Language comprehension: Insights from research on spoken language

Delphine Dahan

Department of Psychology

University of Pennsylvania

and

Fernanda Ferreira

Department of Psychology

University of California, Davis
Language is a system that gives physical form to meaning by combining discrete elements. In order to use language, people must learn the specific ways in which their language expresses meaning, and recruit this knowledge when expressing the meanings they wish to convey or when retrieving those of others. This chapter discusses the architecture and the working of the machinery that enables listeners to derive the meaning of the utterances they hear.

Mainly for practical reasons, research in psycholinguistics has tended to focus on written language. But with the advent of speech technology and the increased availability of tools to edit and analyze speech, as well as to control the presentation of spoken stimuli, work in the area of spoken-language comprehension has taken off. The goal of this chapter is to review how this research has advanced our knowledge of how people comprehend the utterances they are exposed to. Indeed, speech differs from print in a few but significant ways. Some of these differences have brought unique insight to existing debates, while others have led to the discovery of phenomena that pertain to speech specifically.

1. **Prosody and the architecture of the language processor**

   The comprehension of a sentence requires the listener to derive several linguistic structures. Each one operates over specific categories, is guided by distinct rules or principles constraining how categories combine, and makes a unique contribution to the meaning of the sentence. In the context of speech comprehension, the process has been described as a “conversion of acoustic input into meaning”, and is traditionally described as involving separate (i.e., modular) and hierarchically organized stages (see Cutler & Clifton, 1999; Hagoort & Poeppel, 2013, for reviews largely rooted in the traditional approach). According to such a view,
each stage takes as input the outcome of the one that precedes it in order to generate the relevant structure. The data reduction that this architecture entails optimizes the efficiency of the computations taking place at each stage. Before discussing this approach and its alternatives, we briefly describe the nature of the structures generated at each of the stages.

The putative first step takes the acoustic signal as input and infers the phonological structure of the utterance. One aspect of the structure consists of a sequence of sound categories or phonemes. These categories capture the functionally relevant contrasts in the sound system of the language and generalize across contexts, speakers, or speaking genres. In addition to this linearly organized phonemic string, the phonological structure represents the hierarchically organized phrasing and prominence structure of these phonemic elements organized into a prosodic structure.

The sequence of phonemes, in conjunction with its prosodic structure, may then be segmented out to retrieve the morpho-syntactic elements and their relationships. Retrieval of the syntactic properties associated with each of these elements, along with the order in which they appear, gives rise to a structure that captures the groupings of the elements into larger constituents and the relationships that these constituents entertain with one another in a hierarchically organized fashion, i.e., in the syntactic structure of the sentence. Closely related but not identical to the syntactic structure, the semantic structure marks the meaning-based relationships between the morpho-syntactic elements, based on both their syntactic relationships and the semantic properties associated with each element. Thus, the semantic structure relies on the constituency captured in the syntactic structure.

In this traditional model, the meaning of the sentence, represented in its semantic structure, is then integrated into the discourse and its broader context to yield its information
structure. The information structure marks which parts of the sentence meaning are highlighted and brought to the foreground within the broader discourse, and which ones are known or presupposed and left to the background. Information structure also captures some attitudes of the speaker attributed to himself or herself or to the hearer regarding parts of the sentence’s meaning. For instance, Steedman (2014) points out that the sentence ‘You put my trousers in the microwave’ may assert the proposition, in effect marking it as part of the knowledge common to both interlocutors, but may also be used to convey disbelief by marking the proposition as not becoming part of that common knowledge (see also Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014).

The traditional, modular and serial, architecture of the language comprehension system just described has been under attack for several decades (e.g., Rumelhart, McClelland, and the PDP Research Group, 1986) and we refer the reader to other sources for a report on the debate (e.g., Joanisse & McClelland, 2015). Our goal here is to describe the theoretical contributions that research on spoken language comprehension, and prosody in particular, has made to the debate.

Prosodic representations capture the utterance’s phrasing and prominence. For the most part, research on written language has not been concerned with prosody under the assumption that it plays little or no role during reading. Although there is now evidence that people spontaneously generate a prosodic structure to sentences that they read (Breen, 2014; Jun, 2010; Stolterfoht, Friederici, Alter, & Steube, 2007; Swets, Desmet, Hambrick, & Ferreira, 2007), it remains the case that an utterance makes its prosody immediately available to listeners, whereas a written sentence does not. Access to prosodic information is significant because the prosodic structure of a sentence is intimately related to, and constrained by, its phonological, syntactic,
and information structures (e.g., Shattuck-Hufnagel & Turk, 1996; see Dahan, 2015, for a recent review). This relationship is best illustrated by the fact that prosody can reveal aspects of these structures. For example, the phonemically ambiguous phrases “crisis turnip” and “cry sister nip” can be distinguished from each other based on the prosodic rendering that the speaker adopts because a different prosodic phrasing is associated with each of the two sequences of morpho-syntactic elements (e.g., Dilley, Mattys, & Vinke, 2010). Likewise, the sentences “Raoul murdered the man with a gun” and “I asked the teacher who left” have more than one literal meaning and speakers may be able to reveal which one they intend through the sentence’s prosody (Schafer, Speer, & Warren, 2005; Snedeker, & Trueswell, 2003; Schafer, Carlson, Clifton, & Frazier, 2000). For instance, the presence of a prosodic boundary before ‘with a gun’ increases the likelihood that listeners interpret ‘with a gun’ as referring to the instrument of the murder, while the absence of such boundary biases them to interpret ‘with a gun’ as a modifier of the noun phrase ‘the man’. These two interpretations correspond to syntactic structures which differ, among other things, in the way ‘the man’ and ‘with a gun’ are syntactically related. Numerous studies have shown that listeners are exquisitely sensitive to the prosodic cues that differentiate interpretations across the many contrasts that have been examined (e.g., Wagner & Watson, 2010).

The influence of prosody on multiple linguistic structures is a challenge for the traditional approach because the strictly sequential ordering of the stages that it assumes limits the dependencies (or interfaces) between structures it can account for. For instance, the traditional approach captures the fact that the syntactic structure depends on identifying the morpho-syntactic elements of the sentence, and their identity is largely constrained on the retrieval of phonemic categories; likewise, the information structure is often viewed as assigning information
status to the syntactic constituents that have been established at an earlier stage. However, research on prosody has revealed that interfaces between structures cannot be fully captured by the linear ordering that the traditional approach embodies. We describe three kinds of challenges that this research has raised.

First, there is ample evidence that elements of the phonological structure affect the establishment of both the syntactic structure and that of the informational structure. For instance, the informational structure of a sentence is constrained by the presence or absence of prominence on some constituents, elements that are encoded in the phonological structure (e.g., Steedman, 2014). For instance, the utterance “I would love some COFFEE” (with an accent on “coffee”) is ill-fitted to the preceding question “Would you like some coffee?”, but very appropriate following the question “What would you like to drink?” The two contexts differ in the information status of ‘coffee’: The first context mentions it explicitly; by contrast, ‘coffee’ is what is being inquired about in the second context. It is unclear how that distinction is maintained if the mapping of the phonological structure onto the information structure is mediated solely through the syntactic structure because the syntactic structure, as traditionally formalized, lacks the relevant categories.

Another challenge to sequential ordering arises from cases that suggest the influence of a late-built structure on an earlier one. Consider, for example, the ambiguous sentence ‘I asked the pretty little girl who is cold’. In one interpretation of the sentence, ‘who’ is a relative pronoun and ‘who is cold’ a relative clause that specifies which girl is under discussion. In another interpretation, ‘who’ is a wh-word and ‘who is cold’ an embedded question and the object of the asking. As shown by Schafer, Carlson, Clifton, and Frazier (2000), the presence of an accent on ‘who’ biases listeners toward an embedded-question interpretation whereas its absence biases
them toward a relative clause one. This bias, the authors persuasively argued, reflects the fact that in English, focused elements often receive a pitch accent. Because a questioned constituent is by definition in focus, listeners are more likely to interpret ‘who’ as a wh-word standing for a questioned constituent if it carries a pitch accent than if it does not. Importantly here, each interpretation is associated with a different syntactic structure. The example illustrates the fact that the syntactic structure can depend on the information structure (itself informed by the sentence’s prominence pattern and encoded in its phonological structure). This dependency cannot be accounted for if the establishment of syntactic structure strictly precedes that of the information structure. The dependencies between these two structures appear to go both ways. Further evidence of bi-directionality between structures comes from demonstrations that listeners are more likely to hear the presence of a major prosodic boundary in locations where such a boundary is licensed by the syntax, even when the acoustic correlates for the prosodic boundary are weak or absent (Buxó-Lugo & Watson, 2016; Cole, Mo, & Baek, 2010). Thus, the prosodic structure influences and is influenced by the syntactic structure.

Finally, there is mounting evidence for inter-dependencies between structures such that the way a given portion of the sentence is accounted for in one structure affects the way it is accounted for in another structure whose establishment allegedly precedes the former. Examples of such inter-dependencies come from Clifton, Carlson, and Frazier (2006). Their study examined the role that the presence or absence of a prosodic phrase boundary, encoded in the phonological structure of the sentence, has on its syntactic structure. As predicted by the sequential and unidirectional approach to sentence interpretation, Clifton et al. (2006) showed that in sentences such as “Pat or Jake and Lee convinced the bank president to extend the mortgage”, the presence of a prosodic boundary after ‘Pat’ leads people to assume that either one
person (i.e., Pat) or two people (i.e., Jake and Lee) talked to the banker; by contrast, a prosodic boundary after ‘Jake’ leads people to assume that two people talked to the bank, Lee and someone who was either Pat or Jake. An account of this effect within the traditional approach has been offered by Watson and Gibson’s (2005) ‘anti-attachment’ hypothesis, according to which the presence of the prosodic boundary is interpreted as “a strong cue not to attach the upcoming word to the last potential attachment site before the boundary” (p. 285).

However, Clifton et al. also showed that when the sentence was modified to include longer noun phrases to refer to the individuals, as in “Patricia Jones or Jacqueline Frazier and Letitia Connolly convinced the bank president to extend the mortgage”, the effect of the pause location on people’s interpretation was significantly reduced. As proposed by the authors, listeners appear to consider multiple influences that operate on the production of prosodic-phrase boundaries, one originating from the syntactic structure and one associated with the phonological structure and the need to produce phrases of a manageable length. The more the presence of a pause is attributed to the phonological factor, i.e., the length of the phrases, the less it is attributed to the syntactic factor. Such a trade-off relationship is difficult to account for in the traditional architecture because the stages are linearly organized and operate on different categories, with syntactic constituency assumed to be impervious to the constituents’ phonological contents.

Another illustration of inter-dependency across structures is offered in Brown, Salverda, Gunlogson, and Tanenhaus (2015). Their investigation built on Salverda, Dahan, and McQueen (2003), who showed that a syllable is more likely to be interpreted as a monosyllabic word (e.g., pump) rather than the initial syllable of a polysyllabic word (e.g., pumpkin) when that syllable is long than when it is short. This effect is attributed to listeners’ knowledge that syllable
lengthening is an acoustic correlate of a prosodic-phrase boundary and such boundaries are necessarily aligned with word boundaries. But syllable lengthening is an acoustic correlate of other prosodic phenomena, including the presence of a pitch accent. Brown et al. (2015) showed that the tendency to interpret syllable lengthening as marking prosodic phrasing diminished when the context was conducive to interpreting it instead as marking a (contrastive) pitch accent. Because the domain of pitch accents is the stressed syllable, not the prosodic phrase, an accent can fall on a monosyllabic word as well as on the first syllable of a polysyllabic word (provided that the syllable is stressed). In their study, listeners’ attribution of the lengthening of a stressed syllable to a pitch accent rather than to a prosodic-phrase boundary was revealed by a diminished tendency to favor the monosyllabic word when the syllable was lengthened than when it was shortened.

How might a processing model of language comprehension capture interdependencies between linguistic levels? The notion of a unidirectional flow of information between hierarchically organized levels of analysis is difficult to maintain. One possibility is to view the phonological structure, which encompasses both the segmental and prosodic structures, as a representation of the sentence that listeners try to explain by deciding what the other structures need to be in order to motivate this phonological structure. Thus, comprehending a sentence is viewed as a data-explanation process, where listeners aim to select the meaning, represented by all the linguistic structures simultaneously, that best explains the data (i.e., the utterance they hear). There is no data reduction or seriality in the process. It is generative because it aims to discover or infer the underlying structures that generated the data. Such a model provides a natural way to capture the inter-dependencies between the various linguistic structures that listeners aim to establish for a given utterance. Indeed, an explanation favored on one such
structure in order to ‘explain away’ the realization of a sentence may compete with a hypothesis at another level if both claim to account for the same portion or characteristic of the phonetic signal.

This approach is not new: Stevens and Halle’s (1967) ‘analysis-by-synthesis’ belonged to this class of models. However, it has recently gained popularity in cognitive science. As discussed by Clark (2013) within the realm of perception and action, generative models explain how people build an interpretation of the world from input to their senses by selecting a set of hypotheses that best explain the sensory signal. These hypotheses are based on the observer’s knowledge what the signal would be like if it were generated by each hypothesis. Hypotheses represent causal structures that may have generated the signal. This view follows Helmholtz’s (1860) view of perception as a process of knowledge-driven and probabilistic inference. Bayesian models of perception and cognition embody this approach (e.g., Norris & McQueen, 2008; Chater & Manning, 2006). The interactive activation model, originally proposed by Rumelhart and McClelland (1982) and revised over the years, offers one implementation of these ideas (see McClelland, Mirman, Bolger, & Khaitan, 2014, for a recent review, and Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010, for a discussion on the similarities and differences between Bayesian and connectionist models).

Generative models are being increasingly recognized in the psycholinguistic community. An example pertaining to the perception of prosody was offered by Frazier, Carlson, and Clifton (2006) and their ‘rational speaker hypothesis’: Listeners use what they know to be the constraints on a speaker’s production of prosodic structures to infer the underlying causes of the prosodic rendering of a sentence. This approach helps explain the fact that a prosodic-phrase boundary is interpreted as the marking of the closure of a syntactic constituent more if the phonological
phrase before the boundary is short (i.e., containing few words or syllables) than if the phonological phrase is long. According to the ‘rational speaker hypothesis,’ listeners attribute the prosodic-phrase boundary to the need to break a long syntactic phrase into more manageable constituents when uttering the sentence, rather than to the marking of syntactic structure. While the scope of the ‘rational speaker hypothesis’ has remained limited, the underlying principles are very general and in line with the generative model approach.

In summary, the traditional approach to spoken language comprehension which set the stage for current models assumes a sequential, stage-by-stage processing architecture. As just reviewed, the unidirectionality with which information flows in such an architecture is at odds with numerous empirical findings. Generative models offer a viable alternative: In such models, listeners simultaneously identify the linguistic structures that best account for the phonetic and phonological characteristics of an utterance, with trading relationships between structures playing a role in explaining the sensory data.

2. Speaker- and context-specific adaptation and plasticity in speech perception

In the previous section, we discussed prosody, an aspect that research on written language has largely ignored. In this section, we discuss another aspect that speech, but not print, conveys: the voice of the talker and the numerous aspects pertaining to his or her current emotions and attitudes, all encoded in much the same acoustic dimensions as those used to convey phonetic, and thus linguistic, information. This research, as we describe below, has radically changed our understanding of how listeners establish the phonological elements of an utterance. Interestingly, the type of modeling that scholars have gravitated toward in order to capture the new findings takes a data-explanation perspective, the same framework as the one we discussed in the preceding section.
As pointed out above, speech conveys much more than the words that a talker chooses to utter. However, comprehension relies in large part on retrieving the linguistic elements that compose an utterance. Because these elements were assumed to be abstract, context-independent categories, speaker normalization was long assumed to be the process by which stable phonetic units could be retrieved. The goal was to remove the impact of speaker specificity from the acoustic signal (see Johnson, 2005, for a review). However, seminal work by Pisoni and colleagues (e.g., Goldinger, Pisoni, & Logan, 1991; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994, 1995) showed that experience with a given talker facilitates the processing of subsequent speech from the same talker. This finding suggests two things. First, the retrieval of the phonological units of a sentence is dependent on the acoustic details with which these units are produced. Second, because the retrieval of these units becomes more effective with experience with a specific speaker, some form of learning takes place during speech perception. A large body of research has since documented listeners’ remarkable plasticity in accommodating speaker-specific pronunciations (e.g., Clarke & Garrett, 2004; Eisner & McQueen, 2005; Kraljic & Samuel, 2006; Magnuson & Nusbaum, 2007; Maye, Aslin, & Tanenhaus, 2008; Norris, McQueen, & Cutler, 2003).

What changes, exactly, as a result of exposure to the utterances of a given speaker? In a feed-forward model that transforms the acoustic signal by reducing it to a sequence of context-independent phonological categories, learning could only affect the normalization process, i.e., the computations that carry out the data reduction. However, Dahan, Drucker, and Scarborough (2008) provided compelling evidence that the categories themselves are changed: Listeners first heard a talker produce words like ‘bag’ with a vowel that has moved toward and approaches that of ‘beg’, a feature found in some American English dialects and limited to ‘g’ contexts. Another
group of listeners heard standard pronunciations of ‘bag’. When listeners later heard words such as ‘back’ with standard vowels produced by the same talker, which version of ‘bag’ they had heard before affected their recognition of ‘back’: Listeners who had heard the non-standard ‘bag’ words were faster at identifying ‘back’ as ‘back’ than listeners who had heard standard ‘bag’ words. If exposure to ‘bag’ with a non-standard vowel altered the normalization process, one would not expect to see impact of this learning on the interpretation of words with the standard vowel such as ‘back’. If, on the other hand, people modified how they expect words like ‘bag’ to sound based on their experience with their non-standard vowels, ‘bag’ is no longer a plausible interpretation to consider based on hearing the beginning of the word ‘back’ with a different vowel. Thus, it appears that people can modify their expectations of how words (or perhaps smaller linguistic categories such as vowels) are expected to sound when uttered by that talker.

What kinds of architecture may capture such learning? Here Bayesian modeling is a framework that lends itself to this kind of approach, and Kleinschmidt and Jaeger (2015) have recently offered an elegant and computationally explicit model following this formalism. Bayesian models of speech perception describe the process as one in which people infer the phonetic (and, at a higher level, lexical) units of an utterance by selecting the hypotheses that explain the data best. Hypotheses can be conceptualized in listeners’ minds as probability distributions over acoustic dimensions. Distributions are mathematical tools to capture the variability with which a given phonetic unit is acoustically realized within, as well as across, speakers and contexts: Values that this unit can take along these dimensions correspond to probability distributions. In this formalism, listeners adapt to a given speaker or dialect by establishing speaker- or dialect-specific distributions and relying on them when interpreting subsequent speech from the same source.
While listeners benefit from prior familiarity with talkers, we know that they are also perfectly able to understand speech from a speaker of their language they have no prior experience hearing, provided that the talker’s idiolect is similar enough to that of known talkers. Thus, as pointed out by Kleinschmidt and Jaeger (2015), speech perception is a process requiring listeners to recognize speech from a familiar talker, to generalize this acquired knowledge to similar talkers, but also not to generalize, and instead to adapt to novel speakers who differ significantly from familiar ones. What prompts listeners to establish and apply speaker-specific distributions, as opposed to resorting to ‘default’ distributions? Kleinschmidt and Jaeger propose that adaptation is triggered when the default distribution associated with a hypothesis favored by additional contextual cues predicts the observed data with low probability. This so-called ‘error signal’ is what allows such models to ‘update their beliefs’, i.e., to learn and adapt (see Norris, McQueen, and Cutler, 2016, for a similar view).

The data explanation approach naturally explains a host of phenomena that point to the flexibility with which listeners adjust to the context when inferring the phonological units present in a given utterance. For instance, Dilley and Pitt (2010; see also Baese-Berk, Heffner, Dilley, Pitt, Morrill, & McAuley, 2014) showed that whether people perceived the presence of a short function word in a phrase (e.g., ‘leisure and time’ vs. leisure time’) depends on the perceived duration of the fragment of speech comprising the syllable that precedes the potential function word site and the function word itself, if produced. If the fragment is perceived as too long to be a single syllable, given the speech rate, a function word is posited. Conversely, if the fragment is perceived as too short to include a function word, its presence is largely undetected. In the light of this work, one must conclude that the segmental representation of an utterance is only partially determined by the spectral cues present in the signal; equally important are the temporal
characteristics of the utterance to which listeners try to attribute lexical content. Furthermore, listeners do not hypothesize specific segments or phonemes solely based on the support for them in the signal per se. Instead, listeners infer the presence of a specific element based on their beliefs that the element in its context would have resulted in the signal. This finding is reminiscent of the evidence that people adjust their expectations of what a word should sound like given what they know about who said it.

Is there any evidence for plasticity and context-specific adaptation at higher linguistic levels? These ideas are receiving increased attention because they can be viewed as natural extensions of the now widely accepted view that much of the syntactic knowledge that people recruit when establishing the structure of sentences reflects the statistics of the language they have been exposed to (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994). These statistics can be viewed as probability distributions associated with structures. It is thus natural to ask how contextually dependent these statistics may be.

There is some evidence that listeners can learn to comprehend and produce structures that are initially judged to be ungrammatical (Luka & Barsalou, 2005; Kaschak & Glenberg, 2004). Furthermore, evidence for such learning is not limited to unusual structures. Syntactic priming, i.e., the tendency for people to use a structure that they have recently used or heard used, is a phenomenon that illustrates such learning (e.g., Bock, 1986; see Pickering and V. Ferreira, 2008, for a review). Studies that capitalize on syntactic ambiguity (i.e., where more than one structure for the sentence is licensed, given the incomplete fragment of the sentence available to the listener or the reader) showed that repeated exposure to a given syntactic structure can also tip people to anticipate this structure more than they would otherwise (Fine & Jaeger, 2013; Thothathiri & Snedeker, 2008; Wells, Christiansen, Race, Acheson, & MacDonald, 2009), or to
process the repeated structure more rapidly (e.g., Myslin & Levy, 2013). These findings suggest that people adjust the probability distributions associated with structures to reflect their experience most closely when this experience is contextually relevant.

Overall, there is a wide consensus that the representations that listeners bring to bear to their interpretation of speech are flexible and can be temporarily adjusted to contexts. In the next section, we discuss how these representations are dynamically recruited in order to comprehend speech incrementally.

3. Incremental Processing of Spoken Language

Contrary to written language, spoken language is fleeting and the sensory information does not remain available to the hearer for later inspection. Furthermore, listeners do not have control over the rate at which that information becomes available. As recently discussed by Christiansen and Chater (2016), these constraints are some of the reasons why listeners do not wait to have heard the complete sentence to build its structures and meaning; rather, they utilize partial, incomplete information to project these structures. For instance, the first sounds of an utterance may result in positing the presence of the fricative ‘th’ followed by the reduced vowel sometimes called ‘schwa’, and this unstressed syllable may be immediately interpreted as indicating the presence of the definite determiner ‘the’, which itself leads to the projection (or anticipation) of a noun phrase. Incremental processing of speech was perhaps best demonstrated by the work of Marslen-Wilson and colleagues (Marslen-Wilson, 1975; Marslen-Wilson & Tyler, 1980), and later confirmed through different paradigms (e.g., Allopenna, Magnuson, & Tanenhaus, 1998).
Furthermore, because elements of a sentence can entertain long-distance relationships with one another, the spoken signal must be encoded in a way that enables later inspection and manipulation in some type of short-term memory buffer. As argued by Christiansen and Chater (2016), the immediate interpretation and organization of the sentence in terms of constituents at multiple levels of linguistic analysis (i.e., the process they call ‘chunking’) enables this encoding.

Just as it has been proposed in other areas of cognition (Clark, 2013), human listeners are viewed as prediction engines who learn the detailed properties of their environments and use this knowledge to anticipate and predict what can follow. While this ‘educated guessing’ appears naturally helpful before any information has been gained, it is also critical in order to interpret what is being heard because sensory cues and their neural encoding are inherently noisy and can therefore provide only probabilistic information regarding the objects of perception people try to retrieve. Integrating multiple sources of information together (i.e., each of the constraints that the context provides together with the spoken signal) helps compensate for the uncertainty associated with each individual source.

Three questions have been addressed with regard to incremental processing: First, what knowledge do listeners base their predictions on? We know that people interpret partial information within the context in which it is encountered, and this rich set of knowledge constrains people’s predictions. This was illustrated by Altmann and Kamide’s (1999) study showing that people can anticipate which type of objects (or their properties) will be mentioned as the upcoming theme of a verb upon hearing a fragment of a sentence that includes the verb such as ‘The boy will eat the…’. The prediction goes beyond the verb and incorporates semantic properties of the subject and the agent of the verb, as subsequently shown by Kamide, Altmann, and Haywood (2003). Thus, upon hearing ‘the girl will ride’, people are more likely to anticipate
a carousel as the upcoming theme of ‘ride’ than a motorcycle, while the opposite pattern is observed following ‘the man will ride’. The specificity of the anticipations suggests that people anticipate what may be coming at multiple levels of representation.

The second question concerns what happens when partial information is compatible with multiple continuations. An incomplete sequence is compatible with and predicts many possible continuations. Are all continuations considered and entertained? This question has been the subject of much debate. In the field of spoken-word recognition, a consensus prevails: Multiple candidates are considered simultaneously, proportionally to their goodness of fit with the sensory information available (Luce & Pisoni, 1998; Marslen-Wilson, 1990; Marslen-Wilson, Moss, & van Halen, 1996). As more information is accumulated, the fit changes and so does the support for each candidate. There has been little concern for some form of limitation in the resources associated with multiple-candidate consideration, perhaps because the process is assumed to involve passive retrieval of content-addressable information from long-term memory.

In the field of syntactic parsing, most of the work on this question has been done on written language. We nonetheless discuss it here because the theoretical positions that have been proposed may apply to the spoken modality as well. As described above, if people interpret a sentence in an incremental fashion, this means that they project a structure based on partial information. This structure is often viewed as being built and actively maintained in working memory. Because the capacity of working memory is limited, resource limitation has played a major role in constraining theoretical approaches (see Clifton & Staub, 2008, for a review). According to serial approaches, only one parse can be pursued; ambiguity is resolved by choosing which structure to maintain, although the criteria for choosing have varied (e.g., Frazier, 1979; Gibson, 1998). Parallel approaches, on the other hand, do not subscribe to the
notion of resource limitation; instead, in a way that is reminiscent to assumption pertaining to word recognition, these approaches posit that multiple parses are held in memory, each one associated with some likelihood that it is the parse intended by the speaker (e.g., MacDonald et al., 1994; McRae, Spivey-Knowlton, & Tanenhaus, 1998). Note that serial accounts do assume that all (or most) parses are established in order to be evaluated relative to one another. Thus, the divergence between the serial and parallel views rests on how long multiple parses can be held in working memory.

Distinguishing between these two approaches empirically has proven challenging (e.g., Van Gompel et al., 2005; Green & Mitchell, 2006, Sturt, 2007). Levy (2008) proposed a model of syntactic-processing difficulty that blurs the distinction even further: In this Bayesian model, grounded in parallel probabilistic ambiguity resolution, the difficulty of processing a new word is equivalent to the degree to which the set of parses (which, in the model’s formalism, is the probability distribution of structures) that are compatible with the sentence before the word is known differs from the probability distribution of the sentence after the word is known (i.e., a measure of ‘surprisal’). Importantly, Levy showed that this quantity is formally equivalent to the negative probability of the word given the sequence of words that preceded it, a quantity that may be difficult to empirically distinguish from what serial models predict if only one analysis were constructed and entertained at a time.

The third question that incremental processing raises concerns what happens when the most plausible structure (whether the only one pursued or among alternatives) is disconfirmed by later-arriving information, a phenomenon that has been extensively studied in the processing of ‘garden-path’ sentences. Although this research has largely focused on the conditions under which garden-path effects emerge, important insight has come from studies examining the
consequences of these effects on sentence comprehension (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira, 2003; Slattery, Sturt, Christianson, Yoshida, & Ferreira, 2013): When asked about the events described in the sentences they read, people seem to hold true propositions derived from both the original and revised structures. This finding suggests that people do not always revise the linguistic representation of the sentence based on subsequent information. Ferreira proposed that in many cases, people represent the meaning of sentences in terms of ‘good enough’ representations, i.e., representations that may not be the most veridical rendition of the sentence but that a constellation of factors has rendered most plausible. A study by van Alphen and van Berkum (2010) reached a similar conclusion: People retrieved the meaning of a monosyllabic word (e.g., ‘pain’) upon hearing a longer word for which it forms the last syllable (e.g., ‘champagne’) if the semantic context preceding the longer word was more congruent with the short than the long word (e.g., ‘The patient asked the nurse when the champagne [would be cold enough to be served]’), even though that parsing would leave some portion of the utterance (i.e., the first syllable of ‘champagne’) unaccounted for.

Garden-path effects have also been examined in the realm of speech perception and word recognition. The influence of subsequent information on the analysis of earlier portions of speech, a so-called ‘right-context’ effect, have been known since Warren (1970): For instance, when the first consonant of a syllable ending in ‘..ite’ was replaced by a cough, the sound that people reported hearing (e.g., ‘b’ or ‘k’) depended on which word the following context made more plausible. The perceptual basis of the effect, as well as its generality, was later confirmed (see Dahan, 2010, and references therein). Influence of the subsequent context can also be observed when the portion that affects the interpretation of the ambiguous sound comes a few
syllables after it (McMurray, Tanenhaus, & Aslin, 2009) or even after a few words (Connine, Blasko, & Hall, 1991; Szostak & Pitt, 2013), as in phrases like ‘the t/dent in the forest/fender’.

Retroactive effects of this sort can be found even with clearly articulated, undistorted speech. Hearing the pronunciation of a word (‘carrot’) can elicit the consideration of a rhyming alternative (‘parrot’), as reflected in listeners’ transient fixations to the picture of the rhyming word (Allopenna et al., 1998). This result cannot be attributed to the occasional misperception of the initial sound because in all cases, people eventually selected the correct referent. Instead, the rhyming portion of the spoken word ‘carrot’ affects the interpretation of its initial sound. Everything else being equal, the initial sound of ‘carrot’ is more likely to be ‘p’ (due to the influence of parrot) than the initial sound of ‘candle’ is (because there is no such word as ‘pandle’). This is, in effect, evidence that the later portion of a spoken stimulus can affect the interpretation of an earlier portion.

The existence of right-context effects raises the question of what kind of representation people hold in memory that allows them to revise their interpretation given later information. One possibility is that listeners make a series of discrete decisions regarding the presence of phonological (and, more generally, linguistic) elements in the speech signal (Christiansen & Chater, 2016). Such decisions may be subject to revisions if subsequent information renders prior decisions less plausible. This approach assumes that people work toward retrieving the combination of discrete phonological elements that the speaker produced. Because of the arbitrary mapping of sound onto meaning, the goodness of match between a sound and its best-fitting category or categories may interfere with extracting a semantic interpretation of the utterance. This approach is akin to the serial parser discussed earlier. Alternatively, people maintain uncertainty over time, and contextual information acts as an additional source of
constraints that, together, modulate probabilistic evidence for lexical candidates (see Bicknell, Jaeger, & Tanenhaus, 2016, for helpful discussion). As pointed out before, the integration of multiple sources of information may reduce uncertainty and thereby render speech perception more robust to noise and variability. This approach is similar to many of the parallel-parser models.

Although the jury is still out, current evidence favors the latter, probabilistic approach: The impact of subsequent context on the identification of an earlier ambiguous sound varies as a function of the phonetic details of the ambiguous sound: It is greatest when the sound is most ambiguous (Connine et al., 1991; McMurray, Tanenhaus, & Aslin, 2009; Szostak & Pitt, 2013). For the integration to show sensitivity to the phonetic details of the sound requires that people have maintained more than the discrete category it best predicts; instead, people appear to maintain the probabilistic information (the uncertainty) associated with past sensory input for several words (e.g., Szostak & Pitt, 2013; Zellou & Dahan, 2017). For how long and why this representation is maintained are questions that future work must elucidate.

4. Processing of Disfluent Speech

In this section, we discuss a feature unique to spoken language, the presence of disfluencies, and their consequences on comprehension. Spoken and written languages often differ in their register. Indeed, written language has usually benefitted from some editing and the author’s formulation is meant to reveal the intended meaning as effectively as possible. Spoken language, on the other hand, is often generated on the fly: Speakers frequently interrupt, backtrack, and correct their utterances, and listeners must interpret these repairs as best they can. Research is illuminating how listeners cope with, as well as benefit from, disfluencies.
Disfluencies are important not just because they exist and we evidently can deal with them, but also because they reside at the intersection of a number of issues researchers have investigated in psycholinguistics over the last fifty years: incremental processing, revision of erroneous interpretations, integration of different sources of information, and prediction. If language is processed incrementally, then it is to be expected that a disfluency in the input is a piece of language that the listener will process and typically integrate into the ongoing representation of the utterance. Identifying a word or stretch of words as an error on the part of the speaker requires that information at the lexical, syntactic, prosodic, and semantic levels be integrated to support the inference that part of the input does not belong in the ultimate representation for the sentence.

Studies of the comprehension of disfluent utterances have mainly investigated fillers such as ‘uh’, ‘um’, and ‘er’, and these investigations demonstrate that listeners tend to expect speakers to refer to something novel (Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Barr & Seyfeddinipur, 2010) or unfamiliar (Arnold, Kam, & Tanenhaus, 2007) when ‘uh’ precedes a word, compared to a fluent production. This effect is seen even in children as young as two years of age (Kidd, White, & Aslin, 2011). Listeners show a reduced N400 to a less predictable word when that word is preceded by ‘er’ (Corley, MacGregor, & Donaldson, 2007) but not when preceded by a repetition such as ‘burnt my my tongue’ (MacGregor, Corley, & Donaldson, 2010). These findings indicate that fillers lead listeners to infer that the speaker is having trouble formulating her utterance, and therefore lead to the expectation of something new, unusual, or unpredictable. In addition to the significance they can have on what listeners expect to hear next, our research has demonstrated that ‘uh’s and ‘um’s can influence syntactic parsing (Bailey & Ferreira, 2003; Ferreira & Bailey, 2004). Previous work had shown that it is more difficult for
comprehenders to recover from a garden-path the longer they have been committed to the incorrect analysis, as measured in number of words (Ferreira & Henderson, 1991). Our disfluency experiments demonstrated that it did not matter whether words or disfluencies were used to lengthen the garden-path. They also showed that listeners used the presence of an ‘uh’ or ‘um’ to help resolve a garden-path, because listeners tend to assume that fillers preceded a more syntactically difficult constituent. Fillers also seem to direct attention to subsequent material, facilitating its recall (Fraundorf & Watson, 2011).

A more complex challenge arises from disfluencies including self-repairs (‘turn left I mean turn right at the statue’). One model of how self-repairs are interpreted (Ferreira, Lau, & Bailey, 2004; Lau & Ferreira, 2005) assumes that the process is backwards looking: the mechanism begins from the point of breakdown and then moves backward through the input to arrive at the speaker’s intended meaning. Many models designed to automatically recognize self-repairs in speech are also essentially backward-looking: the properties of the repair and the interruption point trigger non-monotonic operations designed to locate the misspoken word (henceforth, the reparandum) and then delete the content. For example, Zwarts, Johnson, and Dale (2010) have developed an automatic disfluency recognition system that is non-monotonic and succeeds in recognizing a self-repair disfluency about five words after the repair has occurred. This level of performance is superior to many other automatic systems, but Zwarts et al. acknowledge that a more robust computational model would identify self-repairs much earlier.

An alternative approach takes advantage of the evidence for prediction during language comprehension and assumes that listeners can anticipate that a repair will occur, and even what its content will be. This approach to the processing of spoken language containing disfluencies
brings together two models that have been independently proposed and empirically tested: (1) the Noisy Channel model of language processing (Gibson, Bergen, & Piantadosi, 2013) and (2) models of focus which assume that contrastive prosody signals a listener to generate a set of candidates that stand in semantic contrast to some critical item. Putting these together, the idea is that listeners treat each word in the input probabilistically and therefore as something that potentially will be repaired. Listeners predict the repair by creating a contrast set similar to the type of contrast set generated during the processing of contrastive focus, in which it has been shown that reference to, for example, a ‘large pitcher’ causes the comprehender to infer the existence of a set of pitchers, one of which is large (Sedivy, 2002). Recent evidence supports these assumptions (Lowder & Ferreira, 2016): listeners displayed a greater tendency to predict the upcoming repair when the speaker signaled that they had made an error (e.g., ‘…some salt uh I mean…’), compared to their tendency to predict the second conjunct when the speaker signaled the presence of a coordination structure (e.g., ‘…some salt and also…’). A follow-up experiment replicated this finding and also examined prediction during the processing of contrastive focus (e.g., ‘…not some salt but rather…’). Results showed that listeners use information about the negated NP to begin generating predictions about the alternate NP—a pattern that resembles listeners’ tendency to generate predictions during the processing of repair disfluencies. Furthermore, as soon as listeners have evidence that a portion of the input was spoken in error, they begin to generate hypotheses for likely repairs. If listeners receive a sequence that is implausible (‘put the milk in the stove…’), they begin the process of repairing the input even before there is any explicit, bottom-up evidence of any error. These findings confirm listeners’ ability to process linguistic input rapidly and integrate linguistic representations formed online with knowledge and expectations generated on the basis of linguistic experience.
To summarize, research on the comprehension of spoken language containing disfluencies reinforces the conclusions that have emerged from more traditional studies of fluent speech. Listeners incrementally process the input, leading them to integrate content that may later need to be removed from the sentence representation due to speaker error. Listeners use both backwards-looking (revision) and forward-looking (anticipation) mechanisms to establish an interpretation that corresponds to the speaker model they have built and which captures the speaker’s communicative intentions. Listeners also capitalize on the information value associated with filled pauses, editing terms, and other disfluencies to draw inferences about the speaker’s goals, thus enabling them to build an interpretation that captures the speaker’s intended meaning.

5. Relationship between comprehension and production

There is a growing interest in addressing the nature of the relationship between speech production and speech comprehension (e.g., Dell & Chang, 2014; MacDonald, 2013) and some of the recent developments in research on language comprehension are relevant to this endeavor. In this last section, we discuss what we take as the two theoretical perspectives put forward on the topic.

At various points in this review, we have described research that adopts a view according to which listeners do not decode spoken language, but instead make inferences about what the speaker intended to say. These inferences apply to both the spoken signal that listeners have heard and what may come next. In this section, we discuss two theoretical approaches that capture this general idea. One approach, which this review has discussed on multiple occasions, can be referred to as the Noisy-Channel model, originally proposed by Shannon (1948) and embodied in many of the Bayesian models of language processing (see Gibson, Bergen, & Piantadosi, 2013). Here, listeners infer what the speaker is likely to have said by combining two
sources of information: prior beliefs about what may and may not be said in the present context, and the likelihood that each possible hypothesis would have given rise to the acoustic signal. The other approach, which we call the Interactive Model of Dialogue, assumes a direct connection between production and perception.

Noisy Channel Models refer to a family of models that assume that processing takes place over a noisy, and therefore variable, communication channel. As alluded to earlier, speech always contains some degree of ambiguity; furthermore, the neural encoding of this signal is noisy. Thus, a rational comprehender whose goal is to recover the intention behind an utterance can make the best guess by combining the sensory data with any prior knowledge of what is likely or unlikely to be said. This idea is the foundation for many models of language processing (Levy, 2008, MacDonald, 2013). The information that is accessed to establish the priors can range from biases related to structural forms all the way to beliefs concerning speaker characteristics (e.g., van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008).

Noisy Channel models imply a tight, albeit indirect, connection between listeners and speakers because the listener’s goal is to infer the speaker’s communicative intention. Models of interactive dialogue assume an even closer connection, one in which interlocutors closely collaborate during dialogue, allowing them to efficiently communicate (e.g., Clark, 1996). As anyone who has attempted to transcribe a real world conversation can attest, the language that people produce in conversations is quite different from the way people express their thoughts in writing: Conversational turns overlap, interruptions occur, many utterances are mere fragments, and errors are frequent. But although transcription might be difficult, conversation is typically quite easy. One reason that conversation is easy despite the apparent challenges, Pickering and Garrod (2004) have argued, is that interlocutors tend to align at every level of linguistic
representation. This insight is captured in the Interactive Alignment model (Pickering & Garrod, 2004). Interlocutors converge on the same referring expressions; they even come to imitate each other’s accents and speech styles (Giles, Coupland, & Coupland, 1991). With respect to syntax, the Interactive Alignment model views the phenomenon of syntactic priming as a reflection of this alignment tendency: speaker and listener come to use the same grammatical forms, even when they communicate across different languages (Hartsuiker & Pickering, 2008). This alignment skill is supposed to arise from people’s tendency to model each other’s minds (see Pickering & Clark, 2014). For the interlocutors’ representations to be compatible, they must be aligned – the representations must be stated over similar vocabularies. This means, in effect, that the machinery involved in the real-time comprehension of an utterance and its working are those involved in the production of the same utterance. This is a stronger claim than that made by Bayesian approaches such as the Noisy Channel model and one that future research will actively examine.

6. References


http://doi.org/10.1006/jmla.1997.2558


http://doi.org/10.1016/S0010-0277(99)00059-1


http://doi.org/10.1017/S0140525X12002440


http://doi.org/10.1016/j.jml.2011.03.004


http://doi.org/10.1073/pnas.1216438110


http://doi.org/10.1111/j.1467-7687.2011.01049.x


http://doi.org/10.1080/01690960444000142


 Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology, 47*(2), 204-238.


http://doi.org/10.1006/jmla.2001.2812


