and Time.* Chicago: University of Chicago Press.

Cutler, A., Dahan, D., & Donselaar, W. V. (1997). Prosody in comprehension of spoken language: A literature review. *Language and Speech, 40*(2), 141-201.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language, 47*(2), 292-314.

Dede, G. (2010). Utilization of prosodic information in syntactic ambiguity resolution. *Journal of Psycholinguistic Research, 39*(4), 345-374.

Ferreira, F., & Anes, M. (1994). Why study spoken language processing? In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics.* San Diego, CA: Academic Press.

Flecha-García, M. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication, 52*(6), 542-554.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language, 34*(6), 709-738.

Fox Tree, J. E., & Meijer, P. J. A. (2000). Untrained speakers' use of prosody in syntactic disambiguation and listeners interpretations. *Psychological Research, 63*(1), 1-13

Fox Tree, J. E., & Schrock, J. C. (1999). Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language, 40*(2), 280-295.

Gernsbacher, M. A., Goldsmith, H. H., & Robertson, R. R. W. (1992). Do readers mentally represent characters' emotional states? *Cognition and Emotion, 6*(2), 89-111.

Glenberg, A. (1999). Why mental models must be embodied. In: G. Rickheit, & C. Habel (Eds.), *Mental Models in discourse processing and reasoning* (pp. 77-90). Amsterdam, Netherlands: North-Holland/Elsevier Science Publishers.

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

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48*, 163-189.

Gumperz, J. J. (1982). *Discourse strategies.* Cambridge: Cambridge UniversityPress.

Halliday, M. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

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

Holle, H., Obermeier, C., Schmidt-Kassow, M., Friedereci, A., Ward, J., & Gunter, T. C. (2012). Gesture facilitates the semantic analysis of speech. *Frontiers in Psychology, 3*(74), doi: 10.3389/ fpsyg.2012.00074

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: gestures as simulated action. *Psychonomic Bulleting & Review, 15*(3), 495-514.

Ibáñez, A., Manes, F., Escobar, J., Trujillo, N., Andreucci, P., & Hurtado, E. (2010). Gesture influences the processing of figurative language in non-native speakers: ERP evidence. *Neuroscience Letters, 471*(1), 48-52.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: eye movements during instructed visual search. *Journal of Memory and Language, 58*(2), 541-573.

Jannedy, S., & Mendoza-Denton, N. (2005). Structuring information through gesture and intonation. In S. Ishihara, M. Schmitz, & A. Schwarz (Eds.), *Interdisciplinary studies on information structure 03* (pp. 199-244). Universitätsverlag Potsdam.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin. Speech and gesture mutually interact to enhance comprehension. *Psychological Science, 21*(2), 260-267.

Kerkhofs, R., Vonk, W., Schriefers, H., & Chwilla, D. J. (2008). Sentence processing in the visual and auditory modality: Do comma and prosodic break have parallel functions? *Brain Research, 1224*, 102-118.

Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language, 40*, 153-194.

Kraljic, T., & Brennan, S. (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology, 50*(2), 194-231

Ladd, D. R. 1996. *Intonational phonology.* Cambridge: Cambridge University Press.

Lau, E. F., & Ferreira, F. (2005). Lingering effects of disfluent material on the comprehension of garden path sentences. *Language and Cognitive Processes, 20*(5), 633-666.

Loehr, D. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology, 3*(1), 71-89.

Lœvenbruck, H., Dohen, M., & Vilain, C. (2009). Pointing is 'special'. In S. Fuchs, H. Lœvenbruck, D. Pape, & P. Perrier (Eds.), *Some aspects of speech and the brain* (pp. 211-258). Peter Lang.

Marmolejo-Ramos, F. (2007). Nuevos avances en el estudio científico de la comprensión de textos. *Universitas Psychologica, 6*(2), 331-343.

Marmolejo-Ramos, F., & Cevasco, J. (2012). *Text comprehension as a problema solving situation.* Manuscript under review.

Marmolejo-Ramos, F., Elosúa de Juan, M. R., Gygax, P., Madden, C., & Mosquera, S. (2009). Reading between the lines: The activation of embodied background knowledge during text comprehension. *Pragmatics & Cognition, 17*(1), 77-107.

McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic Research, 27*(1), 69-89.

Menenti, L., Pickering, M. J., & Garrod, S. C. (2012). Toward a neural basis of interactive alignment in conversation. *Frontiers in Human Neuroscience, 6* (185). DOI: 10.3389/fnhum.2012.00185

Mishra, R. K., & Marmolejo-Ramos, F. (2010). On the mental representations originating during the interaction between language and vision. *Cognitive Processing, 11*(4), 295-305.

Molinari, C., Barreyro, J. P., Cevasco, J., & van den Broek, P. W. (2011). Generation of emotional inferences during text comprehension: Behavioral data and implementation through the landscape model. *Escritos de Psicologia, 4*(1), 9-17.

Ochs, E. (1979). Planned and unplanned discourse. In T. Givon (Ed.), *Discourse and syntax. Syntax and Semantics Series, vol. 12* (pp. 52-80). New York: Academic Press.

Parrill, F. (2010). Viewpoint in speech-gesture integration: linguistic structure, discourse structure, and event structure. *Language and Cognitive Processes, 25*(5), 650-668.

Poyatos, F. (1984). The multichannel reality of discourse: language-paralanguage-kinesics and the totality of communicative systems. *Language Sciences, 6*(2), 307-337.

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences, 1191*, 62-88.

Rico, R., Cohen, S., & Gil, F. (2006). Efectos de la interdependencia de tarea y la sincronía en las tecnologías de comunicación sobre el rendimiento de los equipos virtuales de trabajo. *Psicothema, 18*(4), 743-749.

Roll, M., Horne, M., & Lindgren, M. (2011). Activating without inhibiting: Left-edge boundary tones and syntactic processing. *Journal of Cognitive Neuroscience, 23*(5), 1170-1179.

Schafer, A., Carter, J., Clifton Jr, C., & Frazier, L. (1996). Focus in relative clause construal. *Language and Cognitive Processes, 11*(1/2), 135-63.

Schafer, A. J., Speer, S., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research, 29*(2), 169-182.

Singer, M., & Halldorson, M. (1996). Constructing and validating motive bridging inferences. *Cognitive Psychology, 30*(1), 1-38.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language, 48*(1), 103-130.

Sörqvist, P., & Rönnberg, J. (2012). Episodic long-term memory of spoken discourse masked by speech: what is the role for working memory capacity? *Journal of Speech, Language, and Hearing Research, 55*(1), 210-218.

Speer, S., & Blodgett, A. (2006). Prosody. In M. Traxler & M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics, 2nd Ed* (pp. 505-37). San Diego, CA: Academic Press.

Stubbs, M. (1983). *Discourse Analysis.* Chicago: University of Chicago Press.

Toepel, U., Pannekamp, A., & Alter, K. (2007). Catching the news: processing strategies in listening to dialogs

as measured by ERPs. *Behavioral and Brain Functions, 3*(53), doi:10.1186/1744-9081-3-53.

van den Broek, P. (1990). The causal inference maker: towards a process of inference generation in text comprehension. In D.A. Balota, G.B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension Processes in Reading* (pp. 423-445). Hillsdale, NY: Erlbaum.

van den Broek, P. (1994). Comprehension and memory of narrative texts: Inferences and coherence. In M.A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 539-588). San Diego, CA: Academic Press.

van den Broek, P. (2010). Using texts in science education: cognitive processes and knowledge representation. *Science, 328*(5977), 453-456.

van den Broek, P., & Kremer, K. (2000). The mind in action: What it means to comprehend. In B. Taylor, P. van den Broek, & M. Graves (Eds.), *Reading for meaning* (pp. 1-31). New York: Teacher's College Press.

Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and cognitive processes, 25*(7-9), 905-945.

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referents in time: eye-tracking evidence for the role of contrastive accents. *Language and Speech, 49*(3), 367-92.

Wu, Y. C., & Coulson, S. (2010). Gestures modulate speech processing early in utterances. *Neuroreport, 21*(7), 522-526.

Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics, 38*(3), 329-336.

Zwaan, R. A., & Rapp, D. N. (2006). Discourse comprehension. In M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics, 2nd Ed* (pp. 725-764). San Diego, CA: Academic Press.